Overcoming Challenges in Macromolecular Crystallography

David Scott Tourigny

A dissertation submitted to the University of Cambridge in accordance with the requirements for award of the degree of Doctor of Philosophy.

Trinity College and MRC Laboratory of Molecular Biology

July 2014

31,000 words

Abstract

This thesis is composed of two distinct parts that address some long-standing problems in macromolecular crystallography. Part I is concerned with structural studies of ribosomal translocation, the least well-understood stage of the translational elongation cycle, and consists of experimental results obtained during the first two years of my degree. I present a crystal structure that has advanced our understanding of translocation and indicates a universally conserved mechanism of GTPase activation on the ribosome. I also describe a structure of the ribosome bound to an antibiotic known to interfere with this process, and discuss both structures in light of recent developments within the field. Part II is devoted to theoretical work completed during the final year of my degree. I describe the derivation and implementation of a likelihood-based algorithm that estimates true covariances between different crystals using intensities and standard deviations. This enables missing data to be predicted on the basis of data from other crystals and the relatedness between them. With the combination of theory and experiment being what it is, the reader may find themselves unfamiliar with methods or terminology used in certain places. I have therefore included a lay summary of each main result that is intended to suit a general audience- hopefully this proves useful even if the reader happens to be an expert in all fields.

Acknowledgements

I would like to thank my supervisors Venki Ramakrishnan and Garib Murshudov for their continued support over the course of my degree. It has been a privilege to learn from two outstanding scientists whom I will do my best to emulate during the rest of my academic career.

This work was only made possible with help from members of the Ramakrishnan and Murshudov groups, and in particular I would like to thank Ann Kelley, Israel Fernández, Rob Nicholls and Paul Emsley for their contributions and guidance.

I am indebted to everyone who gave me the opportunity to make it this far and I must take this chance to express my gratitude to Tristan Cogan, Julian Ketley, Peter Moody, John Schwabe and Jon Scott.

Finally, I would like to thank Jess, Simon, Tobi, my family and friends for their love and encouragement.

Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Signed:

Date:

Contents

Lay summaries

0.1 The ribosome in an intermediate state of translocation

Ribosomes are the macromolecules responsible for protein synthesis inside all living cells. They join the building blocks of proteins (amino acids) as these are delivered by transfer RNA (tRNA) molecules in the order specified by a genetic message on messenger RNA (mRNA). A ribosome proceeds with strict directionality along an mRNA transcript whilst the process of translation is carried out with great speed and efficiency. An obvious question is how tRNA and mRNA molecules translocate through the core of the ribosome, but in fact this has taken almost half a century to answer properly. Confirming the so-called hybrid state model proposed in 1968 required an atomic structure of the ribosome trapped in the act of translocation. Using elongation factor G (EF-G), a protein factor that assists with translocation, we were able to crystallise the ribosome in the main intermediate (hybrid) state (Tourigny et al., 2013a). The X-ray crystal structure reveals a dramatic conformational change between the two individual subunits of the ribosome that forces tRNA molecules into the hybrid state. This provides a mechanism by which the fundamental process of translocation is coupled to GTP hydrolysis, a source of the energy required to

power translation.

The hybrid state tRNA is found in a highly distorted conformation on the rotated ribosome and the structure reveals it to be stabilised by the L1 stalk, a component of the larger (50S) ribosomal subunit. In fact, this was the first time that the L1 stalk had been visualised in its entirety as it swings into a closed conformation to form electrostatic interactions with the phosphate backbone of the tRNA. These contacts are to be assumed conserved throughout translocation and are probably an essential feature for stabilisation of the hybrid state. On the other side of the ribosome, the activated state of EF-G is also revealed for the first time by this structure. In comparison with the structure that EF-G assumes in isolation, a region called domain IV has moved towards the tRNA binding site. After GTP hydrolysis is initiated domain IV moves even deeper into the tRNA binding site, preventing back-translocation and forcing tRNAs forward as the ribosome rotates.

Another key insight provided by this study is the actual mechanism by which the ribosome is able to activate GTP hydrolysis on EF-G once the factor has bound. A conserved region of 50S RNA (called the Sarcin-ricin loop) uniquely positions important residues of the protein near a water molecule involved in the reaction. Remarkably, the structure formed at this active site is essentially identical to that of elongation factor Tu, the protein factor involved with the delivery of amino acids and tRNAs. Since these residues are conserved amongst almost all translational factors utilising GTP as a source of energy it is likely that the mechanism we propose it common to all species. In fact, since this process is so fundamental, the mechanism revealed by this structure appears to be shared by all organisms ranging from bacteria to humans.

0.2 How pactamycin analogues interfere with translocation

Over 60% of known antibiotics work by targeting the translational machinery. The most effective antibiotics used in clinical treatment exploit subtle differences between functional sites of bacterial or parasitic ribosomes and those that belong to humans. Atomic structures of antibiotics bound to the ribosome are crucial for understanding how they work on a molecular level. These structures can provide a detailed description of how every atom of an antibiotic interacts with the ribosome, allowing scientists and pharmaceutical companies to design ever-more potent drugs.

The antibiotic pactamycin was first isolated from the bacterium *Streptomyces pactum* as a potential anti-tumour drug, but exhibits inhibitory activity against human ribosomes as well as those of many other eukaryotes and bacteria. This makes it impossible for use as a general antibiotic since treatment will compromise the host. Recently however, biosynthetic and chemically derived analogues of pactamycin have attracted interest as the need for new antibiotics continues to grow. An intermediate in pactamycin synthesis, de-6-methylsalicyly (MSA)-pactamycin, displays equivalent anti-parasitic, antibacterial, and anti-tumour activity to pactamycin, but derivatives of this intermediate have significantly reduced toxicity to mammalian cells. In order to explain this result, we solved the X-ray crystal structure of de-6-MSA-pactamycin bound to the ribosome (Tourigny et al., 2013b).

The structure reveals that de-6-MSA-pactamycin binds at the same site where tRNA must bind before it is released from the ribosome (the E-site). This prevents tRNAs from translocating from the adjacent P-site on the ribosome during translation of the genetic message, blocking the process of

protein synthesis in bacterial cells. Based on the structure, a chemical bond between de-6-MSA-pactamycin and the ribosome is proposed to separate bacteria from mammals, and modifications of the antibiotic at this site have been shown to reduce affinity for mammalian ribosomes even further. Other antibiotic-ribosome interactions explain the potent antimalarial features of this drug. Consequently, the results of the study pave the way for development of new and improved analogues with effective anti-protozoal and possible anti-tumour activity.

0.3 The likelihood method for multi-crystal data processing

It is often not possible to collect an entire set of diffraction data from a single crystal of large macromolecular complexes like the ribosome. This can be due to their susceptibility to radiation damage or weak diffraction in certain directions, and data from more than one crystal must be merged prior to structure solution. Many problems are encountered when combining data from multiple crystals. A common approach to combining data relies on the assumption that different data or 'wedges' can all be related to a single underlying structure. Data from different wedges are then combined in the same way that reflections on different images are put on the same scale. As the level of isomorphism decreases between crystals however, it is often the case that merging will fail due to the incompatibility of variable unit cells. Even if merging is achieved, the statistics of combined data will be poor and a dataset will not reflect the true quality of information contained within each of the individual wedges.

We have developed a novel approach to working with multiple data sets that accounts for differences between poorly-isomorphic crystals based on the

principle of maximum likelihood. This involved the derivation and implementation of a likelihood-based algorithm that estimates true covariances between different crystals using diffraction intensities and standard deviations. The true covariance between crystals enables them to be clustered prior to structure solution to devise a strategy for further processing. All tests indicate that estimates remain robust with noise and are insensitive to how far a data set is from being complete. Once equipped with true covariance, the algorithm then proceeds to predict the values for missing data based on observed data from other crystals and the relatedness between them. For example, on the basis of true covariance the algorithm is used to 'complete' three incomplete data sets from three different crystals and predicted data are then used to solve each structure independently. One consequence is that the algorithm is particularly well-suited to revealing the presence of a ligand bound to a protein where only 50% of the data from that crystal have been observed.

Part I

Ribosome translocation

Chapter 1

Introduction

1.1 The translational pathway

1.1.1 A brief history of the ribosome

During his time spent developing cell fractionation, Albert Claude became the first person to document the particles that would later become known as ribosomes. He coined the term “microsomes” after noticing the mitochondria of chicken embryos contained a large fraction of granular particles that he presumed to be involved in anaerobic glycolysis (Claude, 1943). A decade later, Mary Petermann and Mary Hamilton based the purification of microsomes on the use of analytical ultracentrifugation (Petermann and Hamilton, 1952). Like Claude, they had been interested in the differences between the mitochondria of normal and malignant animal cells that appeared in constitutions of the microsomal fractions. Their discoveries relied on the method of sucrose sedimentation developed by Hogeboom, Schneider and Palade (Hogeboom et al., 1948) that allowed separation of the microsomal material into discrete, individual granules of distinct sizes. Petermann used the word “macromolec-

ules" to describe the particles found to be rich in ribonucleic acid (Petermann et al., 1953).

Research progressed rapidly during the 1950's. In 1955 and 1956, Philip Siekevitz and George Palade used electron microscopy to confirm Claude's hypothesis that microsomes were fragments of the endoplasmic reticulum (Palade, 1955; Palade and Siekevitz, 1956). Paul Zamecnik's work on protein biosynthesis culminated with the observation that radioactively labeled amino acids were predominantly incorporated into the microsomal fractions of rat livers, and his group succeeded in making the first cell-free system capable of protein synthesis using microsomes (Keller et al., 1954; Zamecnik and Keller, 1954). By now the idea that ribonucleic acid (RNA) made up the majority of the microsomal particles responsible for protein synthesis was undisputed, and Richard B. Roberts suggested the term "ribosome" should be used when referring to ribonucleoprotein particles of the microsome fraction (Roberts, 1958). Around the same time in Cambridge, George Gamow founded the "RNA Tie Club" whose aim was "*to solve the riddle of the RNA structure and to understand how it built proteins*" (Watson, 2001). Then came the famous Crick, Brenner *et al*. experiment demonstrating that three bases of DNA (one codon) code for one amino acid in the genetic code (Crick et al., 1961).

In 1955, Crick had proposed his "adaptor hypothesis", which suggested a (then) unknown molecule was responsible for carrying amino acids to the ribosome in the order specified by an intermediate messenger RNA (mRNA) nucleic acid template (Crick, 1958). Zamecnik had recently discovered these adaptors, now called transfer ribonucleic acids (tRNAs) (Hoagland et al., 1958), and using this knowledge Marshall Nirenberg and Philip Leder were able to determine the sequences of 54 out of 64 codons in light of the Crick, Brenner *et al*. paper and work by Har Gobind Khorana (Nirenberg et al., 1965). Enumer-

ation of the genetic code coincided with an understanding of what constituted ribosomes (sucrose sedimentation had again proved a useful tool for characterising these macromolecules). By 1958, Alfred Tissiéres and James Watson had shown that prokaryotic ribosomes with a sedimentation coefficient of 70S (the ratio of a particle's sedimentation velocity to the acceleration that is applied to it) could be dissociated into 50S and 30S particles each containing around 63% RNA and 37% protein (Tissières and Watson, 1958). Alexander Spirin and Charles Kurland then achieved further separation and characterisation of the bacterial 23S and 16S ribosomal RNAs (Kurland, 1960; Spirin, 1961), and a 5S RNA was identified as part of the mature 50S ribosomal subunit in 1963 (Rosset and Monier, 1963). Eukaryotic ribosomes, being significantly larger with a sedimentation coefficient of 80S, are composed of large 60S particles and small 40S particles that consist of 28S (plus 5.8S) and 18S RNAs respectively.

Proteins associated with translation were classified as ribosomal proteins or translational factors according to whether they were permanently affiliated with the mature ribosome or assisted with protein synthesis, respectively. Jean-Pierre Waller initiated the study of the ribosome's protein composition (Waller and Harris, 1961; Waller, 1964) and hypothesised that ribosomal proteins formed a special class of basic proteins that "*quite possibly serves the role of maintaining ribosomal RNA in a suitable conformation for protein synthesis*". This hypothesis concurred with Carl Woese's, Francis Crick's, and Leslie Orgel's suggestions that RNA acted as a catalyst for protein synthesis, which would serve to answer the age-old chicken/egg problem of protein and RNA. It is now understood that ribosomal proteins have little functional role other than stabilising ribosomal RNA, and that the ribosome is the largest known example of an RNA catalyst (a ribozyme) (Cech, 2000). However, it took many more

decades to get to where we are today, and the combined effort of mapping out the translational pathway and gathering atomic resolution structures of the ribosome along the way. The remainder of this introductory Chapter will outline our current understanding of the biochemical pathway responsible for protein synthesis. Figure 1.1, taken from (Schmeing and Ramakrishnan, 2009), gives a beautiful illustration of prokaryotic translation that is separated into three major stages: initiation, elongation, and termination.

1.1.2 Initiation

There are major differences distinguishing the ways in which prokaryotes and eukaryotes initiate protein synthesis on the ribosome. Part I of this thesis is primarily concerned with the translocation event in the elongation cycle (which remains essentially identical in all Kingdoms of life) and so eukaryotic initiation will not be discussed here. It is a complex and poorly characterised process compared with the rest of the translational pathway, and little progress has been made towards resolving numerous controversies in the field. Readers interested in the current state of eukaryotic initiation should consult (Jackson et al., 2010) and references therein. An extensive review of prokaryotic initiation is found in (Laursen et al., 2005).

Prokaryotic initiation involves three protein initiation factors IF-1-3 (Figure 1.1). Initiation begins on the 30S particle of a ribosome separated into individual subunits, bound to IF-3 so as to prevent the 50S particle from re-associating (Karimi et al., 1999; Peske et al., 2005). Initiation requires the ribosome to position a unique species of aminoacyl-tRNA, initiator fMet-tRNAfMet, into the so-called P-site (cf. Chapter 1.1.3) of the 30S over a START codon in an mRNA that is about to be translated. The START codon is nearly always the sequence AUG and N-formylmethionine (fMet) is usually the first

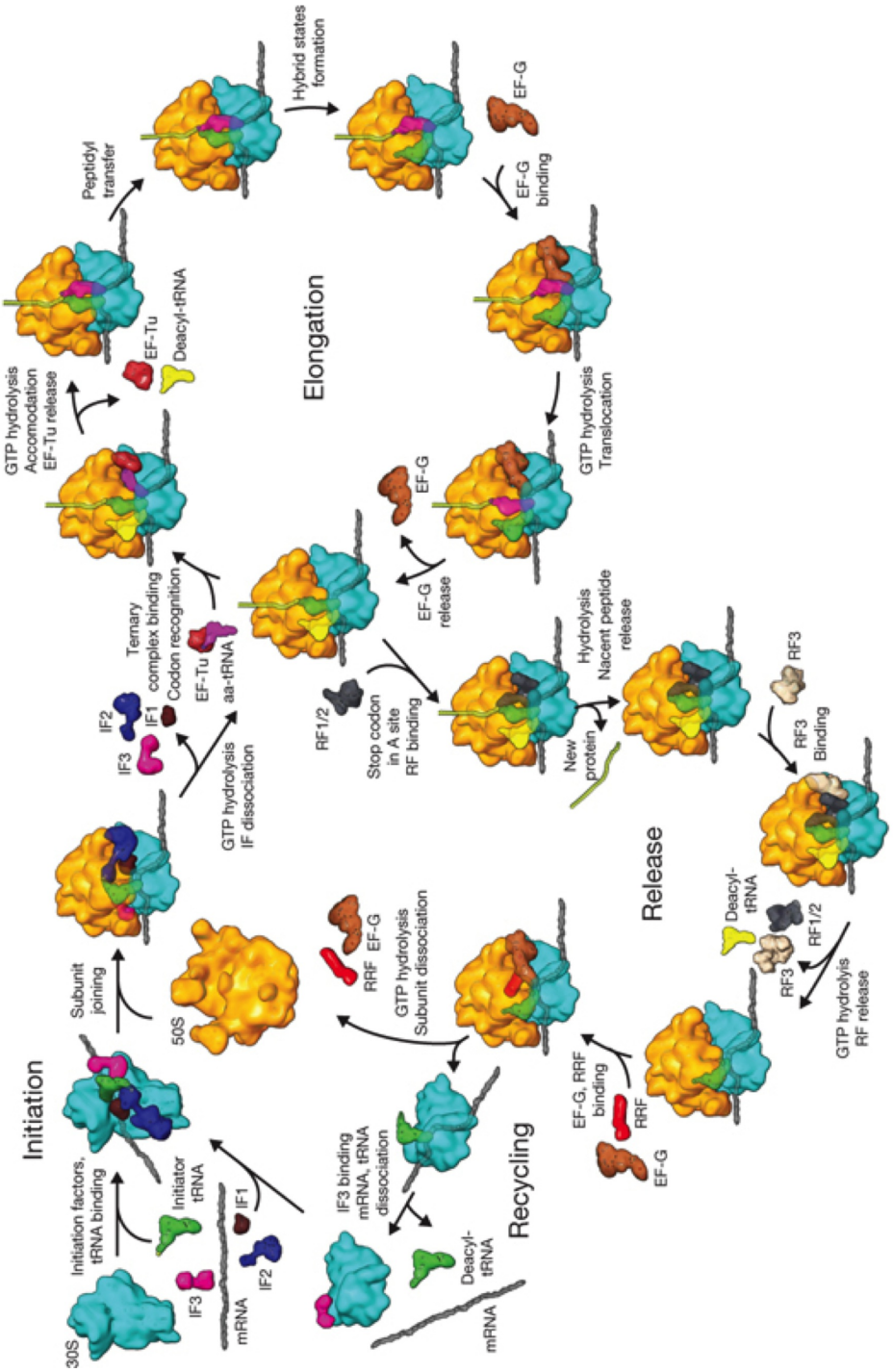


Figure 1.1: An overview of prokaryotic translation

Taken from (Schmeing and Ramakrishnan, 2009).

amino acid to be included into a nascent bacterial protein, even though it may be removed at a later post-translational stage (Sherman et al., 1985). To distinguish an AUG START codon from a conventional AUG methionine codon, the ribosome is able to recognise a six-base consensus AGGAGG Shine-Dalgarno (SD) sequence upstream of the coding region (Shine and Dalgarno, 1974). It is then the first AUG codon to follow the SD sequence that is interpreted as the START codon. Consequently, the SD sequence helps to recruit the ribosome by aligning it with the START codon during the first stage of initiation.

Evidence now points towards the binding of an IF-2-guanosine-5'-triphosphate (GTP) complex occurring just prior to fMet-tRNAfMet recruitment, which may accelerate the process and confer specificity towards the initiator tRNA (Milon et al., 2010). This is contrary to the belief that IF-2 carries fMet-tRNAfMet to the 30S in a ternary complex with GTP (Hershey and Merrick, 2000). IF-1 is also thought to bind the 30S at some stage, although its precise role remains unclear, and the binding of the 30S-IF3 complex to mRNA, IF-1, IF-2 and fMet-tRNAfMet results in the formation of a 30S initiation complex (Laursen et al., 2005). IF-2 is then understood to mediate the recruitment of the 50S ribosomal subunit, which is accompanied by IF-3 release (Antoun et al., 2006; Grigoriadou et al., 2007b; Milon et al., 2008). IF-2 is one of several translational GTPase factors and the 50S subunit is able to act as a GTPase activator to induce GTP hydrolysis on this initiation factor. GTP hydrolysis is followed by the release of IF-1 and IF-2, allowing fMet-tRNAfMet to be accommodated within the peptidyl-transferase centre (PTC) of the 50S subunit in preparation for elongation (Tomsic et al., 2000; Grigoriadou et al., 2007a).

1.1.3 Elongation

This Section will provide a brief introduction to the most conserved part of the translational pathway, which will be discussed again in Section 1.2.2 in light of recent high-resolution structures from the field. Once again the focus is on prokaryotic elongation although elongation is basically identical in all Kingdoms of life (Voorhees and Ramakrishnan, 2013).

After initiation is complete, amino acids are delivered to the ribosome whilst covalently bound to their cognate tRNA in the form of an aminoacyl-tRNA. From early on it was realised that the ribosome must contain more than one tRNA binding site: a peptidyl (P)-site to hold the tRNA attached to the growing polypeptide chain, and an aminoacyl (A)-site to accommodate the aminoacyl-tRNA prior to peptide bond formation (Watson, 1964). There is also a third exit (E)-site that de-acylated tRNAs occupy immediately prior to their release from the ribosome.

The elongation cycle can be roughly divided into three stages (Figure 1.2). During decoding, the ribosome ensures that the correct aminoacyl-tRNA is accepted into the A-site. Aminoacyl-tRNAs are brought to the ribosome in a ternary complex with the GTPase protein elongation factor Tu (EF-Tu) and GTP. Provided the correct aminoacyl-tRNA has been delivered, the ribosome is then able to activate the GTPase activity of EF-Tu in a manner that is presumably similar to the case with IF-2. EF-Tu is released from the ribosome once GTP has been hydrolysed to allow the aminoacyl-tRNA to be accommodated into the A-site and prepared for a new round of elongation. A combination of steady-state kinetic measurements and single-molecule Förster resonance energy transfer (smFRET) experiments have elucidated the many steps involved in decoding (Rodnina and Wintermeyer, 2001; Blanchard et al., 2004), but it has been the structural approaches that have had the biggest impact on our understanding

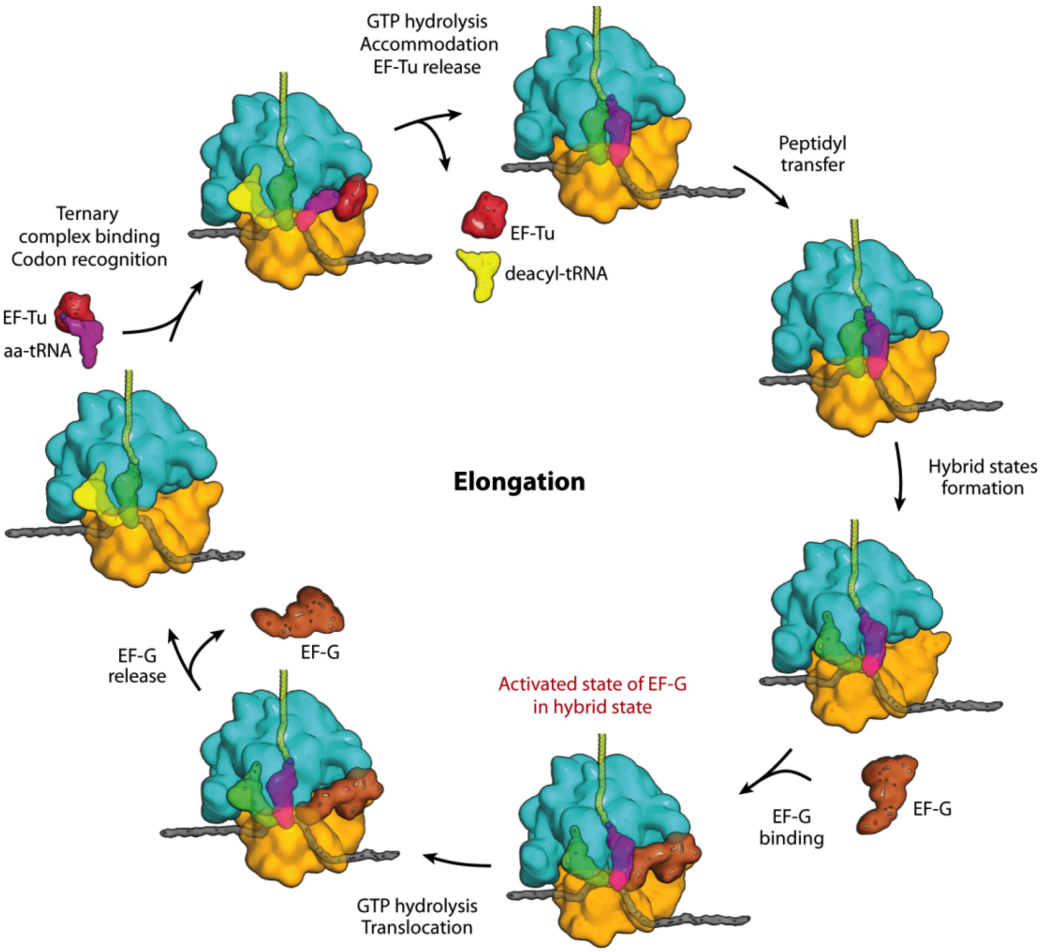


Figure 1.2: An overview of the prokaryotic elongation cycle

Taken from (Voorhees and Ramakrishnan, 2013). The elongation cycle can be divided into three stages: decoding (involving ternary complex binding, codon recognition and tRNA accommodation), petidyl-transfer, and translocation of tRNA across sites and mRNA by a distance of one codon. In red is the activated state of translocation.

of this process. These are discussed in Section 1.2.2.

Accommodation of the aminoacyl-tRNA into the A-site places its α-amino group in position to nucleophilically attack the ester carbon bond of the fMet-tRNAfMet in the P-site. The proceeding peptidyl-transferase reaction is accelerated $\sim$105 fold at the PTC of the ribosome (Sievers et al., 2004), but its pH-independence suggests there to be no general acid/base catalysis involved (Bieling et al., 2006). It has therefore been proposed that the ribosome acts solely by entropic effects (Trobro and Aqvist, 2005; Bieling et al., 2006), although once again these studies have relied on the structural knowledge of the large ribosomal subunit. Once the peptidyl-transferase reaction is complete and the nascent chain has been transferred to the A-site tRNA (leaving a de-acylated tRNA in the P-site), the tRNAs must translocate to P-site and E-site respectively moving up to a distance of $\sim$50 Å. The mRNA must also translocate by a distance of precisely one codon so as to place a new codon in the A-site. Translational GTPase elongation factor G (EF-G) plays a role in assisting with translocation, but for a long time this remained the least well-understood process in the elongation cycle.

Once translocation is complete the de-acylated E-site tRNA is released from the ribosome, which is returned to the beginning of the cycle with an empty A-site and a nascent polypeptide chain in the P-site. The nascent polypeptide is extended one residue at a time as the elongation cycle is repeated until the ribosome encounters a STOP codon. At that point a combination of release and recycling factors terminate elongation and release the fully synthesised protein.

1.1.4 Termination

In the standard genetic code there are three STOP codons that signal the end of the elongation cycle. Like initiation, termination differs significantly between bacteria and eukaryotes. Prokaryotes have two protein release factors (RF-1 and RF-2) that recognise the STOP codons and cleave the nascent polypeptide chain. These are the class I release factors, distinguished from the class II release factor (RF-3) due to their role in termination. Both RF-1 and RF-2 recognise the UAA STOP codon, whereas UAG and UGA are only recognised by RF-1 and RF-2, respectively. Conversely, eukaryotes have only a single release factor, eRF-1, which is unrelated to RF-1 or RF-2 and can recognise all three STOP signals (Frolova et al., 1999; Song et al., 2000).

There are two conserved sequence motifs that enable the prokaryotic class I release factors to recognise STOP codons and sever the polypeptide chain. A PXT tripeptide in RF-1 and an SPF tripeptide in RF-2 confer specificity towards the respective STOP codons (Ito et al., 2000), and both proteins share a GGQ motif with eukaryotes that results in a dramatic decrease in activity when mutated (Zavialov et al., 2002; Mora et al., 2003; Shaw and Green, 2007). Efficient peptide release is dependent on post-translational methylation of the GGQ glutamine (Dinçbas-Renqvist et al., 2000) that is thought to directly coordinate a water molecule for nucleophilic attack on the nascent chain (Weixlbaumer et al., 2008).

The prokaryotic class II release factor RF-3 is known to bind shortly after RF-1/2 to accelerate the disassociation of either class I release factor following peptide release. RF-3 resembles EF-Tu in the GTP state, and so when both release factors are bound to the ribosome simultaneously this is presumed to mimic the decoding stage of elongation where the ternary aminocyl-tRNA-EF-Tu-GTP complex is held prior to GTP hydrolysis (Gao et al., 2007). Like

EF-Tu, EF-G, and IF-2, RF-3 is a translational GTPase, but instead of activating GTP hydrolysis the ribosome has been suggested to induce the exchange of guanosine-5'-diphosphate (GDP) with GTP. This is thought to lead to a conformational change that destabilises the binding of class I release factors and ultimately result in their disassociation from the ribosome. Once RF-3 itself leaves the ribosome, ribosome-recycling factor (RRF), possibly in combination with EF-G (Hirashima and Kaji, 1973) and IF-3 (Singh et al., 2005), is responsible for emptying and separating the individual 30S and 50S subunits in preparation for a new round of translation.

1.2 Structural studies of the ribosome

1.2.1 Structures of the ribosome at atomic resolution

For many decades it was unclear whether, being so large, complex and asymmetric, the ribosome would ever be amenable to X-ray crystallography. Even when the first crystals of the large 50S ribosomal subunit began to appear (Yonath et al., 1980, 1983a,b, 1984) it was not obvious that they would ever diffract to atomic resolution or even if, in principle, an electron density map could be derived from crystals of such a large molecule. By the late-1980's, ribosome crystals that diffracted to moderate resolution were being grown routinely, but it was not until the development of crystallography at cryogenic temperatures that radiation damage caused by high-intensity X-rays could be significantly reduced (Hope et al., 1989).

After a structure of the 30S ribosomal subunit at 5.5 Å (Clemons et al., 1999) and the 50S subunit at 9 Å (Ban et al., 1998) proved that electron density maps could be reliably interpreted, three structures of ribosomal subunits at atomic resolution followed one another in quick succession (Ban et al., 2000;

Schluenzen et al., 2000; Wimberly et al., 2000). The first was of the large 50S subunit of the archaeon *Haloarcula marismortui* whilst the second and third were of the 30S subunit from *Thermus thermophilus*. These breakthrough structures were then themselves succeeded by numerous structures depicting the ribosomal subunits in complex with antibiotics (reviewed in (Yonath, 2005)), tRNA anticodon stem loops (Ogle et al., 2001, 2002), analogues of the peptidyl-transferease reaction intermediate (Schmeing et al., 2005a), and an initiation factor (Carter et al., 2001). The structure of the complete 70S ribosome was reported in 2005 (Schuwirth et al., 2005) and structures of an entire 70S-mRNA-tRNA complex were reported in 2006 (Korostelev et al., 2006; Selmer et al., 2006). Today, the structures of the entire eukaryotic ribosome and its subunits are also known at atomic resolution (Ben-Shem et al., 2010; Rabl et al., 2011; Ben-Shem et al., 2011; Klinge et al., 2011).

The difficulties surrounding the crystallisation of translational GTPases bound to the ribosome will be discussed in Section 2.2, but eventually the problem was solved and structures of the ribosome trapped in various stages of the elongation appeared in 2009 and 2010 (Gao et al., 2009; Schmeing et al., 2009; Voorhees et al., 2010). Together with the insights into decoding and peptidyl-transfer provided by structures of the ribosome with cognate/near cognate tRNA (Ogle et al., 2001, 2002) and transition state analogues (Schuwirth et al., 2005), these form a basis for our current understanding of the elongation cycle.

1.2.2 The elongation cycle from a structural perspective

Our understanding of the elongation cycle just prior to this thesis was reviewed in (Voorhees and Ramakrishnan, 2013), and this Section will largely serve as a summary of the details presented in that paper. For the impact that structural

biology has had on translation as a whole, a suitable reference is (Schmeing and Ramakrishnan, 2009) for the prokaryotic ribosome and (Jenner et al., 2012) for the eukaryotic ribosome.

During decoding initial binding of aminoacyl-tRNA-EF-Tu-GTP to the ribosome is mRNA independent (Rodnina et al., 1996), but the ternary complex will quickly disassociate if the anticodon is not found to be cognate to the A-site codon. Structures of the 30S subunit in complex with codon-anticodon mimics revealed the universally conserved bases of 16S RNA that allow the ribosome to recognise cognate tRNA (Ogle et al., 2001). Residues A1492, A1493, and G530 monitor correct Watson-Crick base pairing at the first and second positions in the minor grove of the codon-anticodon helix. The correct interactions with the helix are dependent on Watson-Crick geometry but allow wobble pairs (e.g. G-U) at the third position, confirming the wobble hypothesis explanation for degeneracies in the genetic code (Crick, 1966). The conformational changes in A1492, A1493, and G530 that occur upon cognate tRNA binding induce a large-scale domain closure of the 30S subunit (Ogle et al., 2002).

There is also a large distortion in the aminoacyl-tRNA that binds to the ribosome with EF-Tu that plays an essential role in decoding. The aminoacyl-tRNA adopts what is known as the A/T state, which is thought to allow it to attain proximity to the A-site codon whilst retaining its interaction with EF-Tu (Schmeing et al., 2009, 2011). Together with the 30S subunit domain closure, tRNA distortion places EF-Tu into a position ready to have its GTPase activity activated by the ribosome. The sequence of events involved in GTPase activation will be discussed in Chapter 2, and are based largely upon the structure of the activated ternary complex bound to the ribosome (Voorhees et al., 2010) together with results presented in this thesis (Tourigny et al., 2013a). Fol-

lowing GTP hydrolysis, release of the inorganic phosphate induces a ∼100° rotation in the GTP-binding domain of EF-Tu with respect to domains 2 and 3. This conformational change weakens the interactions between EF-Tu and the ribosome, and the elongation factor is released allowing the aminoacyl-tRNA to be fully accommodated into the A-site.

An induced fit mechanism involving a series of conformational changes in 23S ribosomal RNA is thought to promote peptide bond formation by exposing the petidyl-tRNA ester to nucleophilic attack (Schmeing et al., 2005b; Voorhees et al., 2009). This allows the α-amino group of the aminoacyl-tRNA to access the petidyl-tRNA ester, and structures of RF-1 and RF-2 bound to the ribosome suggest similar conformational changes are involved in termination (Laurberg et al., 2008; Weixlbaumer et al., 2008). Structures of the 50S subunit implicated 23S RNA residue A2451 as having a catalytic role in peptidyl-transfer (Muth et al., 2000; Nissen et al., 2000), but mutations at this position have almost no effect on the reaction (Polacek et al., 2001; Beringer et al., 2003; Youngman et al., 2004). In light of more recent structures, A2451 is proposed to be involved in a hydrogen-bonding network with the 2'-OH group of A76 of the peptidyl-tRNA (Trobro and Aqvist, 2005; Schmeing et al., 2005a), which is properly positioned to have a direct role in peptide bond formation (Hansen et al., 2002b). Whether the 2'-OH of A76 can actively participate in catalysis remains disputable however, since substituting this functional group with either a hydrogen or a fluorine results in only a modest ∼100 fold reduction in peptide bond formation (Zaher et al., 2011). As noted previously, biochemical results are consistent with the reaction being catalysed by entropic effects, and so the A76 2'-OH may only serve in proton transfer or substrate positioning (Bieling et al., 2006).

Until recently, translocation remained the most poorly understood aspect

of the elongation cycle (Voorhees and Ramakrishnan, 2013). Even though over half a century of work had been dedicated to the study of translocation, details of key intermediates were unknown because atomic resolution data of their structures were still lacking.

1.3 Part I outline

The remainder of Part I is devoted to structural studies of ribosome translocation. The current prevailing view, presented in Chapter 2, resulted from more than 50 years of biochemical experiments being used to interpret a structure of EF-G bound to the ribosome in an intermediate state of translocation (Tourigny et al., 2013a). Section 2.2 describes a solution to the problems faced in obtaining that structure, whilst Sections 2.3 and 2.4 discuss the results and their implications for translation.

With recently acquired structural knowledge, inhibiting the ribosomes of pathogenic bacteria or parasites during translocation will be a future goal of the pharmaceutical industry. The design of antibiotics that interfere with translocation is discussed in Chapter 3, where the structure of a bioactive pactamycin analogue bound to the ribosome is presented in Section 3.2.2. The combined roles that structural and biochemical studies have in advancing our understanding of translocation is summarised with some concluding remarks in Chapter 4.

Chapter 2

The molecular mechanism of translocation

2.1 Prior understanding of translocation

The discovery of the ribosome was followed by researchers finding that the 70/80S RNA-protein complexes could be separated into individual 30/40S and 50/60S components (Littlefield et al., 1955; Tissières and Watson, 1958). When the role that the ribosome played in protein synthesis had become clear, in addition to helping to identify the catalytic and decoding functions of either subunit, separability of the bacterial 30S and 50S particles initiated the first discussions of translocation.

By the mid-1960's it was understood that the ribosome must contain at least two tRNA binding sites (Watson, 1964). It was argued that regardless as to whether peptide catalysis took place on the A- or P-site, at one point or another the nascent peptidyl-tRNA must somehow translocate from the A-site to the P-site in order to repeat a cycle of elongation. Moreover, that the mRNA must travel at the same time by a distance of one codon so as to place the

next codon in the empty(ing) A-site. This raised the fundamental question of how the ribosome, viewed by many as nothing more than an accessorised platform for protein synthesis, would be able to mediate translocation of tRNA and mRNA with such remarkable speed and efficiency as demonstrated by the earliest studies. In 1968, in a two-page letter to Nature, Mark Bretscher proposed an elegant solution to the problem that formed the basis for the next 50 years of research on translocation (Bretscher, 1968).

Bretscher's model for translocation was founded upon a realisation that separability of the ribosome suggested the possibility of an inter-subunit conformational change. Quoting directly from his paper, "*[Some] relative movement of the two ribosomal sub-units it suggested by the universal existence of two separable particles making up a 70S ribosome*" (Bretscher, 1968). He predicted that translocation took place in two steps with an intermediate conformation of the ribosome containing a tRNA that spanned the so-called hybrid A/P-site (and subsequently P/E-site). Formation of this intermediate state was supposed to require a reversible, inter-subunit conformational change in the ribosome that either allowed the peptidyl-tRNA to remain embedded in the A-site on the 30S subunit and first move to the P-site of the 50S (path 2), or first move to the P-site on the 30S whilst remaining in the A-site of the 50S (path1). The two possible pathways with their respective intermediate states are illustrated in Figure 2.1. It was slightly unfortuitous that Bretscher mentioned he favoured the latter, since we now know it is the first scenario (path 2) that occurs during translocation.

Following peptidyl-transfer, translocation of tRNAs on the ribosome were eventually shown to occur spontaneously with respect to the 50/60S subunit, while the anticodon ends and mRNA remain anchored in their original sites in the 30S subunit, resulting in the formation of A/P and P/E tRNA hybrid

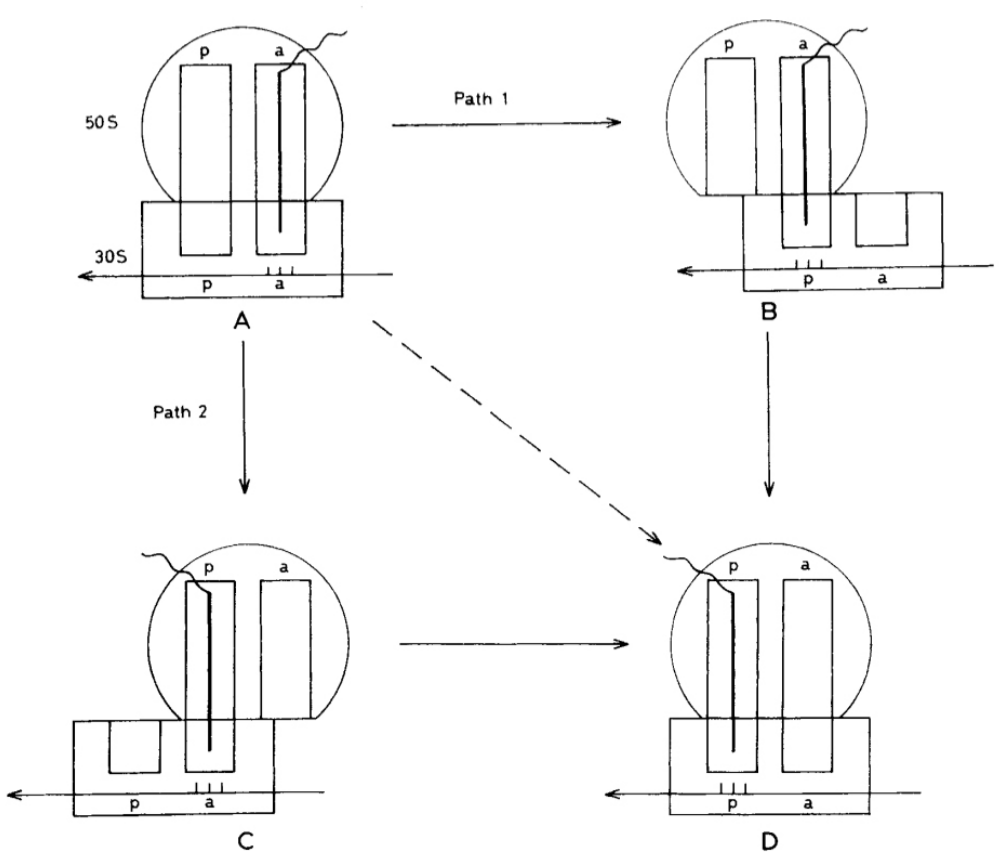


Figure 2.1: Bretscher's hybrid state model

Taken from (Bretscher, 1968).

states (Moazed and Noller, 1989). The intermediate hybrid tRNA states are accompanied by a rotation of the ribosomal subunits relative to one another, together with a series of conformational changes in the L1 stalk and main body of the ribosome (Julian et al., 2008; Agirrezabala et al., 2008). These are the conformational changes predicted in the 1968 letter to Nature. About the same time as Bretscher's prediction, Lipmann argued that the GTPase EF-G, by then known to be associated with the ribosome during translocation, used the energy released upon GTP hydrolysis to "*carry the [ribosome-tRNA] complex one triplet forward*" (Lipmann, 1969). Indeed, during the second step of

translocation EF-G catalyses the movement of mRNA and tRNAs with respect to the 30/40S subunit, thereby placing the next codon of mRNA in the A-site and restoring the ribosome to the canonical, unrotated state.

Various experiments confirmed that EF-G bound to GTP stabilises a rotated state of the ribosome with hybrid tRNAs (Frank and Agrawal, 2000; Dorner et al., 2006; Spiegel et al., 2007; Ermolenko et al., 2007). Crystal structures revealed that EF-G is structurally similar to the ternary complex of EF-Tu, tRNA and GTP, with its domain IV mimicking the anticodon stem-loop of tRNA (Lindahl et al., 1994; Czworkowski et al., 1994; Nissen et al., 1995), and structures of EF-G bound to the ribosome in both the canonical and rotated states were observed by cryogenic electron microscopy (cryoEM) (Valle et al., 2003; Connell et al., 2007; Ratje, 2010). These studies greatly advanced our understanding of the changes in the ribosome induced by EF-G binding, but the lack of high-resolution information meant details of the interactions of EF-G with the the ribosome, and insights into the molecular mechanisms that lead to translocation, were still missing.

It was originally assumed that EF-G simply lowers the free-energy barrier of the spontaneous reaction and that GTP hydrolysis is required to release EF-G from the post-translocated ribosome (Inoue-Yokosawa et al., 1974; Czworkowski and Moore, 1997). The current view, based on kinetic experiments suggesting that GTP hydrolysis precedes and accelerates translocation (Rodnina et al., 1997; Savelsbergh et al., 2003; Chen et al., 2011), is that the rotated-state ribosome plays the role of a GTPase activator for EF-G. Rapid GTP hydrolysis upon ribosome binding is thought to accelerate rate-limiting conformational changes that result in an unlocking of the ribosome leading to translocation (Savelsbergh et al., 2003).

How GTP hydrolysis on EF-G is activated by the ribosome also remained

somewhat controversial, but the mechanism is supposed to be identical to activation of GTP hydrolysis on EF-Tu. Mutation of the highly conserved His84 of EF-Tu to alanine resulted in a 106-fold reduction in catalytic activity (Daviter et al., 2003). The structure of the EF-Tu-tRNA-GTP complex bound to the ribosome showed that His84 in the switch II region was involved in hydrogen-bonding interactions both with A2662 of the sarcin-ricin loop (SRL) of bacterial 23S rRNA and a water molecule positioned for hydrolysis of the γ-phosphate of the GTP analog β-γ-methyleneguanosine 5'-triphosphate (GDPCP) (Voorhees et al., 2010). Suggestions that the histidine might play a role as a catalytic base were later questioned (Liljas et al., 2011; Voorhees et al., 2011; Adamczyk and Warshel, 2011). Subsequently, a structure of RF3 bound to a rotated-state ribosome with a non-hydrolysable GTP analogue (Zhou et al., 2012) placed the histidine in a very different position, suggesting that it was unlikely to play a direct role in catalysis and that any mechanism of GTP hydrolysis is not general, as was first proposed (Voorhees et al., 2010).

Interactions between EF-G and the ribosome were revealed by a structure of GDP-bound EF-G stalled on a post-translocated ribosome (Gao et al., 2009). However, a high resolution of the intermediate state of translocation containing the ribosome with hybrid tRNAs was lacking. This structure was essential for properly explaining the mechanism by which the ribosome is able to activate the GTP hydrolysis on EF-G that ultimately leads to translocation.

2.2 Stabilising the translocation intermediate

As remarked above, a key hitherto missing high-resolution structure in the elongation cycle is that of the ribosome caught in the main intermediate state of translocation. Up to this point, reasons for the failure of achieving the struc-

ture were two-fold. First, until the crystal structure of GDP-bound EF-G stalled on a post-translocated ribosome (Gao et al., 2009), all crystal forms of the entire ribosome from several different species had prohibited the crystallisation of complexes with GTPase translational factors. Analysis of the packing in each of these crystal forms had revealed a shared crystal contact between ribosomal protein L9 and a region of 16S RNA that forms part of the GTPase binding site. To overcome this difficulty, a mutant strain of *T. Thermophilus* in which L9 was truncated to prevent formation of the preferred contact was used to obtain a new crystal form (in space group $P2_1$, see Chapter 5 for a general introduction to crystallography jargon) containing the ribosome with GDP-bound EF-G (Selmer et al., 2012). However, once the first structures of GTPase-bound ribosomes had been solved (Gao et al., 2009; Schmeing et al., 2009) it became clear that ribosomes remained in the canonical, unrotated state in this new crystal form.

For several years it remained unclear whether the new $P2_1$ unit cell would accommodate the rotated conformation of the ribosome or yet another crystal form would be required for the other states of translocation. Additional difficulties were caused by the tendency of crystals being very difficult to reproduce and suffering heavily from radiation damage, probably due to the loss of a crucial crystal contact. Remarkably, whereas two molecules were contained in the asymmetric unit of the first GTPase structures, later structures of EF-Tu trapped with a GTP analogue consisted of only a single unique molecule (Voorhees et al., 2010; Neubauer et al., 2012). In these cases, slight differences in crystal packing aligned the ribosome pair along the two-fold symmetry axis and resulted in the longest direction of the unit cell being reduced by half. Not only were these crystals more resilient to radiation damage, but they also suggested that the $P2_1$ crystal form had some flexibility in ac-

commodating alternative conformations of the ribosome. This was confirmed a year later when the truncated L9 strain was used to solve the structure of RF3 bound to the rotated state of the ribosome (Jin et al., 2011).

Since GTP hydrolysis is thought to be a perquisite for translocation (Rodnina et al., 1997) it made sense to use the non-hydrolysable analogue GDPCP to trap EF-G on the ribosome in the rotated conformation. A simple assay was developed to confirm that inclusion of GDPCP alone was sufficient to trap EF-G on the ribosome (Appendix A.2). Complexes of *T. thermophilus* ribosomes, mRNA, *E. coli* tRNA and *T. thermophilus* EF-G were incubated with either GDPCP or GTP for 20 minutes at room temperature prior to ultracentrifugation over a 1.1 M sucrose solution so that unbound protein remained above the denser supernatant. The re-suspended pellets containing ribosomes and ribosome-bound factors were then analysed using SDS-PAGE analysis. As shown in Figure 2.2, in the presence of GTP EF-G presumably hydrolyses the nucleotide, completes translocation, and is released from the ribosome since only trace amounts of the protein remain bound in the pellet. In the presence of GDPCP however, EF-G remains bound to the ribosome even though the EF-G/ribosome stoichiometry appears much less than one.

The low EF-G/ribosome stoichiometry posed a serious problem for X-ray crystallography where crystallisation of a complex usually relies on the target species making up the majority of the sample. Initially, crystals of the sub-stoichiometric complex were optimised and data were collected to 5.1 Å using a single crystal on the IO4 beam line at Diamond Light Source, Oxford, UK. Even at low resolution, the $F_o - F_c$ difference map was of high enough quality to accurately interpret several distinctive features of the structure. There was clear density for E and P-site tRNAs, mRNA, and an anticodon stem loop in the A-site. Surprisingly, there were also characteristic alpha-helical densities

at the GTPase centre that corresponded well with the superimposed G-domain of EF-G from Gao *et al*. (Gao et al., 2009). However, as indicated by unit cell parameters, the ribosome was found to be in the canonical, unrotated state. In the 2009 structure, domain VI of EF-G protrudes into the A-site and so it was difficult to explain the co-existence of full-length EF-G and an A-site tRNA in the canonical ribosome. The current interpretation is that this complex corresponded to an artificial state of translocation that was perhaps forced into place by crystal packing and use of the non-hydrolysable GTP analogue.

Alongside crystallisation of the complex with full-length EF-G, efforts were made towards the formation of a complex with a truncated version of EF-G. Deletion of domains VI and V of EF-G had been shown to dramatically reduce the speed of translocation without effect on GTP hydrolysis (Rodnina et al., 1997; Savelsbergh et al., 2000). It therefore made sense to use an EF-G mutant lacking these domains to try and shift the equilibrium towards the rotated state. An expression construct containing a truncated version of the *Escherichia coli* EF-G was obtained from M.V. Rodnina and this protein was also expressed, purified, and tested for its ability to bind the ribosome. The truncated protein was difficult to work with and tended to precipitate above concentrations of 8 μM, requiring complex formation in large, dilute volumes to be followed by ribosome pelleting prior to crystallisation trials. It was not overly surprising when complete data collected to 4.3 Å on the PXIII beamline at the Swiss Light Source, Paul Scherrer Institut, Switzerland revealed nothing but ribosomes bound only to mRNA and tRNA in the canonical state.

Up to this point, complexes with full-length EF-G had been formed using two species of tRNA (tRNAPhe and tRNAfMet) and an mRNA that was designed to place a START codon in the P-site and a phenylalanine codon in the A-site. In previous cryoEM structures (Valle et al., 2003; Connell et al., 2007),

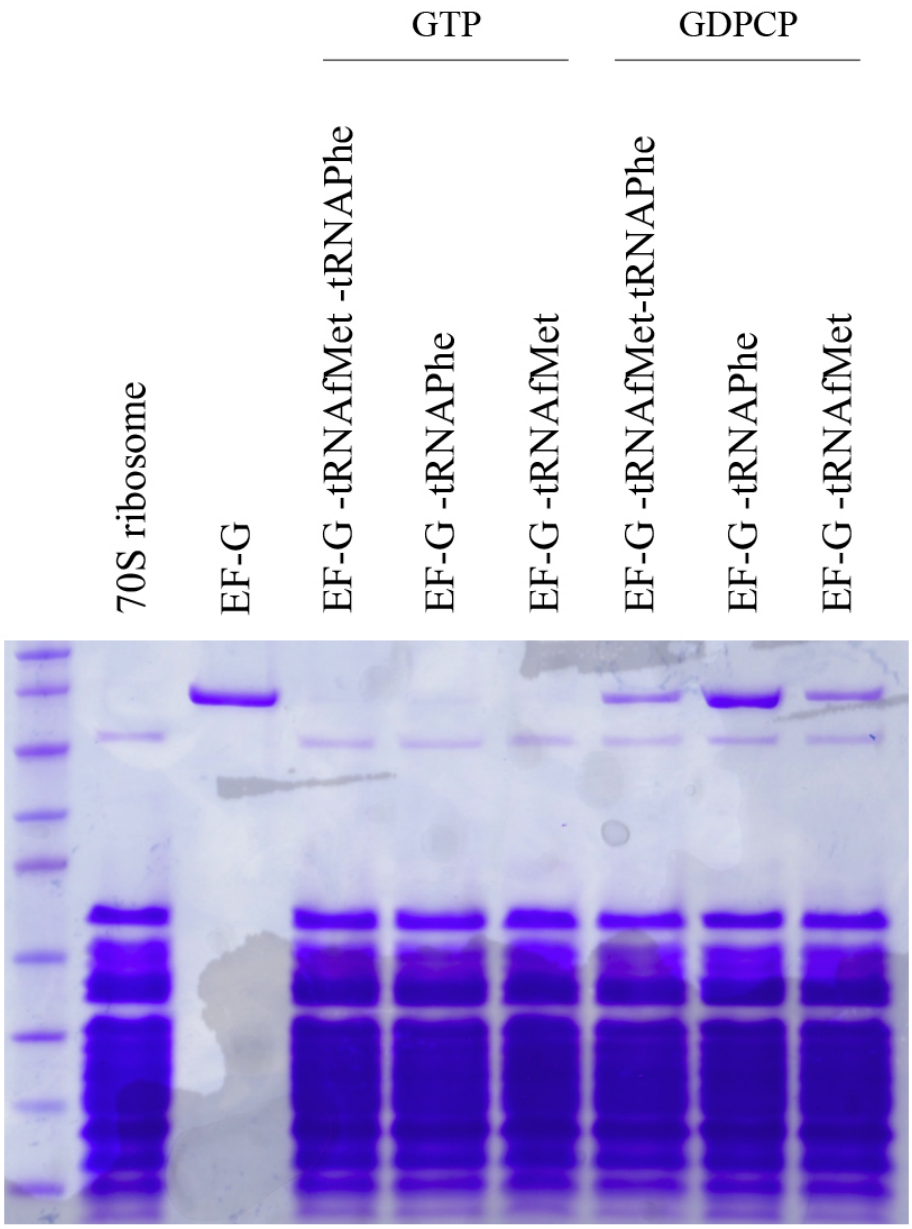


Figure 2.2: Stability of EF-G-ribosome complexes

SDS-PAGE analysis of EF-G-ribosome complexes showing results of the binding assay described in Appendix A.2. EF-G-ribosome complexes were formed together with tRNAfMet, tRNAPhe, or both, either in the presence of GTP or the non-hydrolysable analogue GDPCP. Far left lane contains a protein marker.

the A-site tRNA was left out in order to obtain the intermediate rotated state of the ribosome because in its presence with wild-type EF-G the ribosome proceeds within seconds to the post-translocational canonical state even without GTP hydrolysis (Rodnina et al., 1997; Ermolenko and Noller, 2011). Consequently, the next approach was to leave an A-site tRNAPhe out of the complex formation pathway in an attempt to increase full-length EF-G/ribosome stoichiometry and stability. Figure 2.2 shows that leaving tRNAPhe out of the complex did little to increase stoichiometry. Remarkably however, when tRNAfMet was left out instead of tRNAPhe in a control experiment there was a dramatic increase in stoichiometry and presumed stability of the complex. Optimised crystals of the tRNAPhe complex diffracted beyond 2.9 Å at Swiss Light Source and Diamond Light Source, and the data were used to solve a structure of the rotated ribosome bound to EF-G with GDPCP, consisting of an mRNA with a phenylalanine codon in the P-site, and a tRNAPhe in the P/E hybrid state. The structure lacks an A-site tRNA, but otherwise represents a key missing structure that is the intermediate state of translocation.

2.3 Structure of the translocation intermediate

The results presented in this Section are essentially the same as those described in (Tourigny et al., 2013a).

2.3.1 Overall structure

Full material methods are presented in Appendix A along with crystallographic data statistics in Table A.2. After molecular replacement using the 50S and 30S subunits as search models, the P/E tRNA, mRNA, EF-G and GDPCP were clearly visible in difference Fourier maps (Figure 2.3; mRNA not shown), and

the entire structure of EF-G bound to the rotated ribosome was built and refined (Figure 2.4). Ribosomal RNA of the small subunit required extensive remodelling since, as described below, there are substantial conformational changes associated with 30S rotation and head swivelling. Due to the high-quality difference maps, additional features including extensions on several ribosomal proteins could be modelled for the first time.

The main body of the 30S subunit is rotated ∼7° counterclockwise with respect to the 50S (as viewed from the solvent side) (Figure 2.5). Although the precise rotation angles differ, the inter-subunit interactions and central bridges are similar to those previously seen in the hybrid state with RRF (Dunkle et al., 2011) or RF3 (Jin et al., 2011; Zhou et al., 2012), suggesting a ratcheting motion that is conserved across the translational pathway. The head of the 30S is swivelled by ∼5° as compared to the canonical state (Figure 2.5). Two separate ratcheted states that differ in the degree of head swivelling have been identified by cryoEM of an EF-G- ribosome complex (Ratje, 2010). As displayed in Figure 2.5, the 30S head of this structure has a conformation close to that of the TIPRE state in that cryoEM study (r.m.s.d. of 1.7 Å as opposed to 11.1 Å when compared to TIPOST state). Recently it was shown that the TIPRE state also closely resembles cryoEM reconstructions of ribosomes containing both P/E and A/P hybrid tRNAs after peptidyl -ransfer (Agirrezabala et al., 2012), which is further evidence that the current structure represents a valid model for the main intermediate state of translocation. The head swivel is thought to widen a constriction in the 30S to allow translocation of the P-site tRNA to the E-site (Schuwirth et al., 2005; Selmer et al., 2006; Ratje, 2010). In the rotated state seen here, this constriction is widened by ∼2.7 Å compared to the canonical state, suggesting that further widening must occur at some point to allow translocation of the anticodon stem-loop of tRNA from

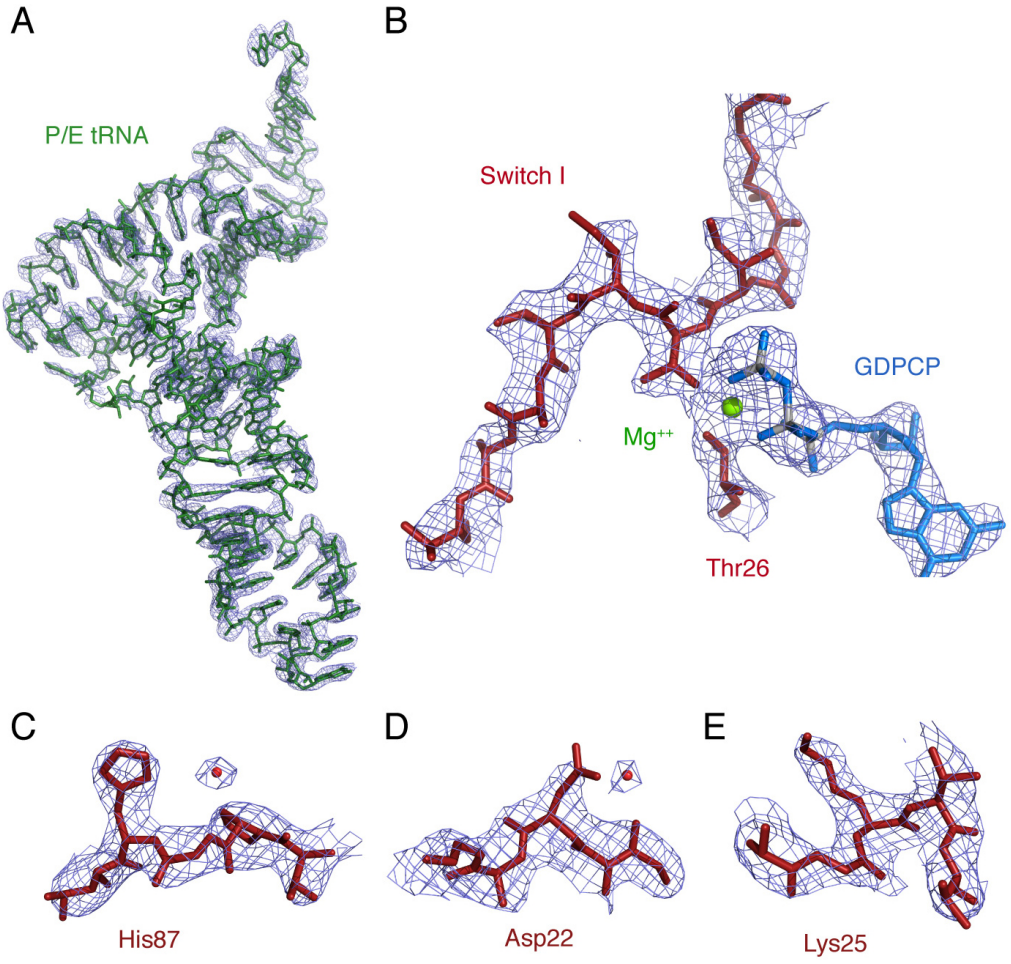


Figure 2.3: Omit difference Fourier maps

Omit difference Fourier maps contoured at 2.4σ obtained after initial refinement with an empty ribosome as a starting model, showing (A) P/E tRNA, (B) switch 1 and GDPCP in the active site, (C-E) key conserved residues in the active site with water molecules.

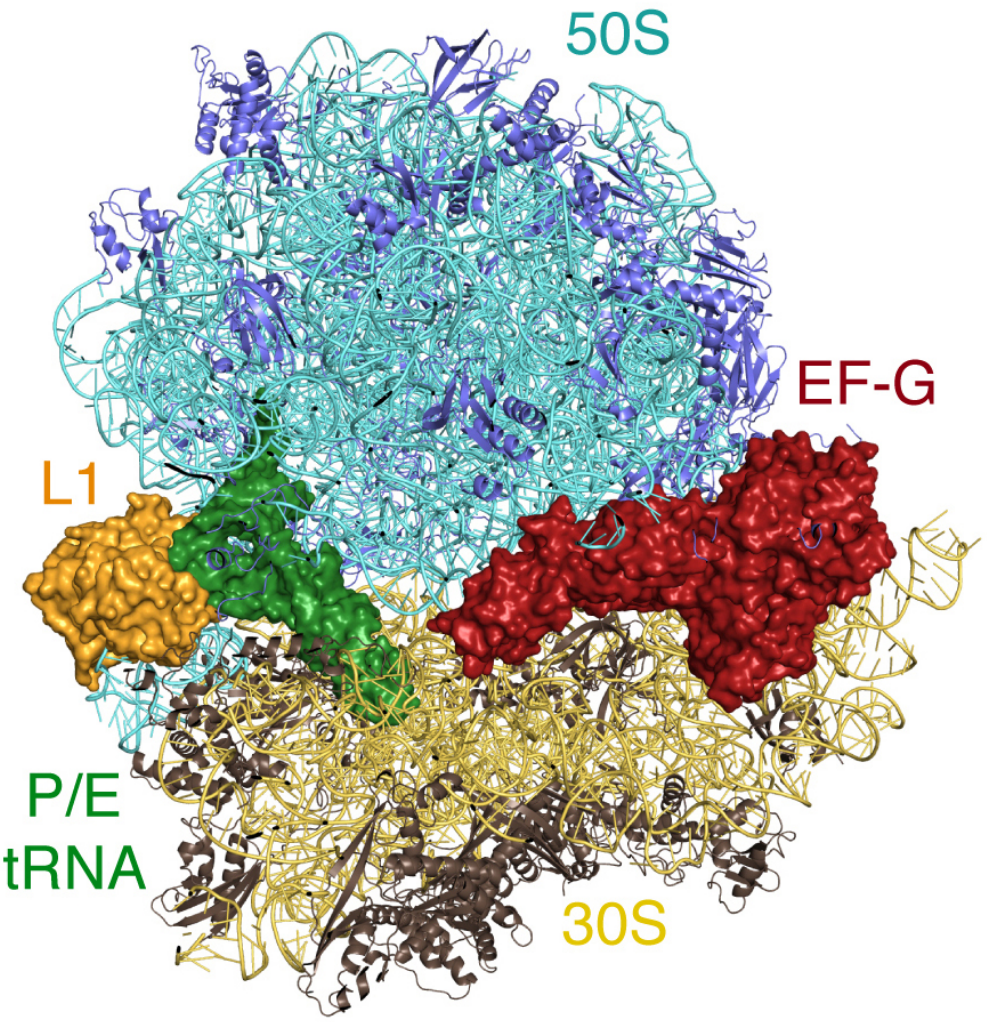


Figure 2.4: Structural overview

Complete structure of EF-G (red) bound to the rotated ribosome with a tRNAPhe (green) in the P/E hybrid state. The RNA components of the large 50S ribosomal subunit are shown in cyan and the protein components in blue. The small 30S subunit is coloured yellow (RNA) and brown (protein), and the L1 protein of the L1 stalk is shown in orange.

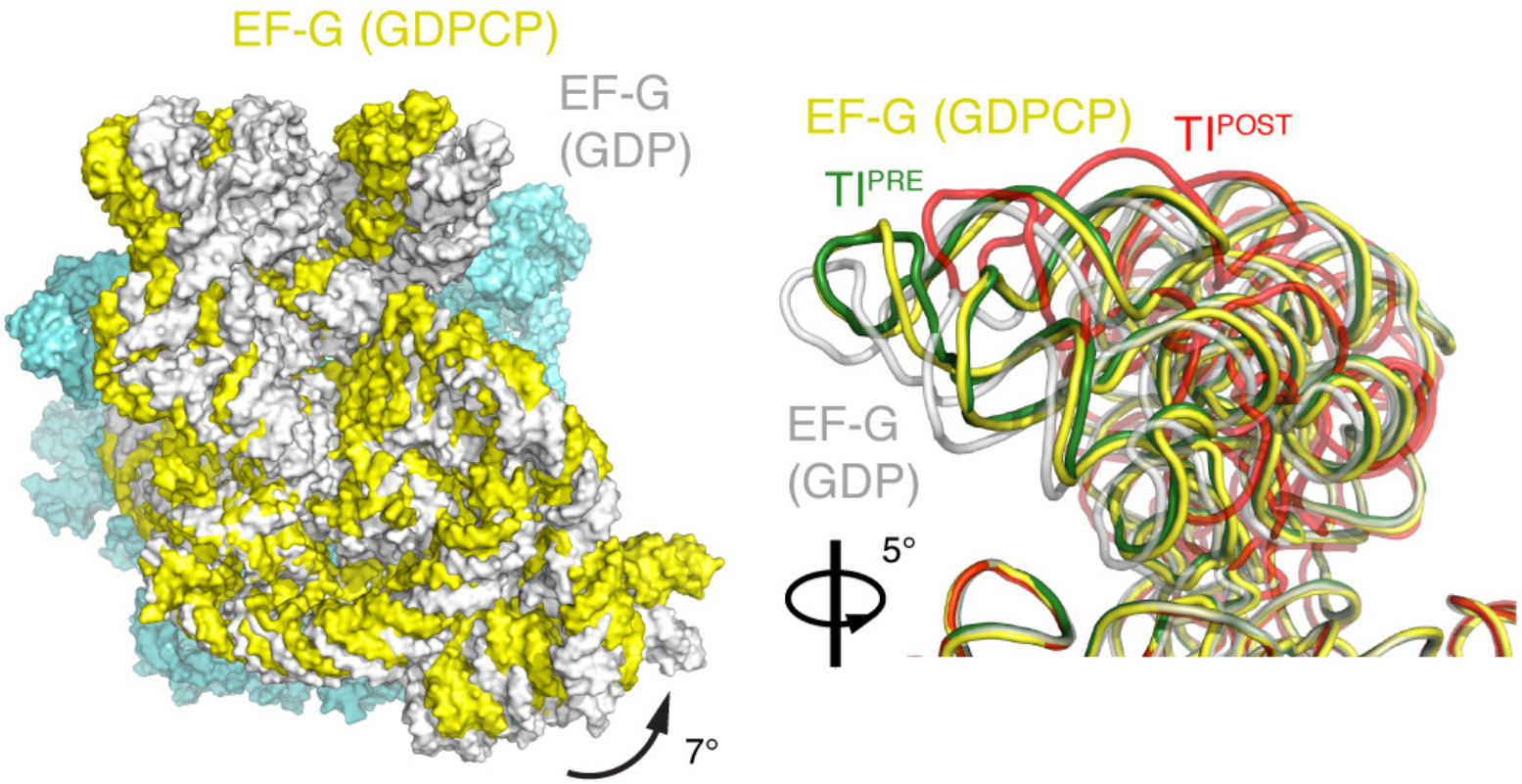


Figure 2.5: Rotation and head swivel of the 30S subunit

Left: Global conformational changes in the 70S ribosome upon GTP hydrolysis as viewed from the perspective of the 23S RNA (cyan) reveals a large rotation of the 30S subunit (yellow). Right: Change in the swivel angle of the head of the 30S in various states of the ribosome, showing the rotated state with EF-G in this study (yellow), the post-translocated state with EF-G and GDP (light gray; pdb code 2WRI) (Gao et al., 2009), the "translational intermediate pre (TIPRE)" state of a cryoEM structure of EF-G with GPDNP bound to a rotated state (green) (Ratje, 2010) and the "translational intermediate post (TIPOST)" state of the same study (red). For clarity, only the ribosomal RNA is displayed.

the P to the E-site. It has been proposed that inter-subunit ratcheting and 30S head swivelling are sequential events that provide directionality to mRNA and tRNA translocation (Guo and Noller, 2012).

2.3.2 Interaction of the L1 stalk with the P/E hybrid tRNA

The tRNAPhe in the P/E conformation is distorted, with a twist in the D-stem of the main body enabling the acceptor arm to swing ∼35° towards the E-site of the 50S subunit, similar to that seen in the hybrid states with RRF (Dunkle et al., 2011) or RF3 (Jin et al., 2011) (r.m.s.d. of 0.8 Å and 1.6 Å respectively). This flexibility allows tRNAs to occupy three very different conformations throughout translation: A/T during decoding, canonical, and hybrid during translocation. Comparative analyses have revealed coevolution between residues of the tRNA D-stem, highlighting in particular seven residues (11-13, 22-24, and 46 in tRNAPhe) that evolve as a single unit (Gutell et al., 1992; Gautheret et al., 1995). With structures of tRNAPhe in all three states available, careful examination of the A/T and canonical tRNAPhe with that in the hybrid state reveals a rigid geometry is retained by these eight residues throughout translation. The vast majority of tRNA nucleotides that are base paired evolve independently of all non-base paired positions and usually only two positions that are base paired evolve as a single unit. On the other hand, the eight nucleotides just described evolve as a complete unit (Gutell et al., 1992; Gautheret et al., 1995) perhaps because this geometry is imperative for tRNA flexibility.

The elbow of the P/E tRNA is cradled by the L1 stalk of the 50S ribosomal subunit, which has pivoted about the base of helix H76 (Figure 2.6) and swung into the fully closed conformation seen in lower resolution studies (Valle et al., 2003; Connell et al., 2007). In structures with a canonical E-site tRNA in

the post-translocational state, the L1 stalk is in a "half-closed" conformation (Gao et al., 2009). Relative to that conformation, the distal part of the L1 stalk has moved inward by ∼25 Å to interact with the P/E tRNA (Figure 2.7), resulting in an angle of ∼17.4° between these two positions. Moreover, there is a distance of ∼37 Å between the closed conformation seen here and the fully open conformation observed in structures of the ribosome with a vacant E-site (Schuwirth et al., 2005). This dynamical nature of the L1 stalk has been studied using two kinds of smFRET experiments and demonstrated to have a mechanistic role during translocation (Fei et al., 2008; Cornish et al., 2009). In the absence of any factor, the L1 stalk fluctuates between half-closed and closed conformations corresponding to non-ratcheted and ratcheted states of the ribosome; binding of EF-G shifts this equilibrium towards the closed conformation of the ratcheted state. The current structure supports the notion that the L1 stalk-tRNA interaction persists throughout translocation (Fei et al., 2008). However, a separate study suggests that hybrid state formation and L1 stalk closure are not tightly coupled (Munro et al., 2010).

A detailed description of the interactions between the L1 protein and tRNA is made possible by the stabilisation of the stalk in the closed conformation, resulting in excellent maps that show side-chain conformations (Figure 2.8). The majority of these interactions are electrostatic, such as Arg59, Arg129 and Arg164 forming salt bridges with the negatively charged phosphate backbone of the tRNA, but there is also a stacking interaction between base C56 and the imino ring of Pro133. Such contacts are probably maintained as the L1 stalk chaperones the P/E tRNA to the E/E conformation during translocation (Fei et al., 2008), since superposing the current structure with that in the post-translocated ribosome structure reveals that the backbone of the L1 protein does not change upon the transition.

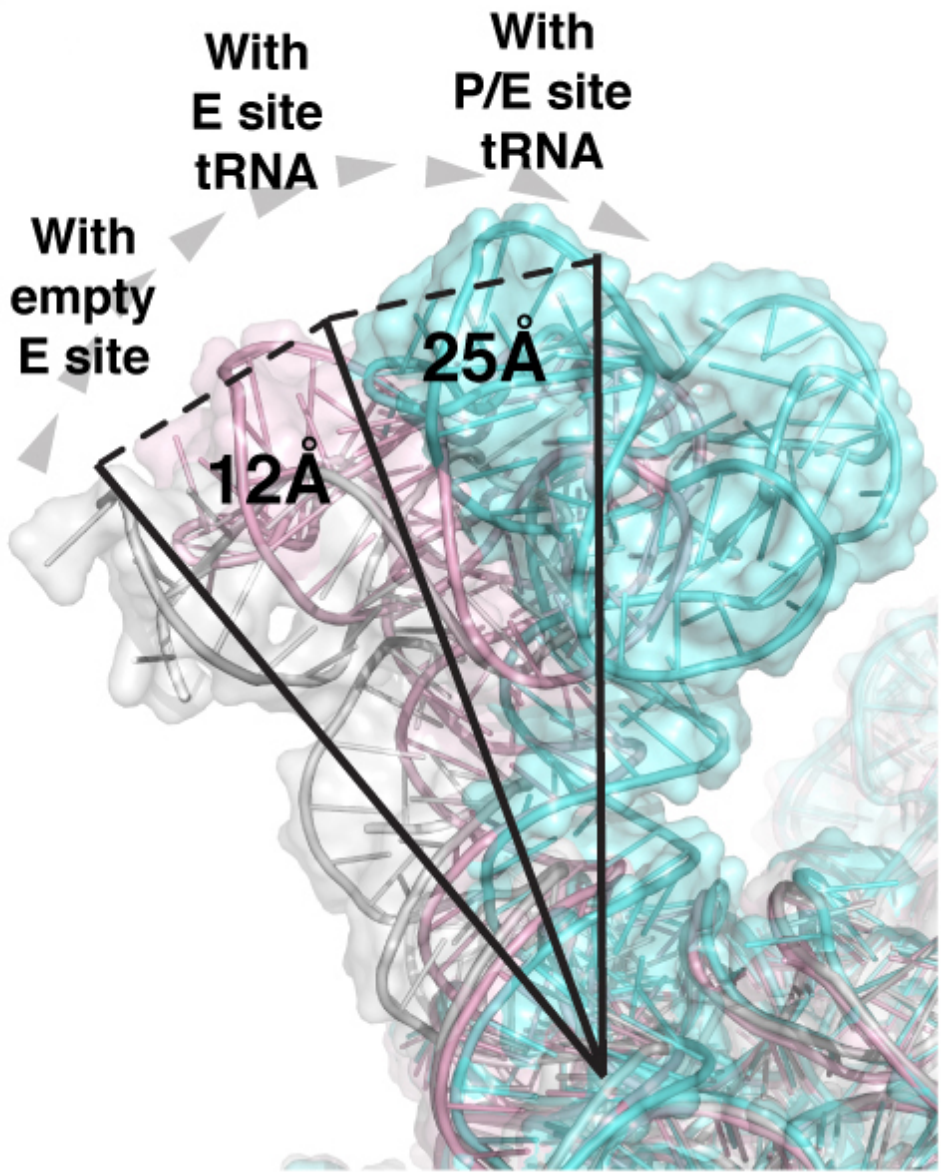


Figure 2.6: Dynamics of the L1 stalk during tRNA translocation

Three distinct conformations of the L1 stalk revealed by superposition of static 23S RNA, showing the open (gray; pdb code 2WA4) (Schuwirth et al., 2005), the half-closed (pink; pdb code 2WRJ) (Gao et al., 2009) and fully closed conformations (cyan; this study). Overlaid lines serve as a guide to the eye and do not represent actual distances.

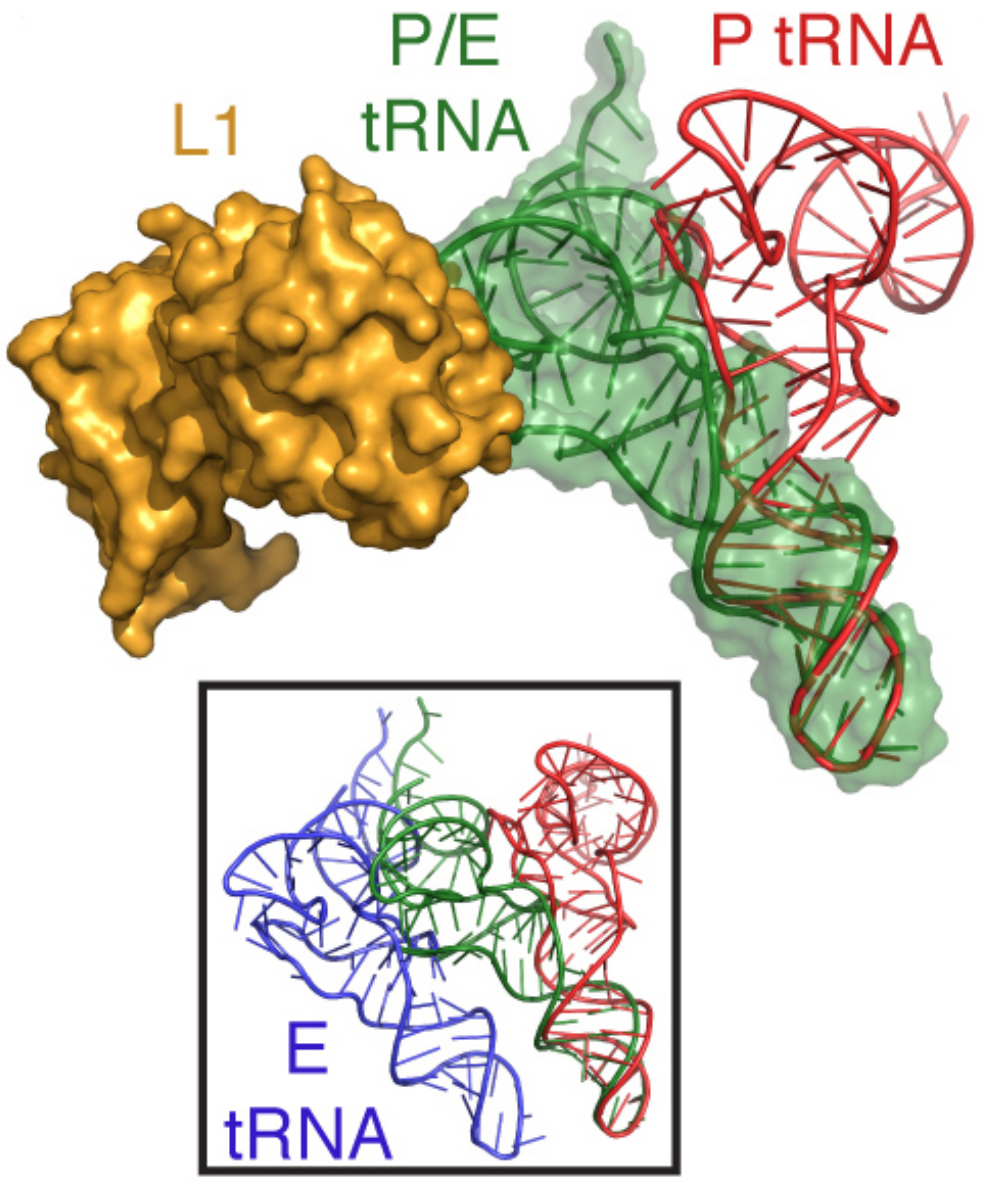


Figure 2.7: Hybrid tRNA conformation

The ribosomal protein L1 (orange) stabilises the tRNAPhe (green) in the distorted P/E hybrid conformation. This lies halfway between the canonical P (red) and E (blue, inset) site conformations.

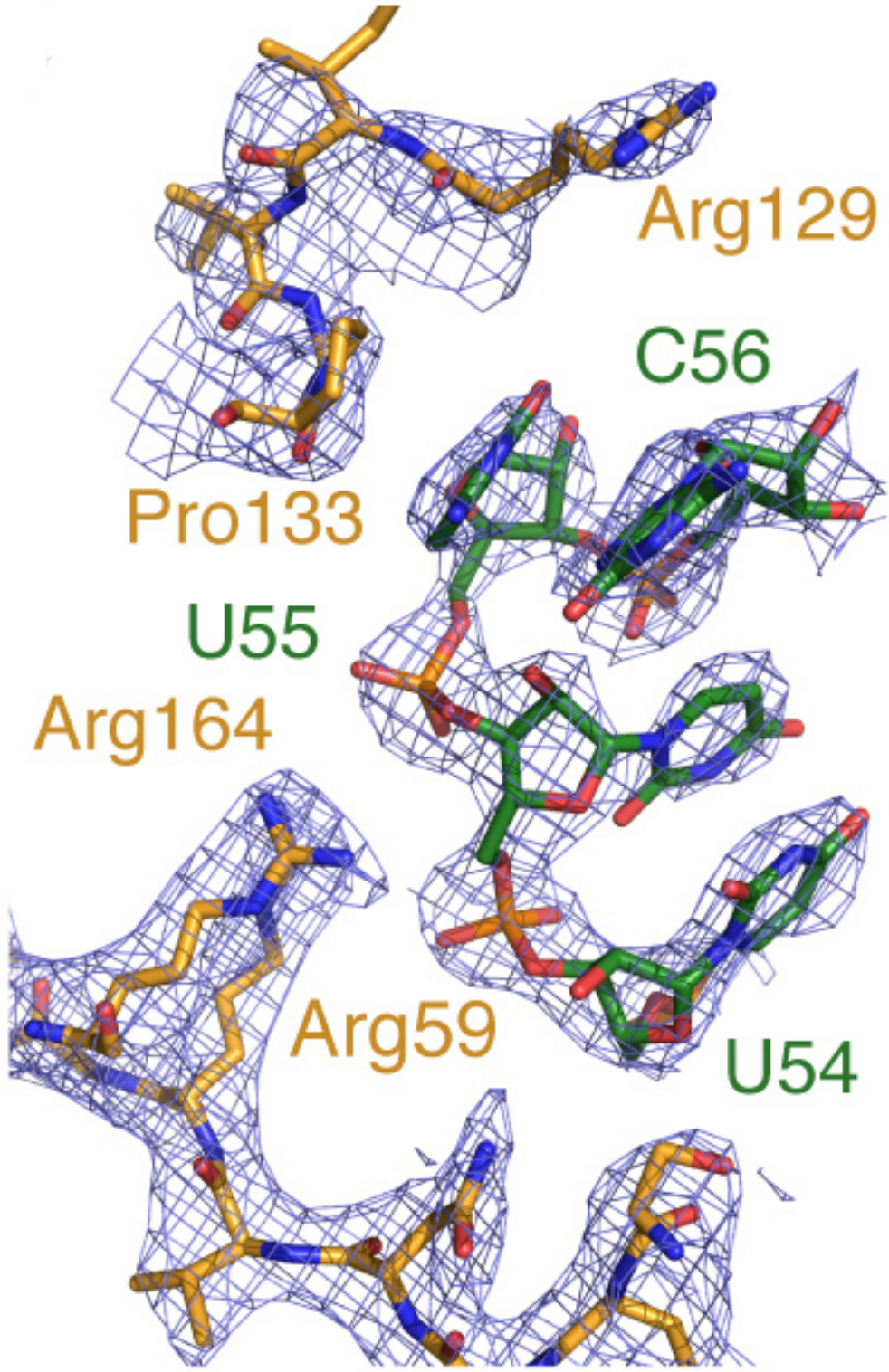


Figure 2.8: L1-tRNA interactions

Details of interactions between the L1 protein (orange) and elbow of the P/E tRNA (green) with labelling of relevant protein and RNA residues. The $F_o - F_c$ difference Fourier map is contoured at 2.5 σand was obtained following refinement with displayed residues omitted from the model.

2.3.3 Interactions of EF-G with L11, L12 and L6

On the other side of the 50S subunit from the L1 stalk, the interaction of EF-G with L6, L11 and the L12 component of the L7/L12 stalk are indistinguishable from those previously described for the post-translocational state (Gao et al., 2009) (Figure 2.9). In particular, the C-terminal domain of one of the L12 molecules is seen interacting with EF-G and the N-terminal domain of L11, and on the opposite side, L6, at the base of the L12 stalk also interacts with EF-G through a flexible C-terminal domain extension. Through these interactions the stalk is thought to play a role in factor recruitment. Based on nuclear magnetic resonance (NMR) data it has been suggested that a hinge region in L7 acts as a molecular switch, initiating closure of the L7/L12 stalk in response to EF-G binding (Bocharov et al., 2004). The antibiotic thiostrepton is known to inhibit EF-G recruitment by disturbing its interaction with protein L11 (Bowen et al., 2005).

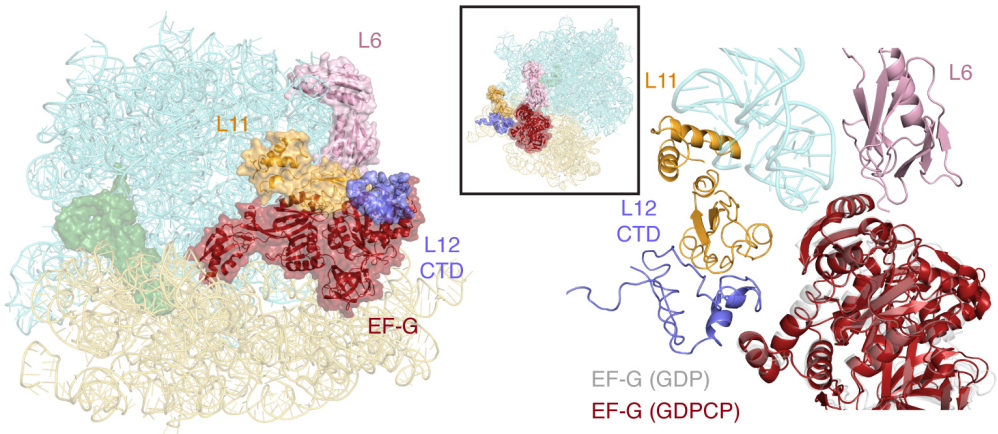


Figure 2.9: Interactions of EF-G with L6, L11 and L12

Interactions of EF-G with ribosomal proteins L11, L6, the L12 CTD near the base of the L7/L12 stalk. A single C-terminal domain of L12 is seen to interact with both EF-G and the N-terminal domain of L11.

2.3.4 Changes in the conformation of domain IV of EF-G

Details of the structural changes in EF-G during translocation can be discerned by superposing domains I and II of EF-G in this structure with those of the isolated factor (Hansson et al., 2005) or in the post-translocational state (Gao et al., 2009). In this superposition, the isolated structure of EF-G would have a conformation of domain IV that would largely avoid a steric clash with A-site tRNA (Figure 2.10 A). Presumably this orientation of domain IV resembles the transient state immediately after EF-G binds to the rotated state and just before translocation occurs in the 30S. In the structure described here, domain IV has moved partly into the A-site and would clash with A-site tRNA (Figure 2.10), which explains why slow translocation can occur even without GTP hydrolysis. Thus ribosome binding alone must promote a conformation of EF-G that partially facilitates translocation. However, the fragmented density and high B-factors for domain IV suggests that it has a dynamical nature, consistent with its requirement for being able to coexist transiently with A-site tRNA.

A comparison with EF-G in the post-translocational state (Gao et al., 2009) shows that the tip of domain IV has moved by another ∼6.6 Å and more fully occupies the A-site (Figure 2.10 B). This further movement is a result of the rotation of the super-domain I-II relative to domains III-V that presumably occurs following GTP hydrolysis. It remains unclear whether domain IV remains outside of the A-site for any extended period of time when EF-G is bound to the ribosome, a state that on the basis of this structure is assumed to be incredibly energetically unfavourable. As revealed in Section 2.4.2 however, this strained conformation has now been visualised at low resolution using cryoEM.

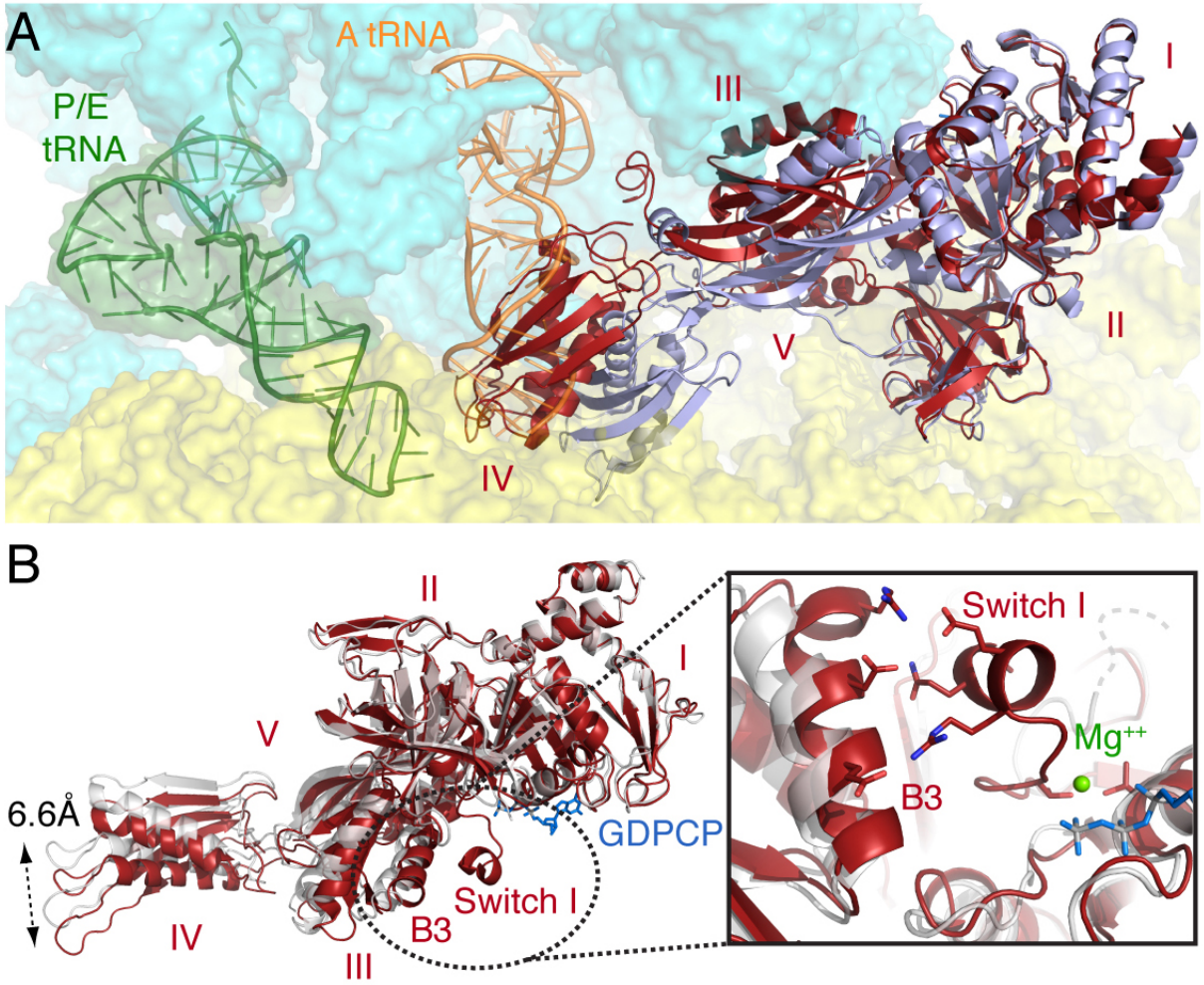


Figure 2.10: Conformational changes in EF-G during translocation

(A) Comparison of isolated EF-G structure (light blue; pdb code 2BV3) (Hansson et al., 2005) with EF-G in this study (red). (B) Comparison of EF-G in this study (red) with that in the post-translocated state (gray; pdb code 2WRI) (Gao et al., 2009) reveals an inter-domain rotation about domain III leading to changes in the orientation of domain IV. Inset (right) shows that in the GDPCP bound EF-G (this study), switch I forms a 3_{10} helix that stabilises helix B3 in an altered conformation.

2.3.5 Changes in the catalytic site

The catalytic site of EF-G shows distinct differences from the post-translocated GDP form (Gao et al., 2009) or the isolated EF-G with guanosine-β,γ-imino-triphosphate (GDPNP) (Hansson et al., 2005) that yield insights into activation of GTP hydrolysis. The switch I region was unresolvable in previous crystal structures of both post-translocated and isolated EF-G, but is ordered in this structure from Met55 onwards. The switch I region (residues 39-66) adopts a single turn of a 3_{10} helix that contacts helix B3 of domain III, as in the isolated structure of the EF-G homologue EF-G-2 in the GTP form, and as also seen at lower resolution by cryoEM studies of a ribosomal complex similar to the structure described here (Connell et al., 2007). The γ-phosphate of GDPCP is surrounded by several highly conserved residues, notably His87 of switch II, and Asp22 and Lys25 in the P loop (Figure 2.11). His87 and Asp22 point away from bound nucleotide in the isolated and post-translocated states of EF-G, but have moved respectively by ∼6.4 Å and ∼3.3 Å (C_α distances) towards the γ-phosphate of GDPCP upon ribosome binding (Figure 2.12) to assume a conformation very similar to that seen before in EF-Tu (Voorhees et al., 2010). As with EF-Tu, the conformation of the activated His87 is stabilised by hydrogen bonding interactions with both A2662 of the SRL, and the catalytic water molecule poised for hydrolysis of the phosphate ester (Figure 2.11). Two Mg^{++} ions positioned by the GAGA tetrad of the SRL stabilise the inward conformation of Asp22 where it coordinates a second water molecule above the γ-phosphate of GDPCP (Figure 2.11). This second water could play a further role in catalysis by donating a hydrogen bond to the γ-phosphate O2. The structure strongly suggests that the change in orientation of Asp22 and His87 upon EF-G binding is part of GTPase activation by the ribosome, and that the mechanism of GTP hydrolysis is essentially the same for both EF-Tu and EF-G.

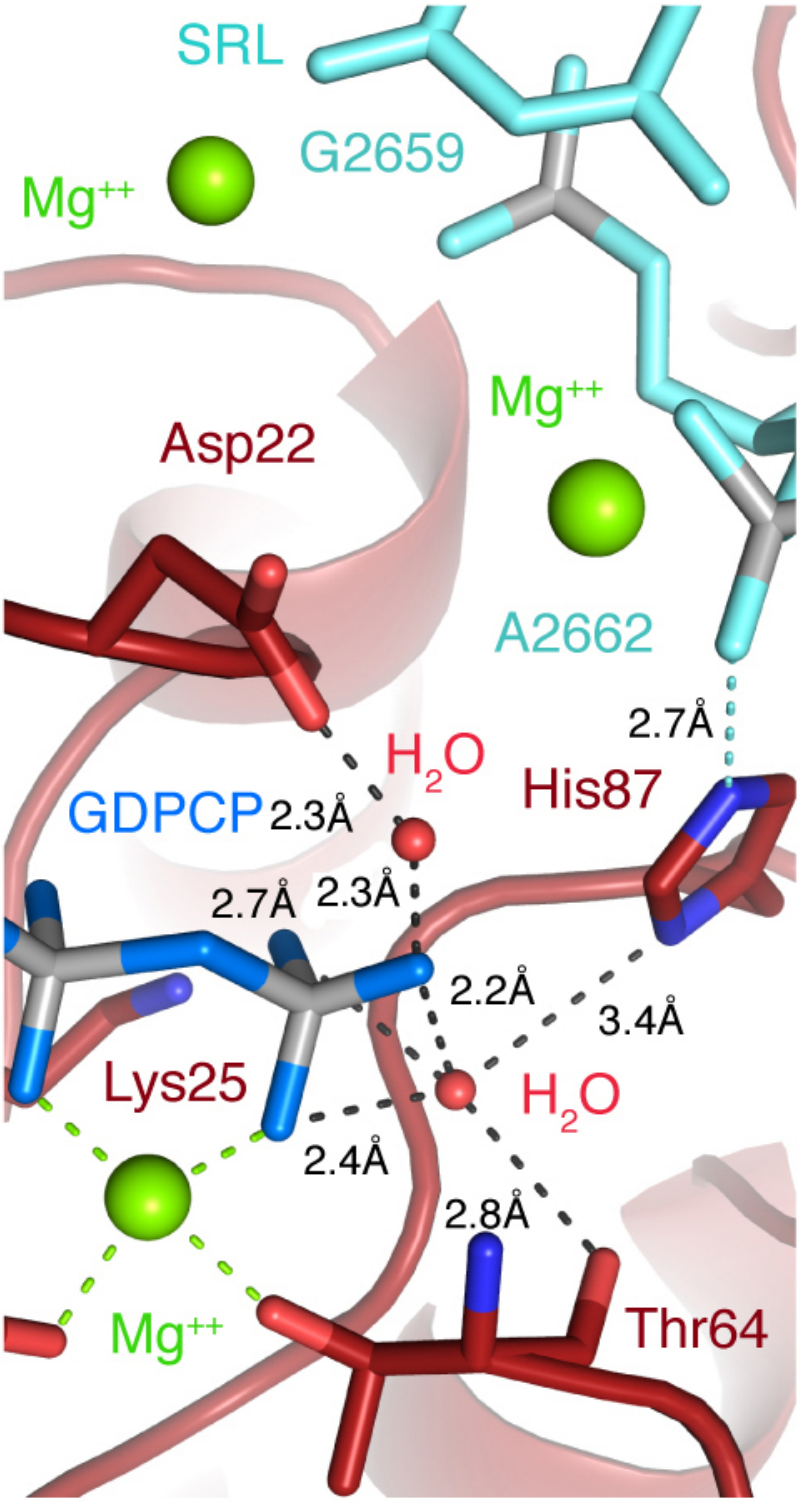


Figure 2.11: The active site of EF-G

Details of the catalytic site around the γ-phosphate of GDPCP (blue) with relevant distances displayed as dashes. EF-G residues and waters are in red, Mg^{++} ions in green, and residues of the SRL are in cyan.

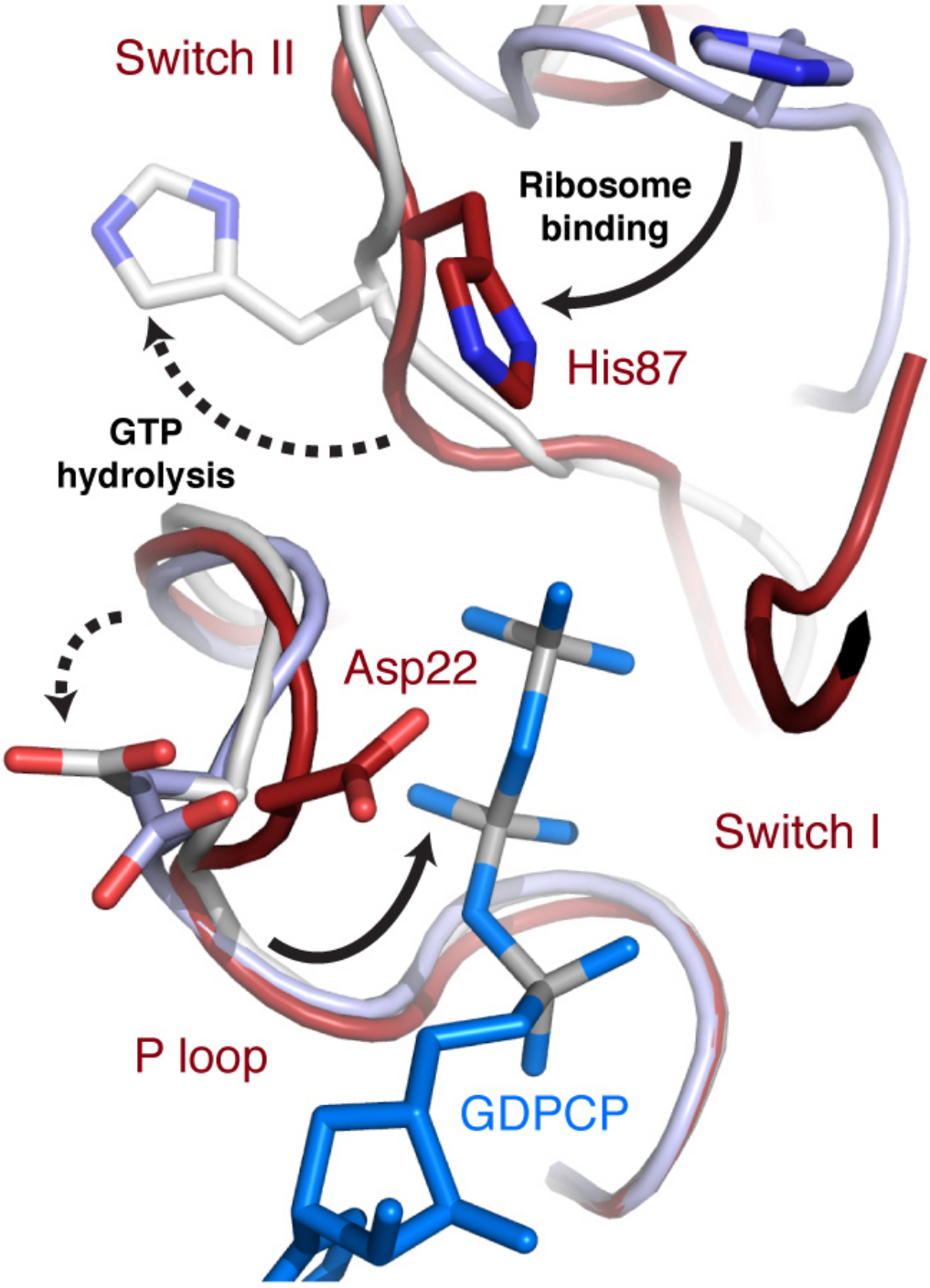


Figure 2.12: Changes in the active site of EF-G during translocation

Comparison of the active site of isolated EF-G with GDPNP (light blue; pdb code 2BV3) (Hansson et al., 2005), EF-G with GDPCP in this structure (red) and EF-G in the GDP post-translocated state (gray; pdb code 2WRI) (Gao et al., 2009) shows that His87 and Asp22 move toward the γ-phosphate of GDPCP on ribosome binding, and away from it upon GTP hydrolysis.

Although the final activated state of EF-Tu and EF-G GTP are highly similar, in EF-Tu the equivalent Asp21 has its activated conformation even in the isolated ternary complex. Thus, different steps may be required to reach the same activated state. Interestingly, the toxin ricin de-purinates A2660 of the GAGA tetrad. It is likely that de-purination of A2660 prevents the surrounding region from adopting the conformation required to bind the metal ions necessary to stabilise Asp22 and neighbouring regions of EF-G in the activated form. EF-Tu does not make these interactions, explaining why ricin only affects EF-G function (Moazed et al., 1988).

While a proposal was made that His87 might be acting as a general base in EF-Tu (Voorhees et al., 2010), the structure is consistent with an alternate mechanism that was proposed subsequently (Liljas et al., 2011). In this mechanism, the negatively-charged environment of the SRL may result in an elevation of the pKa of His87 and stabilise the protonated state of its Nδ, thus enabling His87 to donate a hydrogen bond to the hydrolytic water. The water can in turn donate a hydrogen bond to the carbonyl oxygen of Thr64, and to one of the three oxygen atoms on the γ-phosphate. Under these circumstances, the occurrence of a substrate-promoted catalytic mechanism whereby the γ-phosphate abstracts a proton from the water molecule to generate a hydroxide ion that in turn cleaves the phosphate ester appears feasible. It has also been suggested that the role of the histidine is not to behave as a donor or acceptor of protons at all, but to contribute to an allosteric effect that results in stabilisation of the transition state by the general electrostatic effect of the P loop (Adamczyk and Warshel, 2011). This scenario is compatible with the observation that in EF-G-2, a ribosome-activated GTPase that can substitute for EF-G in polyU-directed protein synthesis in vitro (Connell et al., 2007), the histidine and aspartate have been replaced by tyrosine and glycine respectively

(Figure 2.13).

2.4 Discussion

2.4.1 Implications

The structure reported in this Section provides an atomic model of EF-G bound to the ribosome in a rotated state prior to GTP hydrolysis. It has enabled a complete description of the inward movement of the L1 stalk, stabilisation of the P/E tRNA, and conformational changes in EF-G that are the key steps in facilitating translocation. GTP hydrolysis leads to a series of changes in the switch I, switch II, and P loop regions of EF-G, which result in an inter-domain reorientation about domain III that is expected to promote translocation of any tRNA bound at the ribosomal A-site.

The structure sheds light on the GTPase mechanism of EF-G and on its role in translocation. Globally a striking feature is that the interactions of the L1 stalk with the P/E tRNA appear to be the same as those with the post-translocational E-site tRNA (Gao et al., 2009), implying that the interactions are preserved throughout translocation as previously suggested (Fei et al., 2008). This also suggests that the stabilisation of the closed conformation of the L1 stalk through its interaction with the P/E tRNA is an essential feature of translocation through the stabilisation of hybrid states.

Another large scale movement is the swivelling of the head, which is required to open a constriction that allows passage of the P-site tRNA to the E-site in the 30S subunit (Zhang et al., 2009; Ratje, 2010). It has been suggested before that spectinomycin, an antibiotic that inhibits translocation, may act by inhibiting the movement of the head by binding to a crucial hinge point (Carter et al., 2000; Borovinskaya et al., 2007). The structure shows that in

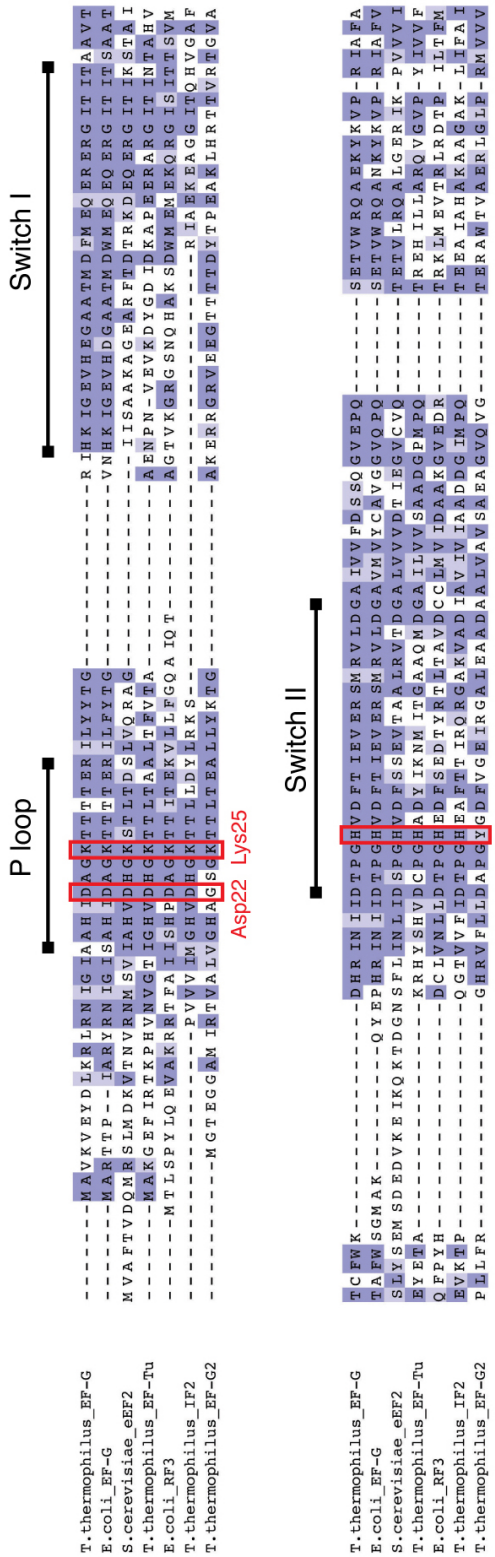


Figure 2.13: EF-G sequence alignment

Sequence alignment of G-domains from several translational GTPases shows conservation of residues Asp22, Lys25 and His87 except in EF-G-2. Regions of the protein discussed in the main text are indicated above.

the rotated state, the swivel angle of the head is such that it would cause a steric clash with spectinomycin, thus supporting this idea.

Remarkably, key residues in EF-Tu and EF-G change conformation in different ways upon binding to very different states of the ribosome to form a nearly identical catalytic site (Figure 2.12), suggesting a common mechanism for activation of translational GTPases by the ribosome. This mechanism also implies that the SRL plays a crucial role in stabilising key residues of the catalytic site in their activated conformations, which would be in keeping with their very high degree of conservation.

Earlier, lethal mutations in the SRL were found not to affect GTP hydrolysis (Shi et al., 2012), suggesting that the SRL does not play a direct role in stabilising the transition state for GTP hydrolysis. However, the interactions with the SRL occurs via phosphate backbone interactions rather than specific bases, so it is possible that in these mutant ribosomes, other nucleotides of the mutated SRL play the role of key residues in the wild-type ribosome.

In contrast to EF-Tu and EF-G, the catalytic site of RF3 on the ribosome appears different; the histidine is far from the γ-phosphate of GTP and makes different interactions with the SRL (Zhou et al., 2012). It is therefore possible that the GTPase mechanism for RF3 is different, or that the structure, which lacks the expected P/E tRNA, does not represent the GTPase-activated state of RF3.

The structure offers some clues into how conformational changes associated with GTP hydrolysis could facilitate translocation. GTP hydrolysis results in changes in switch I, switch II and P loop regions that form an interface between the ribosome, domain III, and GTP. These changes in switch I and II may be communicated to domain III and cause the large movements of the helices that serve to bridge the I-II and III-V super-domains (Figure 2.10).

This would account for the relative change in the orientation of these super-domains upon GTP hydrolysis (Figure 2.10). Deletion of domain III decreases EF-G activity 103-fold on the ribosome (Martemyanov and Gudkov, 2000), supporting the notion that this region may couple GTP hydrolysis to the inter-domain movements that allow domain IV to adopt the favoured conformation of the post-translocational state. Such a conformation may be adopted after tRNA translocation has occurred transiently, allowing domain IV to enter the A-site and prevent a reversal of translocation. Details of the mechanism of action of EF-G will require concerted studies by many complementary techniques.

2.4.2 Recent developments and future directions in the field

After the structure was published (Tourigny et al., 2013a), three separate research groups reported similar results in rapid succession (Pulk and Cate, 2013; Zhou et al., 2013; Chen et al., 2013). The first of these (Pulk and Cate, 2013), reports a pair of novel crystal forms containing *E. coli* ribosomes bound to *E. coli* EF-G in various states of rotation (the asymmetric units of both crystal forms each contain four ribosomes). Each ribosome is in a canonical, partial, or fully rotated state, the latter conformation being almost identical to that of the Tourigny *et al*. structure. In every case EF-G is bound with GDPCP in the absence of tRNA. The catalytic site, together with re-ordering of the switch regions, is identical to that described previously and so this result is in good agreement with the present study. The binding of EF-G in various states of rotation raises the interesting possibility that GTP hydrolysis is not necessarily a prerequisite of translocation and that the additional structures represent true conformational changes in EF-G that occur during translocation. EF-G only partially occupies the canonical ribosome however, suggesting

that this state is disfavoured by EF-G prior to GTP hydrolysis and may only appear as the result of unphysiological conditions in the crystal. Moreover, the absence of any tRNA (that would almost surely help stabilise the rotation of ribosomal subunits in one state or the other) means that the free energy barrier is perhaps increased further in vivo to prevent GTP-bound EF-G from binding a post-translocated ribosome.

The second study also reports a crystal structure of the *T. Thermophilus* ribosome bound to *T. Thermophilus* EF-G (Zhou et al., 2013), although the orientation of EF-G and inter-subunit rotation is distinct from that observed by Tourigny *et al*. and Pulk & Cate. The structure contains a tRNA in what the authors call the pe*/E state, an intermediate between the P/E and E state, and the degree of inter-subunit rotation is comparable to that previously referred to as the TIPOST state (Ratje, 2010). This suggests the structure represents an intermediate between the Tourigny *et al.* structure and the post-translocated state. At the GTPase centre of EF-G His87 is located too far from the catalytic site to coordinate any water near the γ-phosphate of GTP, as the same group reported in the case of RF3 (Zhou et al., 2012). Site-directed mutagenesis experiments have cemented the role of His87 in GTP hydrolysis (Cunha et al., 2013), and so explaining this residue's novel conformation in the TIPOST state makes it difficult to argue that this structure represents the activated state of translocation. More likely is that it appears as a transient state occurring after GTP hydrolysis during reverse rotation of the ribosome. The final structure, reported by Chen et al. (Chen et al., 2013), was initially claimed to be different from the Tourigny *et al*. structure. Since then however, the two have been found identical to within coordinate error.

The structures have assisted with some of the controversy in the GTPase field. Warshel and colleagues used the atomic coordinates to interpret ab

initio quantum mechanical/molecular mechanical calculations simulating GTPase activation (Plotnikov et al., 2013). Interestingly, the results of that study indicate the most likely reaction mechanism involves a concerted water attack of the P-O bond with a two-water proton transfer at the transition state. These findings support the observation of two water molecules near the γ-phosphate of GTP in the Tourigny *et al*. structure. As suggested in Section 2.3.5, in this mechanism His87 does not play the direct role of a proton shuttle but provides an allosteric contribution through donation of a hydrogen bond.

Two cryoEM structures of EF-G bound to the ribosome appeared shortly after the four high-resolution crystal structures. These contain two tRNAs and are supposed to represent the true TIPRE (Brilot et al., 2013) and TIPOST (Ramrath et al., 2013) states of translocation. Due to the energetics of the intermediates, these structures represent only a small subset of the cryoEM samples and so resolutions of the reconstructions are comparably low. However, the global conformations of the ribosome, EF-G and the two tRNAs are distinguishable and agree well with what would be expected from the crystal structures and prior knowledge of the pre-translocated ribosome. A particularly prominent feature of the EF-G-tRNA-tRNA-TIPRE state (Brilot et al., 2013) is the conformation that EF-G is forced to assume outside of the 30S. Domain IV is positioned in the cleft between the 30S and 50S subunits to avoid a steric clash with the A/P hybrid state tRNA, and relaxation of this strained conformation is presumably relaxed upon the GTP hydrolysis event that initiates translocation. Now that an apparent revolution is underway in ribosome cryoEM (Bai et al., 2013; Fernández et al., 2013a; Amunts et al., 2014; Kühlbrandt, 2014), obtaining the true TIPRE state at atomic resolution would be useful to prove conclusively that the catalytic site of EF-G described here is unchanged in the presence of an A/P tRNA.

Chapter 3

Interfering with translocation

3.1 Antibiotics and the ribosome

A wide variety of natural products exhibit toxicity because they target the translational machinery of the cell. Of these compounds, antibiotics that bind selectively to bacterial or protozoal ribosomes are of great clinical significance due to their ability to treat infectious diseases without compromising the host (Poehlsgaard and Douthwaite, 2005; Yonath, 2005). The most effective antibiotics used in clinical treatment exploit subtle differences between pathogen and host ribosomes that can be found at distinct locations within the functional sites. In most cases the target is therefore ribosomal RNA rather than proteins (Gale et al., 1987).

Each functional site of the ribosome targeted by an antibiotic represents a key step in the translational pathway. Antibiotics are known to interfere with decoding, peptidyl-transfer, polypeptide chain elongation, translocation, and also inhibit conformational changes in the ribosome that are required during translation (Yonath, 2005). For example, aminoglycosides such as paromomycin are understood to interfere with decoding and translocation by forming a

tight interaction with the major groove of helix 44 on the 30S subunit. Binding of the antibiotic induces a conformational change in residues A1492 and A1493, which flip out towards the A-site codon to stabilise any tRNA bound at this location (Karimi and Ehrenberg, 1994; Pape et al., 2000). Furthermore, antibiotics can target the protein factors that assist with translation, such as EF-G, which is inhibited by fusidic acid (Bodley et al., 1970a,b), and EF-Tu, which is inhibited by kirromycin (Wolf et al., 1974).

Initial crystal structures of the ribosomal subunits at atomic resolution were followed almost immediately by their structures in complex with different antibiotics (Carter et al., 2000; Brodersen et al., 2000; Pioletti et al., 2001; Schlünzen et al., 2001; Hansen et al., 2002a). This has continued for over a decade and crystallography still serves as the most powerful method for revealing the molecular mechanisms of these compounds by putting biochemical results into perspective. As well as leading to a better understanding of antibiotic selectivity, in some cases high-resolution structures can even explain how site-specific mutations lead to emergence of antibiotic resistance in pathogenic populations. Since X-ray crystallography provides an atomic description of a ligand and its binding site, these structures form a basis for the design of new compounds to help in the fight against resistance.

3.2 How pactamycin analogues interfere with translocation

3.2.1 Pactamycin and its analogues

The aminocyclopentitol pactamycin (Figure 3.1) was first isolated from *Streptomyces pactum* as a potential anti-tumour drug and later found to exhibit

potent activity against many bacteria, archaea, and eukaryotes (Bhuyan et al., 1962; Mankin, 1997). Pactamycin consists of two aromatic rings (6-methylsalicylic acid (6-MSA) and 3-aminoacetophenone) attached to a five-membered ring aminocyclitol unit, which is also linked to a 1,1-dimethylurea moiety (Wiley et al., 1970; Rinehart Jr et al., 1980). Despite initial claims that pactamycin could be used as an anti-tumour drug (Bhuyan et al., 1962), high levels of toxicity displayed towards human cells made it clear this compound was unsuitable for clinical use. A typical approach to this problem is to modify the structure in an attempt to alter the pharmaceutical properties of the drug, but conventional synthetic chemistry proved difficult due to the complexity of pactamycin. Only recently has a complete enantioselective synthesis of pactamycin totaling only 15 steps been possible, and this will augur well for newer analogs (Malinowski et al., 2013; Codelli and Reisman, 2013).

Mahmud and colleagues have now elucidated the biosynthetic pathway of pactamycin (Ito et al., 2009). Briefly, over 50 open reading frames in *S. pactum* DNA are believed to be involved in biosynthesis, with 26 genes (ptmA-ptmZ) forming the core cluster directly involved in the chemical process. Additional genes are thought to transcriptionally regulate the pathway or participate in resistance mechanisms (such as *orf14* and *orf15* that are homologous to translation initiation factor IF-2). Each enzyme *ptmA-ptmZ* has a role in the pathway that has been shown to proceed via an intermediate compound de-6-MSA-pactamycin (Figure 3.1). This compound lacks the 6-MSA ring of the parent molecule and yet displays equivalent antibacterial and anti-tumour activity to pactamycin suggesting that the 6-MSA moiety is not required for cell toxicity (Ito et al., 2009).

Biosynthetic products related to de-6-MSA-pactamycin also inhibit growth of malarial parasites, but with a significant reduction in toxicity to mammalian

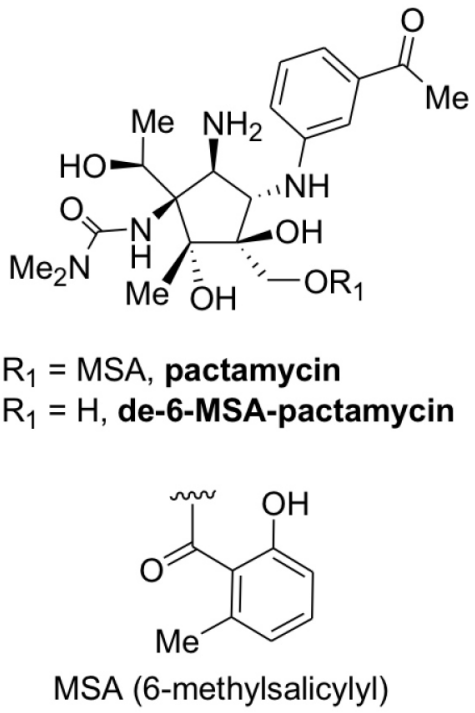


Figure 3.1: Chemical structure of pactamycin and de-6-MSA-pactamycin

Shared structure of pactamycin and its analogue de-6-MSA-pactamycin displayed above, with the chemical group R_1 denoting either a single hydrogen (de-6-MSA-pactamycin), or the 6-MSA ring in the case of the native pactamycin.

cells (Otoguro et al., 2010; Lu et al., 2011). Likewise, semi-synthetic analogs of de-6-MSA-pactamycin, prepared following the first total synthesis of pactamycin (Hanessian et al., 2011) and varying in the nature of the urea or the aniline moieties, exhibit potent in vitro anti-parasitic and anti-tumour activity (Hanessian et al., 2013). As an alternative to the expensive option of synthesising pactamycin analogues in the laboratory, purification and modification of naturally occurring derivatives is an attractive option for the pharmaceutical industry. To select the compounds correctly however, there is a need to un-

derstand the molecular mechanism by which these drugs are able to interfere with translation.

In accordance with biochemical data (Egebjerg and Garrett, 1991; Woodcock et al., 1991), the crystal structure of pactamycin bound to the 30S ribosomal subunit revealed that this antibiotic binds near a highly conserved region of 16S RNA at what is now know to be ribosomal E-site (Brodersen et al., 2000). In the native crystal form of the *T. Thermophilus* 30S subunit, the 3' end of 16S RNA folds round to mimic a genetic message and binds in the mRNA binding cleft (Carter et al., 2000). Binding of pactamycin distorts the RNA at the E-site, causing it to be pushed up towards the back of the 30S and inducing a $\sim$12.5 Å displacement for the last base in the E-site codon. On this basis it was later proposed that pactamycin prevents a codon-anticodon interaction forming at this location, and blocks the translocation of P-site tRNA into the E-site of the 30S (Dinos et al., 2004). This conclusion was supported by the observation that pactamycin inhibits poly(A)-dependent poly(Lys) synthesis, but does not have an effect on initiation as was previously supposed.

Knowing that de-6-MSA-pactamycin maintains its in vitro antibacterial, anti-tumour, and anti-parasitic activities, it was of particular interest to see how the absence of the 6-MSA acid moiety affects its binding to (and inhibition of) the ribosome. Moreover, an atomic structure of de-6-MSA-pactamycin bound to the ribosome would pave the way for the therapeutic development of related compounds.

3.2.2 Crystal structure of de-6-MSA-pactamycin bound to the 30S ribosomal subunit

Determination of the crystal structure of the *T. thermophilus* 30S ribosomal subunit bound to de-6-MSA-pactamycin in the presence of paromomycin has

enabled a detailed description of interactions between pactamycin analogs and the ribosome (Tourigny et al., 2013b). Full material and methods are in Appendix A. Hanessian *et al*. recently reported a complete synthesis of de-6-MSA-pactamycin (Hanessian et al., 2013), and subsequently the compound was purified and concentrated to a level suitable for crystallography. Solubility of the analogue posed a serious problem, but the 2-Methyl-2,4-pentanediol contained within the cryoprotectant proved a convenient solute for suspending the substrate (Appendix A.3). The cryoprotection solution also included the antibiotic paromomycin since this is known to induce stability of the 30S and increase the quality of diffraction (Carter et al., 2000). Data to beyond 3.1 Å were collected on the IO4 beam line at the Diamond Light Source.

Following refinement of the initial atomic model, de-6-MSA-pactamycin was unambiguously placed into electron density identified at the tip of helix 23b (Figure 3.2). This location has previously been described as the binding site of pactamycin (Brodersen et al., 2000). The two distal aromatic rings of pactamycin are known to stack against each other and G693 of helix 23b due to the antibiotic adopting a folded structure mimicking an RNA dinucleotide. This was suggested to result in a displacement of the E-site mRNA. Similarly, the remaining aminoacetophenone moiety of de-6-MSA-pactamycin stacks against the base of G693, where it is stabilized by O6 and N7 forming hydrogen bonds with an amine and ketone on the neighbouring cyclopentitol.

A superposition of this structure with the empty 30S subunit reveals that like pactamycin (Brodersen et al., 2000), de-6-MSA-pactamycin prevents the 3'-end of 16S RNA from folding back on itself to mimic an E-site codon. However, the absence of a 6-methylsalicylic acid moiety on de-6-MSA-pactamycin means the 3' end of the 16S, and presumably the path of mRNA, is displaced to a lesser extent than it would be in the presence of pactamycin ($\sim$8.0 Å

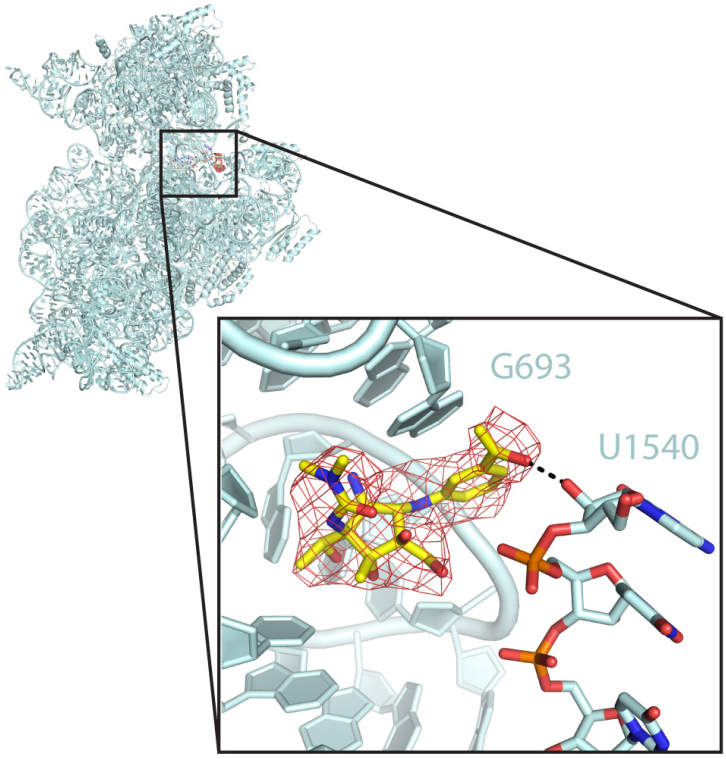


Figure 3.2: The de-6-MSA-pactamycin binding site

Location of the de-6-MSA-pactamycin (yellow) binding site on the 30S ribosomal subunit (cyan) with omit $F_o - F_c$ difference map is contoured at 3 σ.

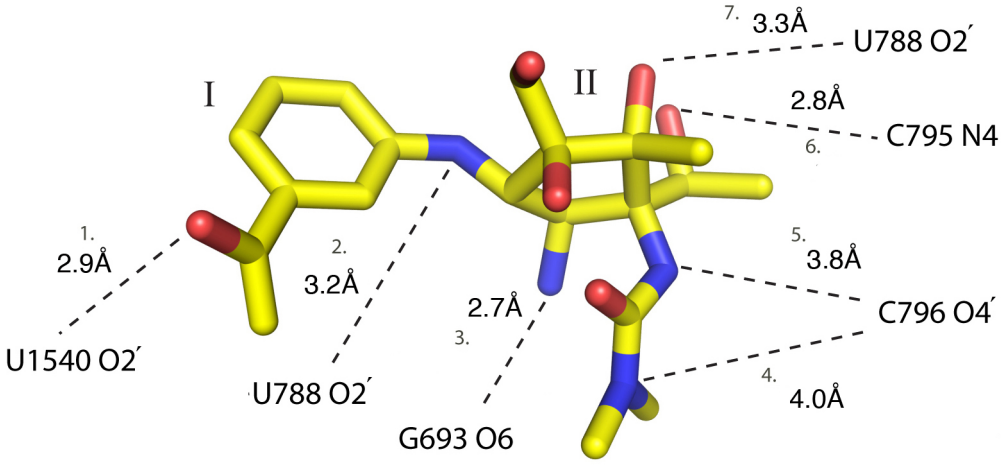


Figure 3.3: Atomic distances between de-6-MSA-pactamycin and the 30S

compared to ∼12.5 Å). This allows base U1540 of the 16S to form a novel hydrogen bond via its O2 and the carbonyl group of the aminoacetophenone ring (interaction 1, Figure 3.3, *E. coli* numbering). Interestingly, replacement of the acetyl group in the aniline moiety of de-6-MSA-pactamycin by fluorine or trifluoromethyl, results in potent in vitro antimalarial activity (Hanessian et al., 2013). It is likely that a hydrogen bond is shared between fluorine and U1540 when such compounds form a complex with the ribosome.

Further hydrogen bond interactions were identified between bases G693 and C796, and functional groups on the extensions of the central ring (Figure 3.3, interactions 2-6, *E. coli* numbering). The N4 of base C795 forms a hydrogen bond with the hydroxyl group on the C7 cyclopentitol atom (interaction 6, Figure 3.3, *E. coli* numbering), which is absent in the antimalarial analog de-6-MSA-7-deoxypactamycin. It would therefore appear that a loss of this hydrogen bond is sufficient to reduce binding of de-6-MSA-7-deoxypactamycin to the mammalian ribosome, enough to lower cell toxicity 10 - 30 fold (Lu et al., 2011). Together, these interactions mean that de-6-MSA-pactamycin forms a tightly bound complex with the ribosome that disrupts base pairing at the E-site of the 30S subunit.

A total of six previously unreported binding sites for paromomycin were also identified in the $F_o - F_c$ difference map of the de-6-MSA-pactamycin complex in addition to the site described in Section 3.1 (data not shown). These sites are distributed non-specifically throughout the 30S, but in each case the antibiotic does not appear to induce any significant conformational change in the surrounding nucleotides. A comparison of published and unpublished structures containing paromomycin and an A-site anticodon stem loop reveals that the presence of a tRNA abolishes binding at every one of these additional sites. However, density of reasonable quality was confirmed at these locations

upon re-examination of the lower resolution structure depicting paromomycin bound to the empty 30S (Carter et al., 2000). Additional interactions between paromomycin and the isolated 30S have been suggested by foot-printing experiments where, interestingly, protection of so-called 'class-III nucleotides' at these sites was reduced in the presence of tRNA (Moazed and Noller, 1987). Despite this, no direct contacts are made between additional paromomycin molecules and the class-III nucleotides. It is likely that the artificially high concentration of paromomycin used for crystallisation causes the antibiotic to bind non-specifically at charged sites throughout the 30S, but it remains unclear why tRNA binding should affect this.

3.3 Discussion

3.3.1 Implications

Although de-6-MSA-pactamycin shares the same binding site as pactamycin, a new collection of antibiotic-ribosome contacts distinguishes this derivative from its parent molecule. A complete understanding of such interactions will aid in the design of new and improved analogs toward the development of effective anti-protozoal and anti-tumour drugs.

3.3.2 Recent developments and future directions in the field

Shortly after completion of this work, a full account of the asymmetric synthesis of pactamycin was published (Sharpe et al., 2013). Total synthesis requires only 15 steps and results in 1.9% overall yield from commodity chemicals and the authors state that the preparation of analogues is underway.

Malaria research has benefitted from recent advancements in cryoEM since

the cytoplasmic ribosome of the disease causing *Plasmodium falciparum* has been determined at 3.2 Å resolution (Wong et al., 2014). The structure has been solved in the presence of the antibiotic emetine, whose mode of binding resembles that of de-6-MSA-pactamycin. In particular, like de-6-MSA-pactamycin, emetine stacks against the parasitic counterpart of G693, but lacks a chemical group needed to fill the space that would be occupied by the 6-MSA ring of pactamycin. The *P. falciparum* structure has been solved in the absence of mRNA, but if present the E-site codon would be displaced by ∼8.0 Å assuming the mode of action were similar to de-6-MSA-pactamycin. Emetine also exhibits potent antimalarial activity against the blood stage of P. falciparum (Matthews et al., 2013), raising the intriguing possibility that this pathogen is particularly sensitive to molecules that localise to the de-6-MSA-pactamycin binding-site. Although emetine remains toxic to human cells these results provide an encouraging framework for the optimisation of compounds to target this site.

Chapter 4

Concluding remarks

Our understanding of translocation has culminated with over 50 years of biochemical and single molecule techniques being used to interpret several recent structures of the EF-G bound to the ribosome in various states of translocation (Gao et al., 2009; Tourigny et al., 2013a; Pulk and Cate, 2013; Zhou et al., 2013; Chen et al., 2013). Together, these have helped to reveal the molecular mechanism by which EF-G is able to control the rotation of 30S/50S subunits and induce conformational changes that accelerate and assist translocation. Moreover, comparison of the model by Tourigny *et al*. (Tourigny et al., 2013a) with a structure of the activated ternary complex bound to the ribosome (Voorhees et al., 2010) has demonstrated that the mechanism of GTP hydrolysis is the same for EF-G and EF-Tu. With this knowledge, scientists in the pharmaceutical industry can develop novel compounds to specifically target the translocational pathway of pathogenic parasites and bacteria. A promising lead compound is de-6-MSA-pactamycin, which has increased potency towards the protozoa ribosome and its atomic interactions with the small subunit are already well documented.

Throughout this Section a consistent theme has been the opportunity of

taking advantage of recent cryoEM developments to study ribosomes at crystallographic resolution. With the quality of several cryoEM structures now matching those obtained by crystallography, X-ray methods are at a point of getting superseded by microscopy. Since single particle reconstructions rely only on nanomolar concentrations of heterogeneous sample, cryoEM techniques are ideally suited to high-resolution structures like that of EF-G bound to the ribosome simultaneously with an A/P hybrid tRNA. Rather coincidently, atomic resolution cryoEM structures began to appear only after the prokaryotic 70S and eukaryotic 80S had been solved by crystallography, but it is already clear that maps are of high enough quality to have been interpreted without atomic models pre-available. It remains to be seen whether cryoEM can be used to yield atomic resolution structures of biological complexes less stable and robust than the ribosome.

Part II

Multi-crystal data processing

Chapter 5

Introduction

5.1 General introduction to crystallography

On 4 May 1912, Arnold Sommerfeld presented a one-page report to the Bavarian Academy of Sciences stating that on 21 April of that year, Walter Friedrich, Paul Knipping, and Max Laue had observed diffraction of X-rays by a zinc sulphide (ZnS) crystal. "*The guiding idea was that interferences arise in consequence of the lattice structure of the crystals, because the lattice constants are ca. 10 x greater than the conjectured wavelengths of the X-rays*" (Forman, 1969). The trio published their interpretation of the results in an 18-page paper that was reprinted a year later (Friedrich et al., 1913), assuming the X-ray source they used had been monochromatic. Conversely, William Lawrence Bragg (with the help of his father William Henry Bragg) interpreted Laue's results assuming reflection of a continuous range of X-rays by planes of atoms in the crystal, with the distance between planes related to the diffraction pattern by a simple formula now known as Bragg's law (Bragg, 1912). This led, not only to a correct structure of the ZnS crystal, but also a pioneering trend of using X-ray crystallography to solve the structures of minerals and metallic com-

pounds (Thomas, 2012). In 1938, W.L. Bragg succeeded Ernest Rutherford as Cavendish professor at the University of Cambridge, encouraging Max Perutz and John Kendrew to begin work on the crystal structures of the proteins haemoglobin and myoglobin.

The next few Sections will serve as an introduction to the theory that, since the Braggs' work in 1912, has led to X-ray crystallography becoming arguably the most powerful tool of modern-day molecular biology. There are many comprehensive texts on the subject (Blundell and Johnson, 1976; Drenth, 2007; Blow, 2002; Lattman and Loll, 2008; Rhodes, 2010; McPherson, 2011; Sheldrick et al., 2001; Rupp, 2010), but this will be a self-contained introduction to macromolecular crystallography with a certain degree of originality.

5.1.1 Crystal geometry and X-ray diffraction

A crystal is defined to be a solid material whose constitute atoms or molecules form a precisely ordered array in three-dimensional space. It is evidently a little ambitious to use the term "precisely ordered" when referring to a physical object, particularly an array of macromolecules, but in practice the molecules of a crystal are arranged in near-perfect periodic array. This is in contrast to an amorphous solid such as glass, whose atoms or molecules do not display periodicity on a long-range scale. The arrangement of molecules within a crystal is characterised by a set of determining parameters. The collection of atoms making up the smallest repeating unit is called the asymmetric unit because no symmetry operations can reproduce the crystal from any of its proper subsets, and the entire crystal structure is determined by the contents of the asymmetric unit combined with a set of symmetry operations. Macromolecular crystals typically contain an integer number of the constitute molecule in the asymmetric unit unless it happens to be a complex containing an internal

symmetry that coincides with the symmetry of the crystal.

In three-dimensional Euclidean space there are 230 distinct symmetries of a periodic configuration, 219 if chiral copies are to be excluded. These are referred to as the 219 space groups or Fedorov groups (Fedorov, 1891), but only 65 represent the possible symmetries of a macromolecular crystal. This is because the 65 possible Sohncke space groups do not contain reflections or inversions as a symmetry operation, which cannot be reproduced by chiral molecules like proteins and nucleic acids. There are at least eight ways of classifying and naming space groups, but this thesis will use the nomenclature introduced by the International Union of Crystallography (Hahn et al., 2005). The international notation consists of up to four symbols $XMNL$ where X is a letter (P,A,B,C,I,R or F) that describes the centring of the Bravais lattice. This is an infinite array of points generated by the translation operators of the crystal and there are 14 different lattices compatible with the 230 space groups. The three numbers MNL define the point group of rotations as viewed down the axis of highest symmetry and also incorporate the degree of rotation of a screw axis. For example, 2_1 is a two-fold rotation combined with a translation of half the lattice vector (a two-fold screw axis), and 3_12 is a three-fold screw axis and a two-fold rotation axis. Consequently, the space group $P2_1$ exhibits primitive (P) centring in the Bravais lattice and a two-fold screw axis down one lattice vector, and $P2_12_12_1$ exhibits primitive centring and a two-fold screw axis down each of the three lattice vectors. Despite an extensive list of possible symmetries, macromolecules display a strong preference for certain space groups over others (Wukovitz and Yeates, 1995).

The components of the so-called unit cell are generally chosen to be the contents of the smallest parallelepiped whose edges coincide with the symmetry elements relating asymmetric units, so that atomic coordinates of the

asymmetric unit can be represented in terms of lattice parameters. The entire crystal structure is then described by the convolution of the contents of the unit cell with the Bravais lattice, which is conveniently represented as a family of two-dimensional planes passing through the crystal. This abstraction was introduced by Bragg, who realised diffraction of X-rays by a lattice was almost equivalent to reflection of the beam from a set of planes through the same lattice points (Bragg, 1912). There are an infinite number of possible sets of planes, labeled by three integers (h,k,l) called the Miller indices, which denote the number of times a particular set of planes intersects one of the three translation operators of the lattice. The distance between neighbouring planes of one particular set decreases as the Miller indices increase, and so it is only for certain values that reflected X-rays constructively interfere. The condition for constructive interference is provided by Bragg's law, which states that twice the product of inter planar spacing d and sine of the angle of incidence θ must be an integer multiple of the wavelength λ:

$$n\lambda = 2d\sin\theta \ , \quad \text{integer } n$$

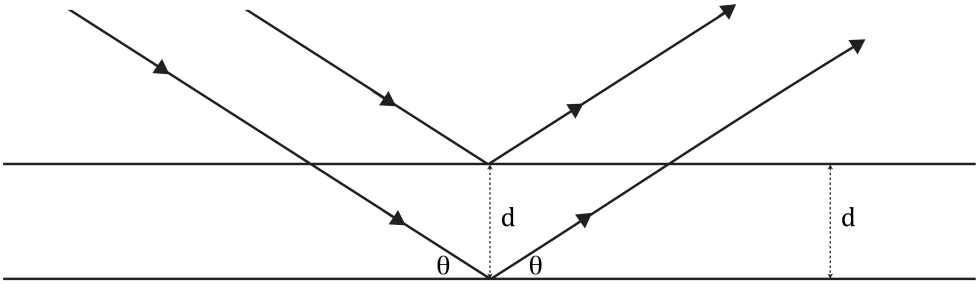


Figure 5.1: Bragg's law

X-rays making an angle of incidence θ with planes separated by distance d passing through a crystal. Bragg' s law gives the condition for constructive interference.

5.1.2 Fourier theory and the phase problem

The set of Miller planes passing through a crystal is an entirely non-physical construction that helps with the interpretation of a diffraction pattern. In reality, X-ray diffraction is a form of elastic scattering from the distribution of electrons in the crystal and the diffraction pattern is proportional to the modulus squared of the scattered quantum mechanical wave function. Mathematically inclined text books on macromolecular crystallography use a classical description to argue the scattered wave is related to the Fourier transform of the crystal lattice and the electron density distribution of unit cell by the convolution theorem. Whilst sufficient for all intended purposes, this argument is only correct up to a first approximation. It is worth giving a correct justification for this approximation seeing as how Part II of this thesis is devoted entirely to working with the intensities of diffracted waves.

Erwin Schrödinger (Schrödinger, 1926) deduced a suitable representation of non-relativistic quantum mechanics that was formalised by Paul Dirac (see (Dirac, 1981)). In this representation, the modulus squared of a quantity called the wave function is a real number interpreted as the probability of finding a particle at a given place and time. The Lippmann-Schwinger equation relates a scattered wave function with a scattering field (Lippmann and Schwinger, 1950) and so the diffraction pattern is proportional to the modulus squared of its solution, which is in general very difficult to obtain for an arbitrary scattering potential. If perturbation theory is used to expand the wave function as a power series the lowest order term in the expansion can be taken as a suitable approximation to the true solution provided the diffraction pattern is recorded far from the scattering event (far with respect to the size of the source). This is called the first Born approximation, which turns out to be the Fourier transform of the scattering potential (Wu and Ohmura, 2011).

The exact form of the scattering potential depends on the nature of the source (in this case X-rays being the incident radiation), scattering from atoms being appropriately described by an atomic scattering factor. This means that to a good approximation the diffracted X-rays are simply the Fourier modes $F(\mathbf{s})$ of the electronic density $\rho(\mathbf{r})$ of the scattering object about its centre of mass

$$F(\mathbf{s}) = \int_{\mathbb{R}^3} d^3r \; e^{2\pi i \mathbf{r}\cdot\mathbf{s}} \rho(\mathbf{r}) \; . \tag{5.1}$$

With these values for the diffracted waves it is possible to reconstruct the scattering density. That is to say, taking the inverse Fourier transform of waves diffracted from a crystal will result in a good approximation to the distribution of electrons in the crystal. From the convolution theorem (Bracewell, 1980), the Fourier transform of the crystal is the product of the Fourier transform of the unit cell and the Fourier transform of the crystal lattice. Diffracted waves are only nonzero at points where the Fourier transform of the lattice does not vanish. By considering the lattice to be an infinite array of points, this occurs only for integral multiples of three basis vectors, $\mathbf{e}^*_1$, $\mathbf{e}^*_2$, and $\mathbf{e}^*_3$, generating the Fourier transform of the lattice in momentum (reciprocal) space

$$\mathbf{s} = h\mathbf{e}^*_1 + k\mathbf{e}^*_2 + l\mathbf{e}^*_3 \; , \quad \text{integer} \;\; h,k,l \; . \tag{5.2}$$

These integer values are precisely the Miller indices (h,k,l) discussed in Section 5.1.1, each spot in the diffraction pattern corresponding to a Fourier mode or *structure factor* with index (h,k,l). The Fourier transform of all structure factors is therefore an infinite sum that converges to the electron density distribution of the crystal

$$\rho(r_1,r_2,r_3) = \frac{1}{V} \sum_{hkl} F_{hkl} e^{-2\pi i(hr_1+kr_2+lr_3)} \; , \tag{5.3}$$

where V is the volume of the unit cell.

From expression (5.1) it is clear that every atom in the crystal contributes to the value of a structure factor. Structure factors are complex-valued functions of atomic coordinates that in general have both real and imaginary parts. The exceptions are centric structure factors with index (h,k,l) sent to minus themselves by a symmetry operation of the space group and are therefore real-valued. Being an observable on the other hand, the diffraction pattern is entirely real, with each spot or intensity I_{hkl} proportional to the amplitude squared of the structure factor, $I_{hkl} \propto |F_{hkl}|^2$. Herein lies the famous phase problem associated with X-ray crystallography. Whereas both amplitudes and phases of structure factors are required to calculate the Fourier series for electron density, it is only the square of their amplitudes that are observed in practice. Although "solutions" to the phase problem are used routinely in macromolecular crystallography, these are only solutions in the sense that they rely on additional experimental information or *a priori* knowledge about the contents of the crystal to provide estimates for the phases. Intractability of the phase problem without prior information is best illustrated by using a discrete version of (5.1) to calculate the amplitude of a structure factor

$$|F(\mathbf{s})| = \sqrt{F(\mathbf{s})F(-\mathbf{s})} = \left(\sum_{n,m} \rho(\mathbf{r}_n)\rho(\mathbf{r}_m) e^{-2\pi i(\mathbf{r}_n - \mathbf{r}_m)\cdot\mathbf{s}} \right)^{1/2} . \tag{5.4}$$

From (5.4) one can see that intensities only contain information about interatomic distances and not the atomic coordinates in a crystal. Direct methods based on Sayre's equation (Sayre, 1952) are used in small molecule crystallography to approximate phases with little prior information, but these have little applicability to the complicated structures found in macromolecular crystallography, except perhaps at high resolution and during substructure solution (Sheldrick et al., 1993). The persistence of crystallographers to find ways around the phase problem is exemplified by a collection of reviews from the

pioneer of direct methods, Herbert Hauptman, spanning three decades and each entitled "The phase problem in X-ray crystallography" (Hauptman, 1983, 1991, 2001, 2008). It has been argued however, that in principle, chemical knowledge (i.e. what possible structures make chemical sense) provides enough information to solve the phase problem (Bricogne, 1993).

5.1.3 Likelihood methods in crystallography

Solutions to the phase problem in macromolecular crystallography will not be reviewed here since they are not of primary concern to Part II (chapters devoted to the subject can be found in all of the texts mentioned previously). Typical methods can be classified according to whether they are experimental or a form of molecular replacement. For experimental methods, a good pedagogical review is (Taylor, 2010). Molecular replacement is based on a method developed by Michael Rossmann and David Blow (Rossmann and Blow, 1962) that relies on the atomic model of a homologous macromolecule (homologous to that in the crystal) being used to calculate initial phases for the structure factors prior to refinement. Rossmann has put together a collection of historical papers on the subject (Rossmann, 1972). During molecular replacement, the phase problem reduces to finding the correct position and orientation for copies of a homologous model in the unit cell.

Once additional experimental information or a molecular replacement-based approach has been used to approximate phases of the structure factors, the crystallographer will embark on a journey of refinement in order to improve the quality of an initial atomic model. There are many criteria used to improve a model, such as restraints on bond geometry and stereochemistry, but the main driving force is that under these restraints the correlation between structure factors amplitudes calculated from the model (F^c_{hkl}) and

those observed experimentally (F^o_{hkl}) increases over the refinement period. The conventional measure of agreement between model and data is given by the R value

$$R = \frac{\sum_{hkl} \; ||F^o_{hkl}| - |F^c_{hkl}||}{\sum_{hkl} \; |F^o_{hkl}|} \; , \tag{5.5}$$

but if left entirely to the refinement programs of today even an unacceptable model can still yield a very low R value (Kleywegt and Jones, 1997). This problem of "over fitting" is partially solved using a R free value (Brunger, 1992) that is calculated from a subset (usually 5%) of observed structure factor amplitudes left out of the refinement procedure. If the R free value decreases along with the R value during a cycle of refinement this is taken as a sign that refinement is progressing in the right direction. Older refinement programs based on least-squares algorithms attempt to improve the model by minimising a weighted difference between observed and calculated amplitudes over the entire set of structure factors. Whilst these programs have proved to be enormously powerful in macromolecular crystallography it turns out, for reasons to be discussed, that newer, likelihood-based algorithms are superior in the majority of cases. Likelihood-based algorithms are also used routinely in experimental phasing and molecular replacement, but here refinement will be used as an illustrative example.

Maximum-likelihood is a method for estimating the parameters of a probability distribution on the basis of observed data. For example, given a statistical model $p(\{|F^o_{hkl}|\}; \text{atoms})$ for the conditional distribution of observed structure factor amplitudes provided the atomic coordinates of a crystal, maximum-likelihood selects the atomic coordinates that maximise the probability that amplitudes are those observed. This requires a precise formula for the conditional probability distribution for data given the model, which must be approximated in practice by using simplifying assumptions about the nature of

experimental errors and properties of structure factors. Even with these assumptions, maximum-likelihood is better suited for the refinement of macromolecular crystal structures because conditions necessary for the least-squares method are not satisfied (Pannu and Read, 1996). Application of maximum-likelihood to protein structure refinement was originally suggested by Randy Read (Read, 1990) and Gérard Briconge (Bricogne, 1991, 1993), and this was followed by three different implementations (Pannu and Read, 1996; Dodson et al., 1996; Murshudov et al., 1997). Compared with least-squares methods the likelihood-based approach achieves more than twice the improvement in average phase error (Pannu and Read, 1996). This is because least-squares maximisation incorrectly assumes that the distributions between observed and calculated structure factor amplitudes are Gaussian with a standard deviation that is independent of the atomic model. Maximum-likelihood generalises least squares by removing this assumption.

If errors in different observations are assumed to be independent, the conditional probability density or likelihood function to be maximised during refinement in REFMAC (Murshudov et al., 1997) is a product of the conditional probability for each observed reflection

$$p(\{|F^o_{hkl}|\};\{|F^c_{hkl}|\}) = \prod_{hkl} p(|F^o_{hkl}|;|F^c_{hkl}|)\ . \tag{5.6}$$

In practice it is convenient to minimise the negative log of the likelihood function (LLK) since this is equivalent to maximising the likelihood function, but involves a sum over (h,k,l) rather than a product

$$LLK = -\sum_{hkl} \ln p(|F^o_{hkl}|;|F^c_{hkl}|)\ . \tag{5.7}$$

To derive the functional form of LLK, errors in different atoms are assumed independent and the distribution of structure factors takes the form introduced

by Luzzati (Luzzati, 1952) and generalised by Read (Read, 1990)

$$p(F_{hkl};F^c_{hkl}) = \begin{cases} \frac{1}{\pi\Sigma_{wc}} \exp\left(-\frac{|F_{hkl}-F^{wc}_{hkl}|^2}{\Sigma_{wc}}\right) & \text{acentric ,} \\ \frac{1}{\sqrt{2\pi\Sigma_{wc}}} \exp\left(-\frac{|F_{hkl}-F^{wc}_{hkl}|^2}{2\Sigma_{wc}}\right) & \text{centric ,} \end{cases} \tag{5.8}$$

where F^{wc}_{hkl} is the calculated structure factor weighted appropriately and Σ_{wc} a weighted variance. The conditional probability for an amplitude given the model is then obtained by integrating over the unknown phase difference between observed and calculated structure factors. A third assumption is that measurement error is normally distributed with variance σ^2_{hkl}, so the final conditional probability for an observation can be obtained by convoluting the distribution of true amplitudes with a Gaussian. In REFMAC (Murshudov et al., 1997) the target distribution takes the form

$$p(|F^o_{hkl}|;|F^c_{hkl}|) = \begin{cases} \frac{2|F^o_{hkl}|}{2\sigma^2_{hkl}+\Sigma_{wc}} \exp\left(-\frac{|F_{hkl}|^2+|F^{wc}_{hkl}|^2}{2\sigma^2_{hkl}+\Sigma_{wc}}\right) I_0\left(\frac{2|F^o_{hkl}||F^{wc}_{hkl}|}{2\sigma^2_{hkl}+\Sigma_{wc}}\right) & \text{acentric ,} \\ \frac{2}{\pi\sigma^2_{hkl}+\pi\Sigma_{wc}} \exp\left(-\frac{|F_{hkl}|^2+|F^{wc}_{hkl}|^2}{2(\sigma^2_{hkl}+\Sigma_{wc})}\right) \cosh\left(\frac{|F^o_{hkl}||F^{wc}_{hkl}|}{\sigma^2_{hkl}+\Sigma_{wc}}\right) & \text{centric ,} \end{cases} \tag{5.9}$$

where $I_0(X)$ is the zeroth order modified Bessel function of the first kind. Parameters of the likelihood function are typically refined in resolution bins, but can also be expressed as continuous functions of resolution. Due to the advantage that likelihood-based algorithms have over conventional least-squares, the newest version of REFMAC (Murshudov et al., 2011) is now one of the most popular software tools for the refinement of macromolecular crystal structures.

5.2 Processing data from multiple crystals

5.2.1 Data reduction

Following on from the brief introduction to macromolecular crystallography, this Section will detail current methods for processing data contained within the crystallographic diffraction images. This stage of the crystallographic protocol amounts to space group assignment, integration of diffraction intensities, and scaling and correction of data. Here the focus will be on the latter. Two excellent reviews on modern data processing are by Wolfgang Kabsch (Kabsch, 2010b) and Phil Evans (Evans, 2005).

Estimation of intensities is preceded by a refinement of unit cell parameters required predict the location and Miller indices of spots. The intensity of a reflection can be distributed across one or more images and so the 'partiality' of an observation must also be accounted for. Once the experimental parameters have been refined the shape of a reflection can be described using a Gaussian model involving the standard deviations of the reflecting range and beam divergence (Kabsch, 2010b). This defines a region of the image over which to integrate. A predicted intensity distribution in a given region is then estimated by minimising a function accounting for background contents, actual contents, and variance of pixels in that domain.

Requirement for the scaling of diffracted intensities observed in an experiment arises because of physical factors that affect the relative values of the same observations made on different images. When combined, these factors all depend on the variability of the X-ray beam, detector, and way in which the crystal is rotated during data collection. Diffraction patterns are sensitive to fluctuations in beam intensity and the calibration of detectors, and the crystal lattice is also susceptible to radiation damage that can mean different reflec-

tions change at different rates from one image to the next (see also Section 5.2.2). Scaling procedures that account for these factors model the corrections as a function of time (i.e. image number or rotation angle) and apply a different scale factor for each image by using equivalent observations on each frame. Once a scale factor has been applied, the agreement of equivalent intensities form a basis for whether certain measurements should be included or rejected from the next round of data processing (Evans, 2005). Most scaling models are based upon the method of Kabsch (Kabsch, 1988) and include a term that weights the overall scale factor for a reflection according to the rotation angle between the adjacent images. So-called B factors are introduced in a similar fashion in order to provide a resolution-dependent correction for various factors including radiation damage. Both scale factor and B factor terms are taken to be smooth functions of rotation angle. Additional terms included in the overall scale factor account for absorption in the secondary beam direction and diffuse scattering that causes long tails on reflections (Evans, 2005).

Parameters of the overall scale factors are determined by minimising the differences between symmetry-related observations over all the collected images. The function minimised is typically

$$\Psi = \sum_{hkl} \sum_{i} (I_{hkl,i} - g_{hkl,i}\langle I_{hkl}\rangle)^2/\sigma^2_{hkl,i} \,, \tag{5.10}$$

where $I_{hkl,i}$ is the ith observation of reflection I_{hkl}, $g_{hkl,i}$ is the associated scale factor, and $\langle I_{hkl}\rangle$ and $\sigma_{hkl,i}$ are the weighted mean and standard deviation of that measurement (Hamilton et al., 1965; Howell and Smith, 1992). Ψ is minimised over all (h,k,l) to obtain a value for each parameter, and so the appropriate scale factor can be applied to obtain the 'true' intensities. This procedure can then be repeated if required, using true intensities from the previous cycle to update new scaling factors (Kabsch, 2010b).

A number of different measures have been introduced to assess the quality of data and suggest a resolution cutoff. R factors taking into account multiplicity of observations have been suggested by several crystallographers (Diederichs and Karplus, 1997; Weiss and Hilgenfeld, 1997; Weiss, 2001), but currently there is no universally accepted measure for deciding on including or rejecting observations past a given resolution. A typical resolution cutoff is when the average signal (I/σ_I) falls below 2.0, but including data beyond this limit has been proven to improve the quality of a final atomic model (Karplus and Diederichs, 2012). Recently the correlation of an observed dataset with the underlying true signal (CC^*) has been suggested as a single statistic for deciding whether data is useful or not. However, this still requires a user-defined cutoff that can seem to vary arbitrarily from one experimenter to the next. There remain major cases of user indecision and conflicts within the crystallographic community (Evans and Murshudov, 2013).

5.2.2 Requirement for multiple crystals

Although protein crystallography is now routinely used in laboratories worldwide, there is still the problem of obtaining well diffracting crystals containing large or insoluble macromolecules and their complexes. A great deal of time needs to be dedicated to overcoming the challenges posed during each stage of the crystallisation, data collection, and the structure solution process, particularly for membrane proteins (Carpenter et al., 2008), viruses (Fry et al., 1999) and large molecular complexes like the ribosome (Brodersen et al., 2003). Another significant obstacle is radiation damage to the crystal caused by absorbed X-rays leading to chemical deformations of the crystalline structure and, ultimately, the demise of intensities in the diffraction pattern (Holton, 2009; Garman, 2010). That smaller crystals suffer more from radiation damage was

confirmed in a recent study where it was demonstrated that damage-limiting scattering power is proportional to crystal volume and inversely proportional to the molecular weight of the asymmetric unit (Holton and Frankel, 2010). To overcome the problem of radiation damage it is common practice to combine data from many different crystals into a single dataset (Riekel et al., 2005). However, merging data from multiple crystals poses further problems during data reduction.

Merging data from highly isomorphic crystals can be viewed as a simple extension of the scaling procedure described in Section 5.2.1, assuming multiple batches of different images from the same crystal. The user will typically 'force' the scaling program to a particular set of unit cell parameters that he choses to be 'true', and this will have little effect on the final statistics since unit cells vary only slightly from one batch to another. The variation between batches of images from highly isomorphic crystals does not much exceed the variability of one image to another, and scaling models are well equipped to deal with these cases. In fact, merging several complete datasets from multiple isomorphic crystals will increase multiplicity and strengthen the signal-to-noise ratio of combined observations (Liu et al., 2012). When combining data from less-isomorphic crystals however, it is often the case that merging will fail due to the incompatibility of variable unit cell parameters and contents (Foadi et al., 2013). Even if merging is achieved, the statistics of combined data will be poor and a dataset will not reflect the true quality of information contained within each of the individual image batches. The root of this problem is that conventional scaling assumes batches were collected from the same crystal whilst only the factors discussed in Section 5.2.1 are used to account for variability between related observations.

Several techniques have been developed to optimise the quality of a data-

set collected from multiple crystals. Data from around 400 crystals were combined in a remarkable effort using only a few images per crystal to solve the recent structure of a lipid G-protein coupled receptor (Hanson et al., 2012). In that study, low-resolution data were used as a reference and subsequently combined high-resolution data were rejected if the R factor remained high after scaling. To assist with the selection of images to be combined, James Foadi and colleagues have developed the software BLEND that sorts data according to a clustering scheme (Foadi et al., 2013). A similar method has been introduced by Giordano *et al.* (Giordano et al., 2012). Although these approaches help with the task of manually selecting data to be combined, they do not make use of valuable information contained within each batch of images.

5.3 Part II outline

The remainder of Part II is devoted to the derivation and implementation of an intensity-based likelihood algorithm that provides an estimation of the true covariance between any number of crystal structures. Covariance is a measure of how much two random variables x_i, x_j change together and is defined to be

$$\Sigma_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \ , \tag{5.11}$$

where angled parentheses denote an expectation value. By true covariance or correlation is meant the physical covariance between unobserved structure factors not affected by measurement error, rather than the observed covariance or correlation between intensities. In general, when measurement errors are not correlated with the values being measured, observed covariance is the sum of the covariance describing the true relationship and the covariance de-

scribing measurement error

$$\Sigma_{\text{observed}} = \Sigma_{\text{true}} + \Sigma_{\text{measured}} \,. \tag{5.12}$$

Figure 5.2 illustrates how the observed correlation between two simulated data sets decreases as measurement error increases. Gaussian noise with zero mean and increasing variance has been used to model how measurement error affects the observed correlation between two identical data sets. In the case of crystallographic data the true covariance is the covariance between corresponding structure factors of two or more crystals, but experimental error is introduced during the measurement of intensities. Measurement error is quantified by the σ_I's however, and so from (5.12) one can see that in principle it should be possible to obtain an estimate for true covariance by a method that accounts for observed covariance and the σ_I's correctly. This is the maximum likelihood method furnished with the appropriate statistical model.

The approach differs dramatically from that of Foadi *et al.* (Foadi et al., 2013) and Giordano *et al.* (Giordano et al., 2012), who do not account for the presence of experimental noise or general coordinate error. A key result is that the true covariance between two crystal structures is much higher than that calculated from standard sample covariance (as illustrated by Figure 5.2), indicating much more information from related crystals can be used than previously thought. On this basis, a method of predicting values for unobserved data points ('completing' data sets) has been constructed and incorporated into the algorithm. Once again this approach differs from the convention of combining data from multiple crystals (i.e. assuming a single underlying crystal structure), instead producing a complete reflection list for all the different crystals for which data are observed. This allows the refinement of individual structures with a different structure corresponding to each crystal.

Test cases are presented to demonstrate that true covariance estimation

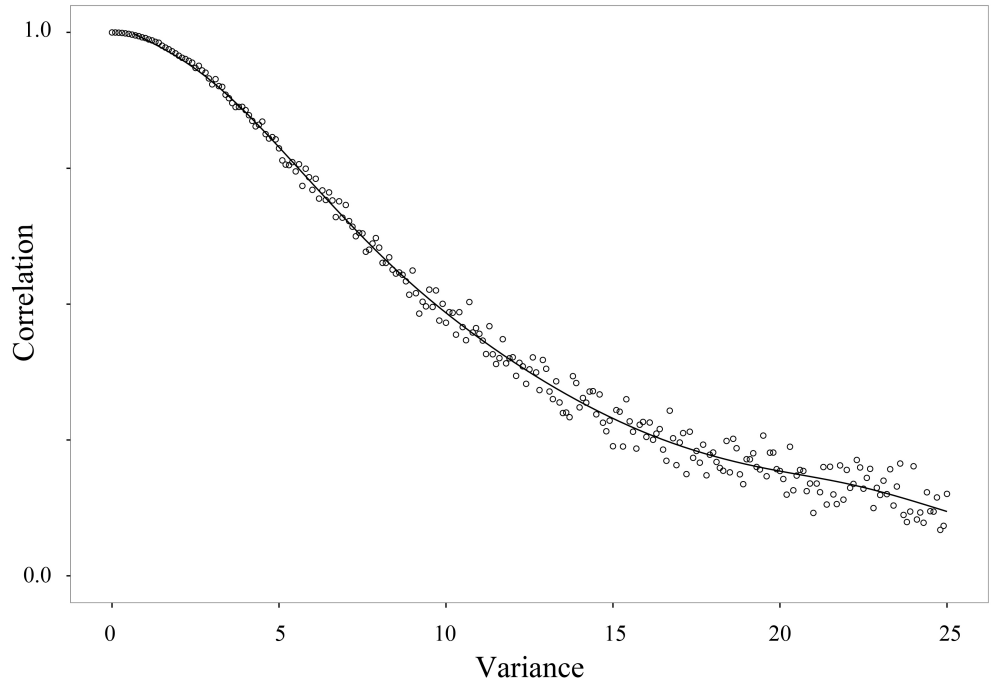


Figure 5.2: Effect of experimental noise on observed correlation

As the variance of the Gaussian noise measurement term increases, the observed correlation decreases to zero even though the true correlation between the two sets of data is 1.0.

remains robust under experimental noise and partial incompleteness of a data set. It is shown that true covariance estimation has a wide range of applications including assistance with a data processing strategy and assessing whether or not a ligand is bound inside a crystal. Part II finishes with a discussion about implications and future developments of the method, including a generalisation of the entire theory to the case of multi-crystal structure refinement.

Chapter 6

The multi-crystal likelihood method

The purpose of the algorithm is to estimate true covariance between crystals for which a user has multiple sets of related, partially overlapping (but not necessarily complete) diffraction data. Once the true covariance matrix $\mathbf{\Sigma}$ has been estimated correctly it can be used to cluster crystals and devise a strategy for further data processing. It may also be used to come up with an estimate for values of missing data.

6.1 The algorithm

6.1.1 Probabilities for related intensities

Here, the probability of observing intensity values from N different crystals is derived given the covariances between them. Read (Read, 1990) gave an interpretation of variance Σ in the generalised Luzzati distribution of acentric

structure factors between two related structures

$$p(F;G) = \frac{1}{\epsilon\pi\Sigma}\exp\left(-\frac{|F-G|^2}{\epsilon\Sigma}\right) , \tag{6.1}$$

which proves a convenient starting point at which to introduce some notation. The intensity factor ϵ depends on space group and the variance is in general a function of resolution given by

$$\Sigma = \sum_n \langle |f_n \exp(2\pi i \mathbf{s}\cdot\Delta\mathbf{r}_n) - g_n|^2\rangle . \tag{6.2}$$

In this expression, f_n and g_n are the scattering factors of the nth atom in crystals with structure factors F and G, respectively. The vectors $\Delta\mathbf{r}_n$ represent the coordinate differences between the two crystals, and angled parentheses denote the expected value with respect to the probability distribution. This expression for Σ arises through the central limit theorem considering f_n, g_n, and $\Delta\mathbf{r}_n$ to be random variables giving independent contributions to the difference between structure factors.

The generalised Luzzati distribution (6.1) can be extended to N crystals $\{C_i\}$ $(i = 1, 2, ..., N)$ by assuming existence of a hypothetical, unobserved crystal C_0 to which each crystal C_i is related. By the central limit theorem, when there are a sufficient number of independent differences between crystals, the conditional distribution of structure factors $\{F_i\}$ (F_i corresponding to C_i) can be written as a complex multivariate Gaussian distribution

$$p(\{F_i\};F_0) = \frac{1}{\epsilon\pi^N \det(\mathbf{\Sigma})}\exp\left[-\frac{1}{\epsilon}(\mathbf{F}-\mathbf{DF_0})^\dagger \mathbf{\Sigma}^{-1}(\mathbf{F}-\mathbf{DF_0})\right] , \tag{6.3}$$

where $\mathbf{F}$ is the vector of N structure factors, $\mathbf{DF_0}$ a vector of expectation values, and $\mathbf{\Sigma}$ is the $N \times N$ covariance matrix. The conditional distribution of amplitudes $p(\{|F_i|\};|F_0|)$ is then obtained by multiple integration over all the unknown phase differences. Maximum likelihood is insensitive to a variable

transformation from $|F|$ to $|F|^2$ (Pannu and Read, 1996), and so $\sqrt{I_i}$ can be substituted for $|F_i|$ and $\sqrt{I_0}$ for $|F_0|$ in $p(\{|F_i|\};|F_0|)$ to obtain $p(\{I_i\};I_0)$. The marginal distribution $p(\{I_i\})$ is then

$$p(\{I_i\}) = \int dI_0 \; p(\{I_i\};I_0)\cdot p(I_0)\;, \tag{6.4}$$

where $p(I_0)$ is a Wilson distribution (Wilson, 1949) with parameter Σ_0. Measurement error can be approximated as Gaussian, which is sufficient when errors are assumed independent, and so the μth observation of an intensity I_i from crystal C_i is normally distributed about I_i with variance $\sigma^2_{i,\mu}$

$$p_\mu(I_i^o;I_i) = \frac{1}{\sqrt{2\pi\sigma^2_{i,\mu}}}\exp\left(-\frac{(I_i^o-I_i)^2}{2\sigma^2_{i,\mu}}\right)\;. \tag{6.5}$$

The marginal probability distribution for a specified set of observations $\{I^o\}$ is therefore

$$p(\{I^o\}) = \int_{\mathbb{R}^N} d^N I \prod_i\prod_\mu p_\mu(I_i^o;I_i)\cdot p(\{I_i\})\;. \tag{6.6}$$

Evaluation of the integral in (6.6) is clearly not possible and so an approximation to $p(\{I^o\})$ is required.

Considering the case of merged data when the requirement for the Greek indices is no longer necessary, from the form of the distribution (6.6) it follows that

$$\langle I_i^o\rangle = \mathbf{\Sigma}_{ii}\;, \tag{6.7}$$

and

$$\langle I_i^o I_j^o\rangle - \langle I_i^o\rangle\langle I_j^o\rangle = \sigma_i^2\delta_{ij} + (\mathbf{\Sigma}_{ij})^2\;, \tag{6.8}$$

where δ_{ij} is the Kronecker delta. It is possible to build a multivariate Gaussian approximation using these values, which takes the form

$$p(\mathbf{I^o}) = \frac{1}{\sqrt{(2\pi)^N\,\det(\hat{\mathbf{\Sigma}})}}\exp\left(-\frac{1}{2}(\mathbf{I}^o-\langle\mathbf{I}\rangle)^T\hat{\mathbf{\Sigma}}^{-1}(\mathbf{I}^o-\langle\mathbf{I}\rangle)\right)\;, \tag{6.9}$$

where $\langle \mathbf{I} \rangle$ has components $\mathbf{\Sigma}_{ii}$, and $\hat{\mathbf{\Sigma}}$ has diagonal elements $\sigma_i^2 + (\mathbf{\Sigma}_{ii})^2$ and off-diagonal elements $(\mathbf{\Sigma}_{ij})^2$. Here intensities have been normalised so that $\epsilon = 1$. The distribution given by expression (6.9) will be considered as a suitable approximation to $p(\{I^o\})$. Constructing the marginal distribution for the case of missing data (i.e. when only a subset of crystals have representatives for a particular I_{hkl}) is straightforward because for multivariate Gaussians one only needs to drop the irrelevant variables from the mean vector and covariance matrix. Proof follows from the definitions of multivariate normal distributions and linear algebra. Consequently, the probability of observing values $\{I_i^o\}$ corresponding to intensity I_{hkl} from crystals $\{C_i\}$ is given by $p(\mathbf{I^o})$, and is uniquely determined by the covariance $\mathbf{\Sigma}$ between them.

6.1.2 Likelihood function and derivatives

In this Section, the distribution $p(\mathbf{I^o})$ is to be interpreted as a conditional probability for observing values $\{I_i^o\}$ (corresponding to a particular reflection) given the relatedness of crystals $\{C_i\}$. Distances between crystals are quantified by the covariance matrix $\mathbf{\Sigma}$, whose off-diagonal elements $\mathbf{\Sigma}_{ij}$ can be related to the coordinate differences between crystals C_i and C_j by the formula

$$\mathbf{\Sigma}_{ij} = \sum_{n,m} \langle | f_{i,n} f_{j,m} \cos(2\pi \mathbf{s} \cdot [\mathbf{r}_{i,n} - \mathbf{r}_{j,m}]) | \rangle \ . \tag{6.10}$$

Clearly $\mathbf{\Sigma}$ is a matrix-valued function of resolution, which reflects the fact that differences become more apparent at higher resolution (e.g. helices that look identical at 5 Å can be distinguished by side chains at 2 Å). With data from N different crystals $p(\mathbf{I^o})$ may be treated as a likelihood function of the covariance matrix $\mathbf{\Sigma}$. Over all reflections I_{hkl} the negative log-likelihood function to be minimised (ignoring constants that do not affect minimisation) takes the

form

$$LLK = \sum_{hkl} \ln\det(\hat{\boldsymbol{\Sigma}}_{hkl}) + (\mathbf{I^o}_{hkl} - \langle \mathbf{I}_{hkl} \rangle)^T (\hat{\boldsymbol{\Sigma}}_{hkl})^{-1} (\mathbf{I^o}_{hkl} - \langle \mathbf{I}_{hkl} \rangle) \,. \tag{6.11}$$

By choosing $\boldsymbol{\Sigma}$ so that LLK is at a minimum the algorithm yields the best estimate for the true covariance between crystal structures.

To stem the onslaught of indices that will inevitably appear in the following calculations it is necessary to introduce some indexing conventions. Since LLK is linear in terms involving different (h,k,l), from here on in Miller indices will be dropped and letters i,j,k,l will be *reserved exclusively for labelling components of vectors and matrices* (i.e. i,j,k,l will now refer to different crystals). Unless explicitly stated, all calculations will be assumed to include a summation over Miller indices. If a particular reflection must be referred to this will be done using the symbol $\mathbf{h}$ and the resolution bin to which it belongs will be labelled by m.

Quasi-Newton methods for minimisation of LLK with respect to the components of $\boldsymbol{\Sigma}$ require first derivatives and an approximation to the Hessian matrix of second derivatives (Nocedal and Wright, 1999). The expectation values of mixed second derivatives of LLK vanish

$$\left\langle \frac{\partial LLK}{\partial \langle \mathbf{I} \rangle_i} \frac{\partial LLK}{\partial (\hat{\boldsymbol{\Sigma}}^{-1})_{jk}} \right\rangle = \left\langle \frac{\partial^2 LLK}{\partial \langle \mathbf{I} \rangle_i \partial (\hat{\boldsymbol{\Sigma}}^{-1})_{jk}} \right\rangle = 0 \,, \tag{6.12}$$

and so the Fisher information matrix

$$\mathcal{I} = \left\langle \frac{\partial LLK}{\partial \boldsymbol{\Sigma}_{ij}} \frac{\partial LLK}{\partial \boldsymbol{\Sigma}_{kl}} \right\rangle = - \left\langle \frac{\partial^2 LLK}{\partial \boldsymbol{\Sigma}_{kl} \partial \boldsymbol{\Sigma}_{ij}} \right\rangle \tag{6.13}$$

makes for a simplified approximation of the Hessian. To evaluate derivatives correctly, the symmetric property of the covariance matrix ($\boldsymbol{\Sigma}_{ij} = \boldsymbol{\Sigma}_{ji}$) must be accounted for. For a symmetric matrix $\mathbf{A}$ there exists identities

$$\frac{\partial \det \mathbf{A}}{\partial \mathbf{A}_{ij}} = \det(\mathbf{A})[2(\mathbf{A}^{-1})_{ij} - \delta_{ij}(\mathbf{A}^{-1})_{ij}] \,, \tag{6.14}$$

and

$$\frac{\partial(\mathbf{A}^{-1})_{ij}}{\partial \mathbf{A}_{kl}} = -(\mathbf{A}^{-1})_{ik}(\mathbf{A}^{-1})_{lj} - (\mathbf{A}^{-1})_{jk}(\mathbf{A}^{-1})_{li} + \delta_{kl}(\mathbf{A}^{-1})_{ik}(\mathbf{A}^{-1})_{lj} \; . \quad (6.15)$$

Using these with the chain rule and formulae from Section 6.1.1, after some algebraic manipulations one finds first derivatives can be expressed as

$$\frac{\partial LLK}{\partial \boldsymbol{\Sigma}_{ij}} = 2(2-\delta_{ij})\boldsymbol{\Sigma}_{ij}[(\hat{\boldsymbol{\Sigma}}^{-1})_{ij} - \hat{\mathbf{S}}_{ij}] + 2\delta_{ij}\sum_{k}(\boldsymbol{\Sigma}^{-1})_{ik}(\boldsymbol{\Sigma}_{kk} - \mathbf{I}^{\mathbf{o}}{}_{k}) \; , \quad (6.16)$$

where $\hat{\mathbf{S}} = (\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{S}\hat{\boldsymbol{\Sigma}}^{-1})$, $\mathbf{S}_{ij} = (\boldsymbol{\Sigma}_{ii} - \mathbf{I}^{\mathbf{o}}{}_{i})(\boldsymbol{\Sigma}_{jj} - \mathbf{I}^{\mathbf{o}}{}_{j})$, and that $\boldsymbol{\mathcal{I}}$ has elements

$$\left\langle \frac{\partial^2 LLK}{\partial \boldsymbol{\Sigma}_{ij}\partial \boldsymbol{\Sigma}_{kl}} \right\rangle = \delta_{ij}\delta_{kl}(\hat{\boldsymbol{\Sigma}}^{-1})_{ii} + 4\boldsymbol{\Sigma}_{ij}\boldsymbol{\Sigma}_{kl}(2-\delta_{ij})(2-\delta_{kl})(\hat{\boldsymbol{\Sigma}}^{-1})_{ki}(\hat{\boldsymbol{\Sigma}}^{-1})_{lj} \; . \quad (6.17)$$

Due to the symmetry of the covariance matrix, there are only $N(N+1)/2$ independent parameters that need to be refined. Numerical optimisation of $\boldsymbol{\Sigma}$ using a quasi-Newton method to minimise LLK is described in Section 6.2.3.

6.1.3 Starting values and prediction of true intensities

Numerical optimisation relies on an initial guess $\boldsymbol{\Sigma}^0$ at the covariance matrix $\boldsymbol{\Sigma}$. Owing to the relation $I \propto |F|^2$, the matrix of sample covariances between intensities ($\boldsymbol{\Sigma}^I$) is not a suitable choice for $\boldsymbol{\Sigma}^0$ because the latter should be representative of the covariance between structure factors. Instead $\boldsymbol{\Sigma}^I$ can be used to form a sample correlation matrix $\boldsymbol{\rho}$ with elements

$$\boldsymbol{\rho}_{ij} = \frac{\boldsymbol{\Sigma}^I_{ij}}{\sqrt{\boldsymbol{\Sigma}^I_{ii}\boldsymbol{\Sigma}^I_{jj}}} \; . \quad (6.18)$$

Diagonal elements of the starting covariance matrix $\boldsymbol{\Sigma}^0$ can then be approximated by the average intensity in each resolution bin, and off-diagonal elements by

$$\boldsymbol{\Sigma}^0_{ij} = \boldsymbol{\rho}_{ij}\sqrt{\boldsymbol{\Sigma}^0_{ii}\boldsymbol{\Sigma}^0_{jj}} \; , \quad i \neq j \; . \quad (6.19)$$

This yields covariances that are on the same scale as structure factors and prove to be suitable starting values for numerical optimisation. However, minimisation appears unstable if starting covariance values are less than $\sim$ 0.15, and so if $(\mathbf{\Sigma}^0_m)_{ij} < 0.15$ then $(\mathbf{\Sigma}^0_m)_{ij}$ is currently set to the value $(M-1)(\mathbf{\Sigma}^0_{m-1})_{ij}/M$, where M is the total number of resolution bins. This appears a sufficient precaution for all intended purposes.

Once $\mathbf{\Sigma}$ has been optimised it can be used to predict the true values of intensities $\mathbf{I}$ not affected by measurement error. The conditional probability for intensities given the observed reflections $\mathbf{I^o}$ is

$$p(\mathbf{I};\mathbf{I^o}) = \frac{p(\mathbf{I},\mathbf{I^o})}{p(\mathbf{I^o})} \,. \tag{6.20}$$

From (6.6) $p(\mathbf{I^o})$ does not depend on $\mathbf{I}$, therefore $p(\mathbf{I};\mathbf{I^o})$ can be taken to be proportional to the joint probability distribution $p(\mathbf{I},\mathbf{I^o})$. Assuming a Gaussian prior for $p(\mathbf{I},\mathbf{I^o})$,

$$p(\mathbf{I};\mathbf{I^o}) \propto \exp\left[-\frac{1}{2}(\mathbf{I^o}-\mathbf{I})^T\mathbf{C}(\mathbf{I^o}-\mathbf{I}) - \frac{1}{2}(\mathbf{I}-\langle\mathbf{I}\rangle)^T\tilde{\mathbf{\Sigma}}^{-1}(\mathbf{I}-\langle\mathbf{I}\rangle)\right] \,, \tag{6.21}$$

where $\tilde{\mathbf{\Sigma}}_{ij} = (\mathbf{\Sigma}_{ij})^2$, and $\mathbf{C}_{ij} = \delta_{ij}c_{ij}/\sigma_i^2$ with $c_{ij} = 1$ if reflections i and j have been observed and $c_{ij} = 0$ otherwise. Taking the vectorial derivative of (6.21) with respect to $\mathbf{I}$ yields

$$\frac{\partial p(\mathbf{I};\mathbf{I^o})}{\partial \mathbf{I}} \propto -[\mathbf{C}(\mathbf{I}-\mathbf{I^o}) + \tilde{\mathbf{\Sigma}}^{-1}(\mathbf{I}-\langle\mathbf{I}\rangle)]p(\mathbf{I};\mathbf{I^o}) \,, \tag{6.22}$$

and the best value for $\mathbf{I}$ is found by setting the right-hand-side equal to zero and rearranging to obtain

$$\mathbf{I} = (\mathbf{C}+\tilde{\mathbf{\Sigma}}^{-1})^{-1}(\mathbf{CI^o} + \tilde{\mathbf{\Sigma}}^{-1}\langle\mathbf{I}\rangle) \,. \tag{6.23}$$

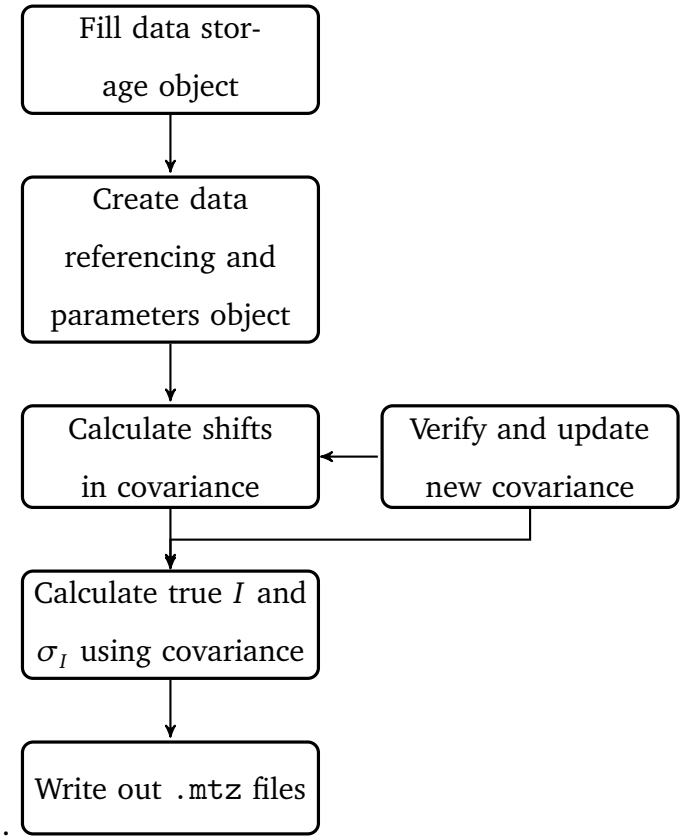
The individual 'true' σ_i^2 can be approximated by diagonals of the inverse of the second derivate of (6.21)

$$\sigma_i^2 = [(\mathbf{C}+\tilde{\mathbf{\Sigma}}^{-1})^{-1}]_{ii} \,. \tag{6.24}$$

6.2 Implementation

Implementation of the algorithms described in Section 6.1 consists of approximately 5000 lines of *C++* that is logically divided into components that match the system diagram in Figure 6.1. This Section will outline basic features of the program.

Figure 6.1: System diagram summarising main tasks of the program



6.2.1 Working with crystallographic data

Crystallographic data is conveniently handled using the 'clipper' library (Cowtan, 2010). This library defines a large class hierarchy for object-oriented program-

ming that includes tools for reading and writing `.mtz` files. Clipper defines a wide range of objects, but only input/output and data objects will be described here. Full documentation can be found on Kevin Cowtan's website (Cowtan, 2010).

The basic information required to describe a crystal (space group, unit cell etc.) is contained within the reflection indexing class `clipper::HKL_info`. Objects of this class hold a reference to each reflection contained in a file (in the form of a `clipper::HKL_info::HKL_reference_index` object) and therefore several `clipper::HKL_info` objects are needed for data corresponding to different crystals. Upon calling the program a `clipper::HKL_info` object is constructed for each crystal by looping over input files with the function `read_mtz_info()`.

```
clipper::HKL_info read_mtz_info(string & file)
{
clipper::HKL_info hkl_list;
clipper::CCP4MTZfile mtz;
mtz.open_read(file);
mtz.import_hkl_info(hkl_list,0);
mtz.close_read();
return hkl_list;
}
```

This function also illustrates how to read `.mtz` files using clipper functions contained within the class `clipper::CCP4MTZfile`. The list of reflections is then used as an argument for the function `read_mtz_data<T>()` to create a `clipper::HKL_data` object that stores the actual data. This function creates a `clipper::HKL_data` object containing data of a type specified by contents of the angled parentheses. For example, the following segment of code uses the list of reflections created above to enter I and σ_I values into a

clipper::HKL_data object.

```
string label = get_columns(columns,"J","Q")[0][0];
clipper::HKL_data<clipper::data32::I_sigI> I_sigI;
I_sigI =
read_mtz_data<clipper::data32::I_sigI>(file,label,myhkl);
```

Data from each input file are stored in a unique MULTICRYST::CRYSTAL object held by a single MULTICRYST::CRYSTALS object. There is a single indexed entry for each crystal that contains the clipper data types found in Figure 6.2 provided these can be created from the input file. There is also a string identifier and a series of bool variables that are set according to whether each of the corresponding data types is present for that crystal. Public functions are used to initialise the MULTICRYST::CRYSTAL objects and return each data type upon request. Once a MULTICRYST::CRYSTALS object has been created from the MULTICRYST::CRYSTAL objects it is used for initialisation of the internal class structure.

6.2.2 Classes and structures

The master object through which all data must be accessed belongs to the MULTICRYST::ALL_INDICES class (Figure 6.3). Upon initialisation, the program performs a loop over all crystals to generate a clipper::HKL_info object containing a list of all reflections. It is this list that is used to reference data in the MULTICRYST::CRYSTALS object. A second loop is performed over the entire list to see which crystals have a representative for each reflection. For every reflection an object of type MULTICRYST::HKL_INDEX is created and this contains a MULTICRYST::HKL_INSTANCE object for every crystal that has an observation. The data structure MULTICRYST::HKL_INSTANCE contains an index to a clipper::HKL_info::HKL_reference_index ob-

ject and crystal number.

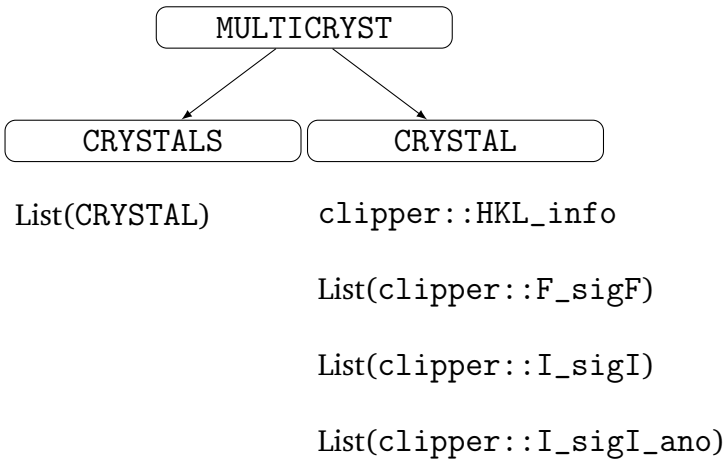
Reflections are sorted into resolution bins by appending an integer m to each `MULTICRYST::HKL_INDEX` object. Raw and normalised data relating to a particular observation can then be referenced through an index $\mathbf{h}$ and crystal number i, using the function `get_label()` to confirm an observation of $\mathbf{h}$ corresponding to that crystal really exists. As a precaution against outliers, any I or σ_I values greater than five times the interquartile range are excluded from further data processing by increasing m by the total number of bins, allowing for a simple method of outlier recovery. At this point it should be noted that when observed I's and σ_I's are missing from the input files (`bool I_sigI_exists` set to zero) their values are approximated from observed $|F|$'s and σ_F's using

$$I \approx |F|^2 \,, \qquad \sigma_I \approx 2|F|\sigma_F + \sigma_F^2 \,. \tag{6.25}$$

A normalised covariance matrix corresponding to $\mathbf{\Sigma}^0$ is created for each resolution bin in the manner described in Section 6.1.3, and the number of reflections common to crystals i and j in each bin are stored for later use.

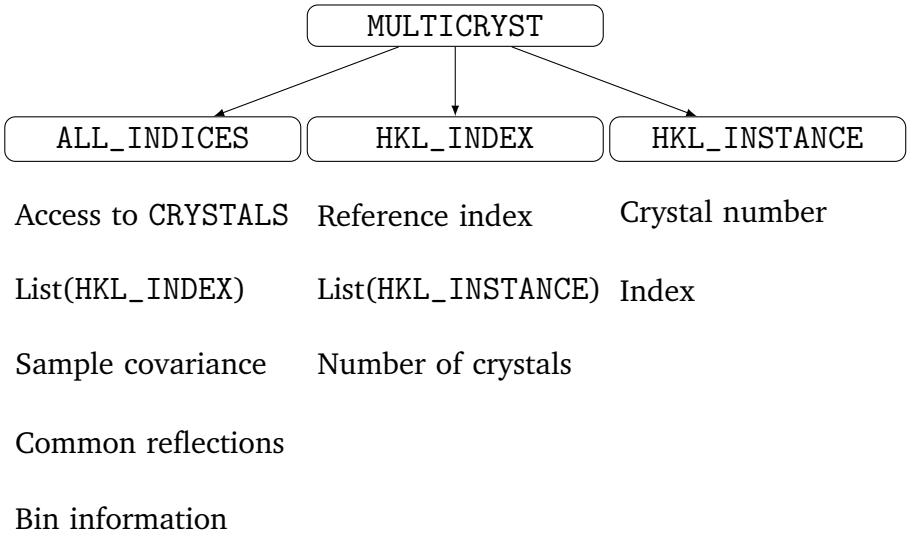
Matrices $\mathbf{\Sigma}^0$ are passed to an object of class `MULTICRYST::PARAMETERS` that is updated during the minimisation procedure (Figure 6.4). Member functions of this class are able to calculate the matrices $\hat{\mathbf{\Sigma}}$ taking a data referencing object as an argument, and perform operations on sparse matrices via the function `pack_unpack()` (Figure 6.5). This function allows all parameter matrices to remain $N \times N$ dimensional throughout the entire minimisation procedure regardless the number of crystals with representatives in each bin. Both indices $\mathbf{h}$ and m can be used to access the whole(elements) of the covariance matrix $\mathbf{\Sigma}$, m being obtained from $\mathbf{h}$ using the function `bin_number()`. For a software-inclined reader, the *C++* header file containing the class structure described here can be found in Appendix B.

Figure 6.2: Schematic of the data storage classes



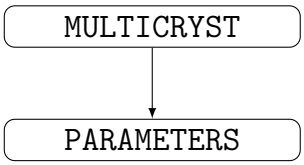
.

Figure 6.3: Schematic of the data referencing classes



.

Figure 6.4: Schematic of the parameter class



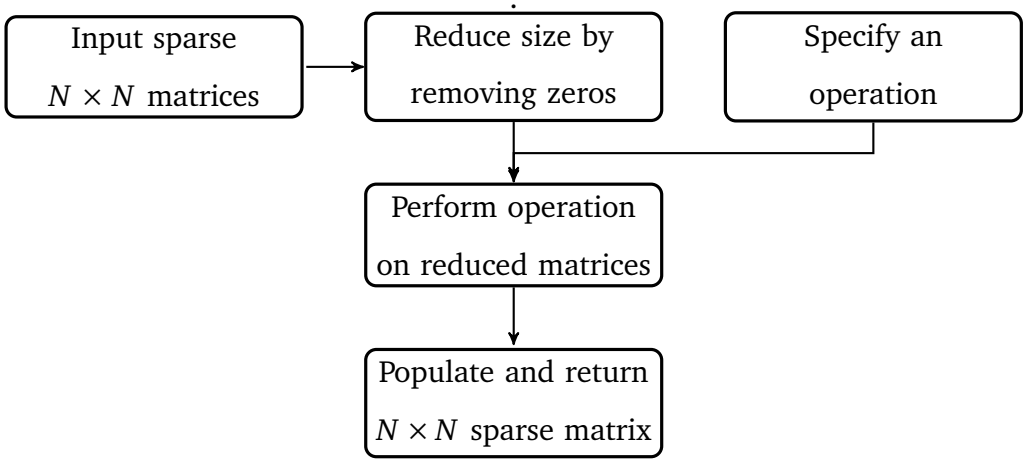
Covariance matrix

Bin information

.

Figure 6.5: System diagram for function `pack_unpack()`

.



6.2.3 Minimisation

The `MULTICRYST::PARAMETERS` and `MULTICRYST::ALL_INDICES` objects are passed to the function `ML_estimation()` that performs minimisation of LLK using quasi-Newton methods. Initially, a $\hat{\mathbf{\Sigma}}$ matrix corresponding to each reflection $\mathbf{h}$ is calculated from the initial covariance and σ_I's using the formula (6.8). Each matrix is inverted taking the Moorse-Penrose pseudoinverse (Moore, 1920; Penrose, 1955) using the `pack_unpack()` function, and the $\hat{\mathbf{\Sigma}}^{-1}$ matrices are then used with the referencing and parameter objects to cal-

culate the Fisher information and first derivatives of LLK in each resolution bin. These calculations are performed using the formulae (6.16) and (6.17) with additional terms included to ensure the covariance flows smoothly as a function of resolution. These arise from the constraint

$$\Phi = \sum_{\mathbf{h}} \sum_{ij} \sum_{m' \neq m} \lambda_{m(\mathbf{h})} \frac{((\mathbf{\Sigma}_{m(\mathbf{h})})_{ij} - (\mathbf{\Sigma}_{m'})_{ij})^2}{(m(\mathbf{h}) - m')^2} \,, \tag{6.26}$$

which is added to LLK as a sort of Kernel smoother (Nocedal and Wright, 1999). Here, $\mathbf{\Sigma}_m$ is the covariance matrix in bin m (written explicitly as a function of reflection $\mathbf{h}$) and the $\lambda_{m(\mathbf{h})}$ are positive constants, currently set to 1.0 for all bins. The first derivate of Φ is

$$\frac{\partial \Phi}{\partial (\mathbf{\Sigma}_m)_{ij}} = 2 \sum_{\mathbf{h} \in m} \sum_{m' \neq m} \lambda_m \frac{(\mathbf{\Sigma}_m)_{ij} - (\mathbf{\Sigma}_{m'})_{ij}}{(m - m')^2} \,, \tag{6.27}$$

and the second derivative is

$$\frac{\partial^2 \Phi}{\partial (\mathbf{\Sigma}_m)_{ij} \partial (\mathbf{\Sigma}_m)_{kl}} = 2 \delta_{ik} \delta_{jl} \sum_{\mathbf{h} \in m} \sum_{m' \neq m} \frac{\lambda_m}{(m - m')^2} \,. \tag{6.28}$$

In this way, refinement of each bin depends on the refinement of all others and minimisation is over all resolutions simultaneously.

With the above constraints added to LLK, minimisation proceeds via an unconstrained optimisation procedure with an additional subroutine to be described shortly. The objective of unconstrained optimisation is to attain the minimum of LLK as a function of parameters $\mathbf{\Sigma}_{ij}$ with knowledge only of LLK and its derivatives (or approximation of) at a given point. Ultimately, all optimisation algorithms search for a point at which the gradient of the target function vanishes, but differ on a choice of two fundamental strategies of moving from one point (given by $\mathbf{\Sigma}^0$, say) to a new iterate. The direction of movement or 'shift' chosen here is the Newton direction (Nocedal and Wright, 1999). This direction is obtained expanding LLK as a Taylor series to second

order in the shift $\mathbf{p}$

$$LLK(\boldsymbol{\Sigma}^0 + \mathbf{p}) \approx LLK(\boldsymbol{\Sigma}^0) + \mathbf{p}^T \nabla LLK(\boldsymbol{\Sigma}^0) + \frac{1}{2}\mathbf{p}^T \nabla^2 LLK(\boldsymbol{\Sigma}^0)\mathbf{p} \,. \tag{6.29}$$

By setting the gradient of $LLK(\boldsymbol{\Sigma}^0 + \mathbf{p})$ to zero, assuming that $\nabla^2 LLK(\boldsymbol{\Sigma}^0)$ is positive definite, the (quasi)-Newton direction is the value of $\mathbf{p}$ that minimises $LLK(\boldsymbol{\Sigma}^0 + \mathbf{p})$

$$\mathbf{p}_{n+1} = -(\nabla^2 LLK_n)^{-1} \nabla LLK_n \approx -(\mathcal{I}_n)^{-1} \nabla LLK_n \,, \tag{6.30}$$

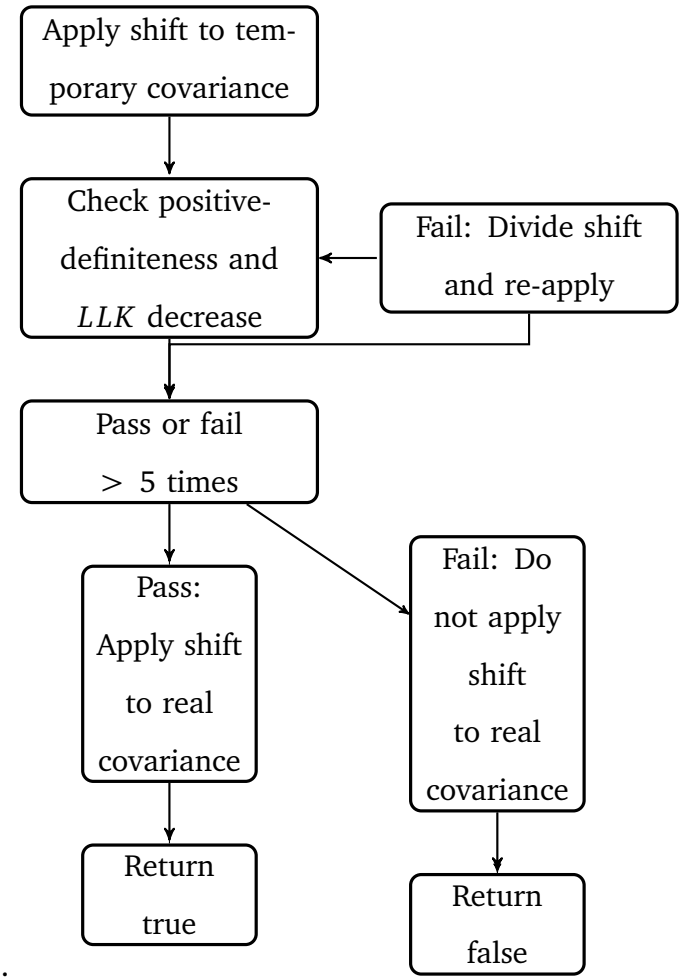
where the index n denotes the nth iteration of the optimisation process. The (quasi)-Newton method is appropriate when the shift does not cause overly large changes in LLK. Here $\mathcal{I}_n$ is used in place of $\nabla^2 LLK_n$ because calculation of the Hessian is unnecessarily cumbersome and does not always yield a positive-definite matrix.

The shift to be applied to each covariance matrix is passed to the function `subroutine()` (Figure 6.6) that is used to verify the resulting new covariance matrix remains positive definite and LLK decreases. If either of these conditions are not satisfied then the magnitude of the shift is reduced and the test is repeated, continuing this process up to a total of five times and rejecting the shift entirely if each time it fails. This subroutine is run for all bins with the results being stored as a vector of Boolean variables containing one element for each bin. Provided that at least one bin has completed the subroutine with success, the $\hat{\boldsymbol{\Sigma}}^{-1}$ matrices are updated using the new covariances and another cycle of minimisation is undertaken. The minimisation procedure is not considered to have converged until every bin fails the test simultaneously.

6.2.4 Output files

Following minimisation, each unique element of the final covariance matrices is written out to a `.txt` file and the parameter and referencing objects are then

Figure 6.6: Schematic of the function `subroutine()`



passed to the function `True_I_sigI()` for estimation of true I's and σ_I's. Using the formulae (6.23) and (6.24), these are calculated by treating outliers as observed and returning their bin labels to original values by subtracting the total number of bins from m.

The result is passed to the function `Write_out()` responsible for writing `.mtz` files. A unique file is created for each crystal and populated with the corresponding I and σ_I values by looping over all bins in which a crystal has representative reflections. Writing `.mtz` files using the clipper library is

not entirely straightforward and requires creation of a `clipper::HKL_info` object specifying space group, unit cell dimensions and resolution. The only data currently written to the file are `clipper::data32::I_sigI` objects. Firstly a list of reflections to be included in the file is created in the form of a `clipper::HKL` object and subsequently data are imported from `double` format. Files are exported and the program terminates.

6.3 Functionality demonstration

In this Section test cases are presented to assess performance of the program. The focus is on true covariance estimation and predicting values for missing data.

6.3.1 Estimation of true covariances

Four test cases were used to asses the performance of the covariance estimation algorithm.

Case i - dealing with noisy data

To demonstrate how the algorithm copes with large and noisy data sets, files containing raw intensities from structures of the 30S ribosomal subunit bound to various anticodon stem loops (Fernández et al., 2013b) were used with the permission of Israel Fernández. Figure 6.7 shows an example of the covariance between two structures differing by a single base pair in the A-site. The starting covariance $\mathbf{\Sigma}^0$ (as calculated in Section 6.1.3 via the sample covariance) tends towards zero as the bin number increases, seeming to suggest that the structures are unrelated at higher resolution. To a naive user this might indicate that little or no high-resolution information from one crystal structure could be used to infer properties of the other, but once true covariance

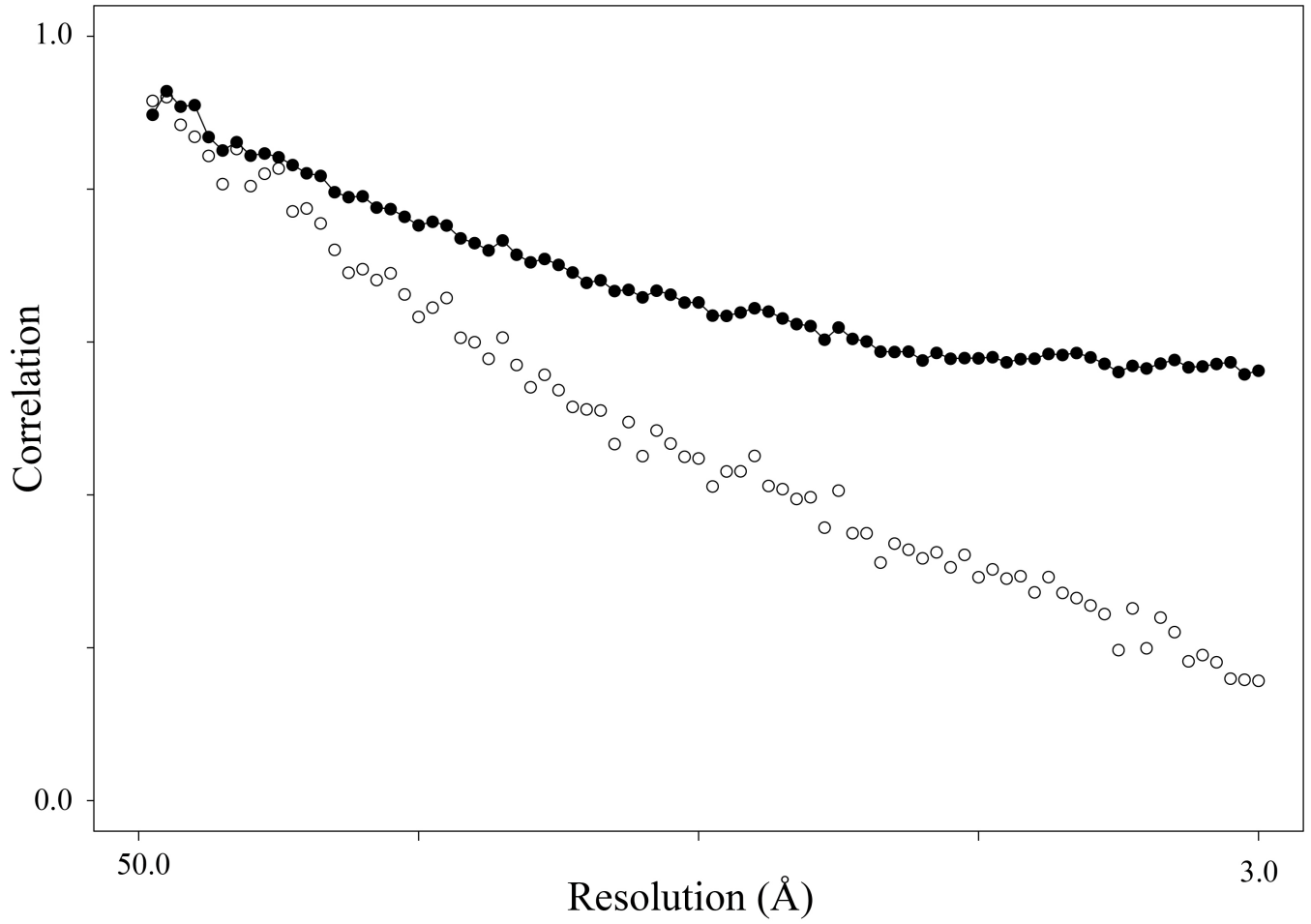


Figure 6.7: Covariance estimation from noisy data

Covariance between two related 30S ribosome crystals plotted as a function of resolution bins. The starting covariance corresponding to $\boldsymbol{\Sigma}^0$ (open circles) decays to zero rapidly as resolution increases to the right, but the true covariance corresponding to $\boldsymbol{\Sigma}$ (filled circles) reaches a plateau around 0.6.

between structure factors has been estimated correctly the actual relatedness becomes clear.

An estimate of the true covariance between structure factors demonstrates that these ribosome crystals are indeed related at higher resolution, with a correlation coefficient of ~ 0.6 in the outer-most resolution bin. This level of correlation suggests that the information contained within data from one of the crystals may be used to help solve the structure of the other, something that is not necessarily apparent from a conventional analysis of sample covariance without prior knowledge of the structures. The reason that sample covariance (or $\mathbf{\Sigma}^0$) is a poor estimation for true covariance is attributed to the fact that the ribosome is subject to low signal-to-noise ratios at higher resolution. Effectively, the covariance estimation algorithm is able to separate the true signal from noise by accounting for measurement errors in the form of the likelihood function. This will be useful for the general case of noisy data, the ribosome being just one extreme case.

Case ii - detecting structural variation

The second test case used unpublished data sets of two different protein complexes (referred to here as protein A and protein B) that are cofactors for HIV-1 invasion. These were provided by David Jacques and Leo James with the understanding that details of the structures would remain confidential. For each protein, three data sets corresponding to the same point mutations (mutants 1-3) were used to estimate true covariances between the mutant structures of both A and B. For example, the true covariances between mutants of protein A are displayed in Figure 6.8. In both cases, mutant 1 and mutant 2 were highly correlated (~ 0.8 in outer bin) whereas mutant 3 was less well-correlated with the other two (~ 0.6 in outer bin). Indeed, for both proteins, careful examination of the structures reveals mutant 3 induces a subtle conformational change

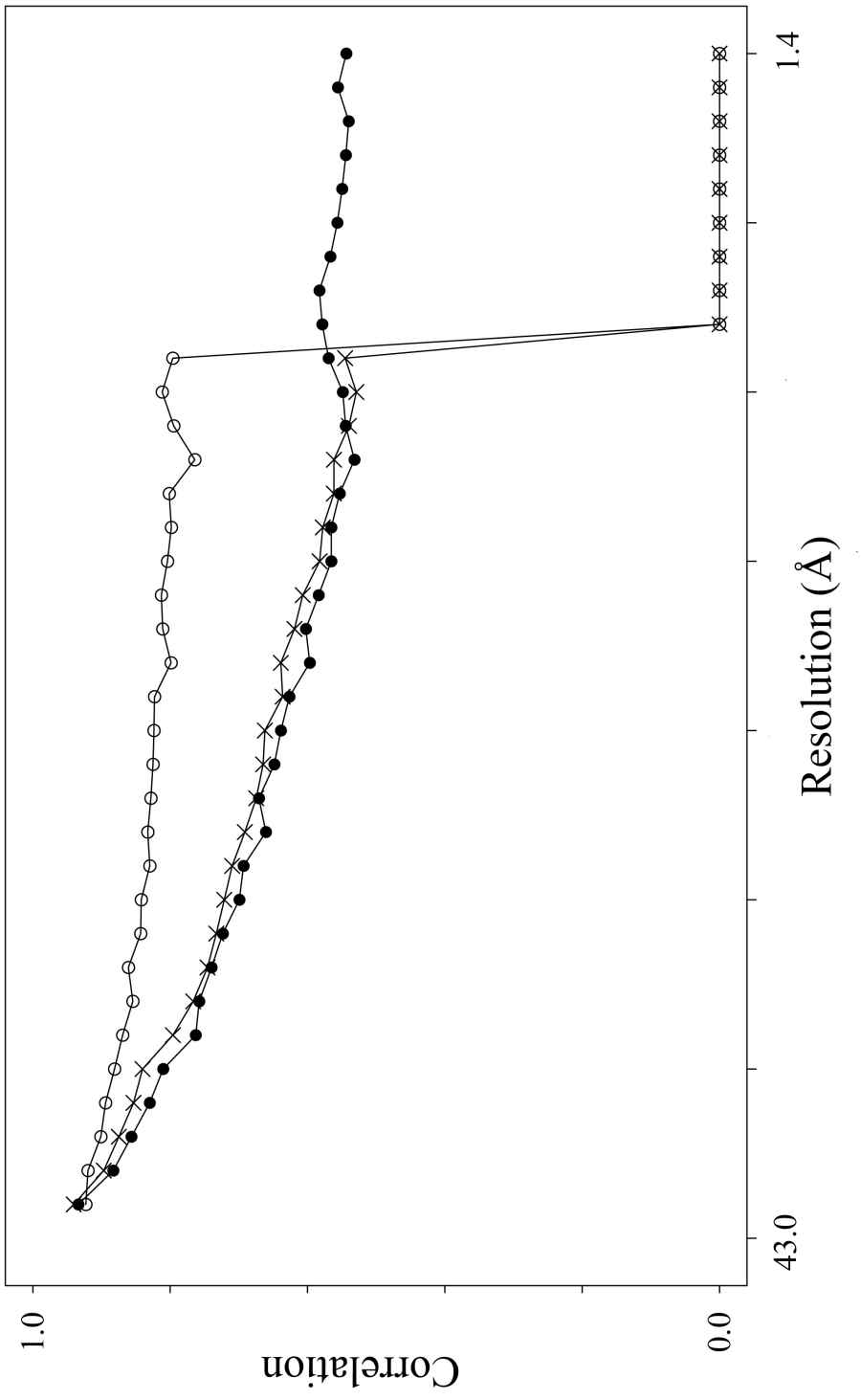


Figure 6.8: Detecting subtle conformational changes using true covariance

Covariance between three point mutants of protein A plotted as a function of resolution bins. Covariance between mutant 1 and mutant 2 is plotted as open circles, between mutant 1 and mutant 3 as filled circles, and between mutant 2 and mutant 3 as crosses.

across 5-6 residues compared with the other two mutants. In particular, the point mutation 3 alters the conformation of a loop and causes additional electron density in that region to become apparent (Jacques & James, unpublished). This case demonstrates that true covariance estimation can detect slight differences between crystals that are separate from randomly distributed coordinate error. In particular, if a user was looking to combine two data sets or interpret maps on the basis of the same structure, he/she would be better placed to do so using mutant 1 and 2 together whilst excluding mutant 3 from that analysis. In this sense true covariance could be used to cluster data sets according to the relatedness between the crystals from which they were derived. This particular application is also well-tailored towards high-throughput drug screening that relies on an advanced automation of data processing (Blundell et al., 2002).

Case iii - crystal clustering

To cluster data sets it was necessary to derive a single matrix **Cov** for the 'overall' covariance rather than many $\mathbf{\Sigma}$ matrices corresponding to different resolutions. To obtain a single value $\mathbf{Cov}_{ij}$ for the overall covariance between crystals i and j, $\mathbf{\Sigma}_{ij}$ was numerically integrated over resolution bins containing reflections common to all crystals. The approach was used to cluster four isomorphic crystals in space group $P6$. **Cov** was calculated from the estimated covariances in the manner just described, and the resulting correlation matrix was used to calculate Euclidean distances between structures using the function `dist()` in the software package *R*. The distance metric was then taken as an argument for hierarchal cluster analysis in *R*, and the results are represented by dendrograms in Figure 6.9.

In space group $P6$ there are two possible choices of indexing reflections, by (h,k,l) or $(h,k,-l)$ respectively. Figure 6.9A displays distances calculated

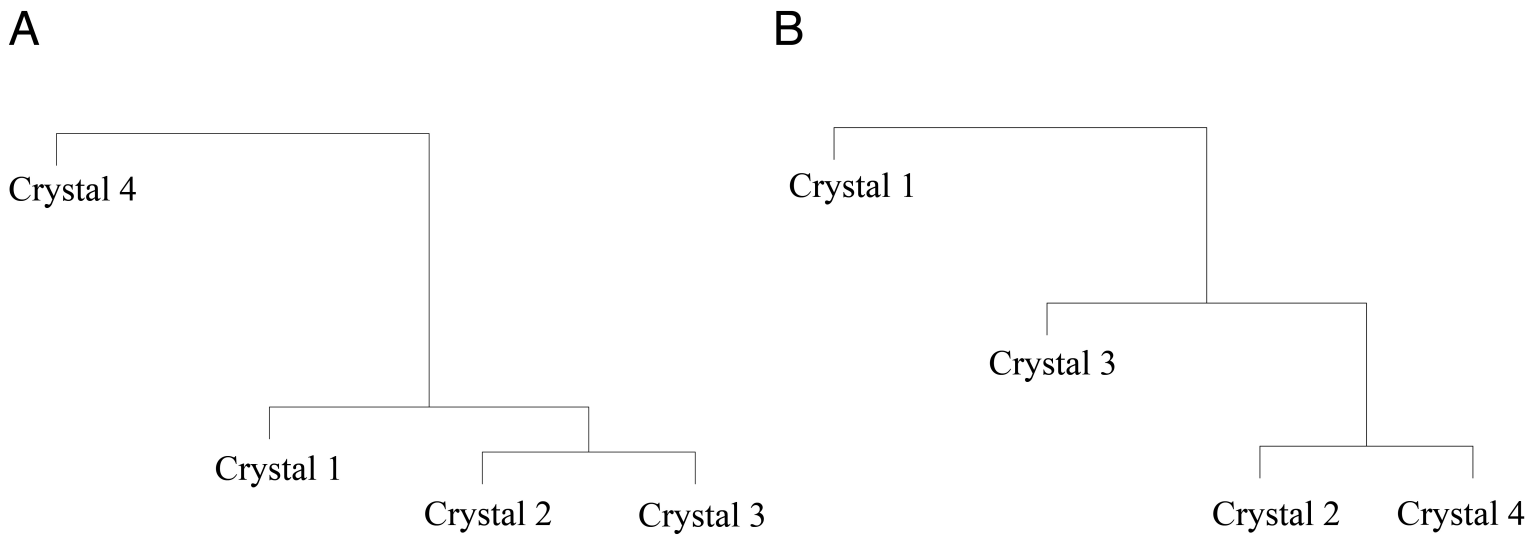


Figure 6.9: Hierarchal cluster analysis using true covariance

Dendrogram representing Euclidean distances between four crystals in space group *P*6. (A) Dendrogram calculated using a choice of indexing for crystal 4 that is inconsistent with that of the other three crystals. (B) Dendrogram of the 'true' relatedness between the four crystals as calculated using a consistent choice of indexing.

using a choice of indexing for crystal 4 that is inconsistent with crystals 1,2, 3, and results in crystal 4 being sent to a cluster far from the other crystals. When a consistent indexing scheme is used amongst all crystals however (Figure 6.9B), hierarchal cluster analysis of the overall correlation matrix reveals that crystal 4 is in fact closely related to crystal 2. The dendrogram in Figure 6.9B reveals the true distances between crystals. This example demonstrates the importance of a consistent indexing scheme when multiple options are available to the crystallographer. Whilst the choice of indexing is arbitrary when working with a single data set, an inexperienced (or even sometimes experienced) user may inadvertently compare multiple data sets using different indexing schemes and arrive at the wrong conclusion about which data to combine. True covariance will highlight what data are using different indexing schemes because overall correlation will be comparable to that expected from randomly generated data.

Case iv - dealing with incomplete data

Finally, to demonstrate how the algorithm copes with incomplete data sets, reflections chosen at random were deleted from both of the two highly correlated data sets from protein A mutants 1 and 2. The program was used to estimate again the true covariance between protein A mutants 1,2, and 3, using 100%, 70%, or 50% of the total data from mutant 1 and mutant 2. The covariance between mutant 1 and mutant 2 resulting in each case is plotted in Figure 6.10 that shows the algorithm returns the same covariance values regardless the portion of data missing. This trend was reproduced for all possible combinations of the test cases described above (data not shown). Covariance estimation from incomplete data is more representative of the general case that would be encountered by an experimenter having to collect data from multiple crystals, where radiation damage prevents collection of an en-

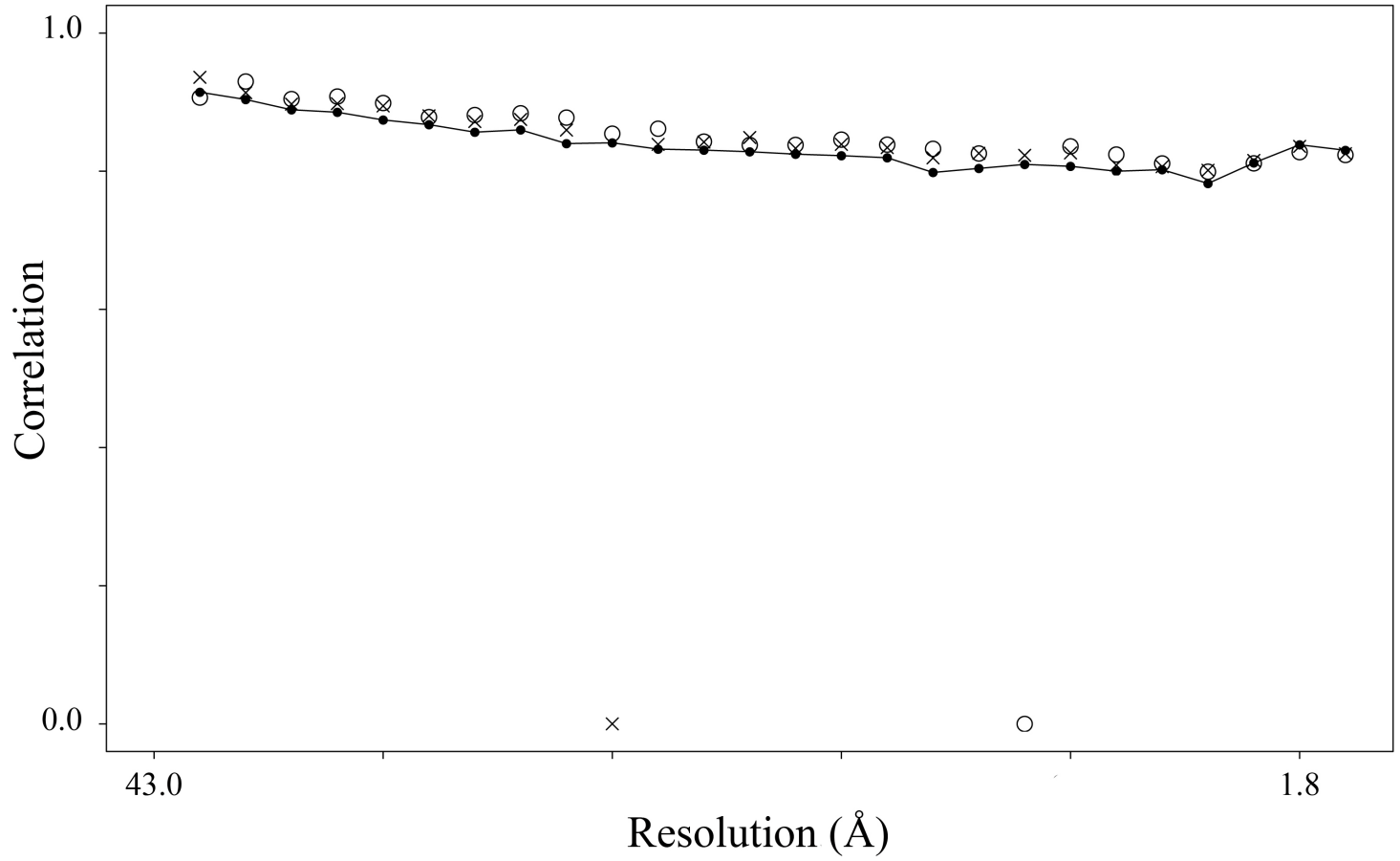


Figure 6.10: Robust covariance estimation from incomplete data

Covariance between point mutants 1 and 2 of protein A plotted as a function of resolution bins. Covariance estimated using 100% of the data set is plotted as filled circles, using 70% of the data as crosses, and using 50% of the data as open circles. Zero values arise due to a lack of common reflections in that resolution bin.

tire set from any one crystal. Besides, the real power of the algorithm is to give a measure of relatedness between crystals without ever needing to solve a structure. The fact that the algorithm remains robust in these situations makes it well-suited for general use during the data collection process.

6.3.2 Prediction of true intensities

To test how accurately the formulae outlined in Section 6.1.3 estimate true intensities and variances, output files containing predicted I and σ_I values were converted to files containing $|F|$'s and σ_F's using the program CTRUNCATE that also runs a script to append an R free flag to 5% of reflections (Stein and Ballard, 2009). For each data set tested, molecular replacement with PHASER (McCoy et al., 2007) was completed using the appropriate search model (not the finally solved structure) and in each case PHASER returned a single solution that matched with those obtained from native data sets. The solutions were refined using 25 cycles of restrained refinement in REFMAC and then statistics and maps were assessed for a measure of performance.

Examples using data from three structures of the same protein bound to different ligands (provided by Jacques & James) are displayed in Table 6.1, where the program was run each time with data corresponding to the undisclosed ligands 1, 2, and 3. Table 6.1 shows the performance of native data compared with predicted data under the same structure solution protocol, using either 100%, 70%, or 50% of the native data from each crystal in the analysis to predict true intensities and variances. Data were removed at random using the program SFTOOLS, written by Bart Hazes. Data predicted using 100% of the native data performed similarly to the native data themselves, and in each case maps and R free values indicated the program successfully predicted a more complete data set for each crystal from as little as 50% of

the total observed reflections. Although *R* free values increase slightly as the number of observed reflections decreases this is expected because predicted data will never truly substitute for observed data. These values remain within or just slightly above a ∼5% margin, suggesting the information content of predictions is comparable to that of the native data.

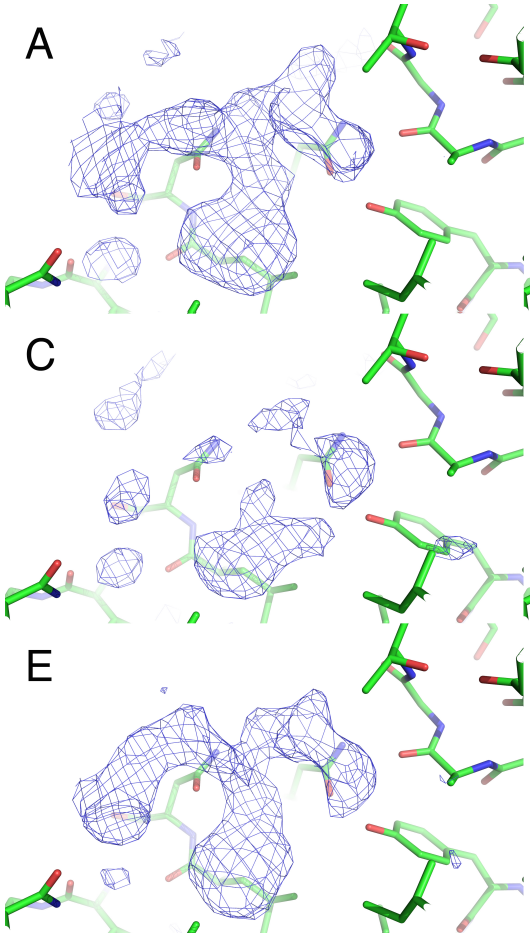
Data prediction relies on true covariance to account for inter-crystal differences, and so predicted data should contain enough information to discern structural features that are unique to a particular crystal. To test whether this was indeed the case, complete data were predicted for each of the ligand-bound crystal structures using only 50% of their native data. Complete data from one of the other two ligands or the Apo crystal structure were used alongside the 50% wedge to estimate the remaining intensities and variances, and all possible combinations are presented in Figures 6.11, 6.12 and 6.13. Of the test cases presented, predicted data for ligand 3 performs significantly better than 50% native data alone since the ligand is barely visible when the latter was used for refinement (Figure 6.13B). A data set completed taking into account Apo data (Figure 6.13C) yields a difference map for ligand 3 that would enable the ligand to be identified and refined (Figure 6.13A). This demonstrates that information from highly correlated structures can be used to infer missing data for others. It follows that difference maps are less revealing when generated from data predicted for ligand 3 using ligands 1 or 2 (Figures 6.13D and E, respectively), since these two crystals have a lower overall correlation with the first (Figure 6.13F).

In the cases of ligands 1 and 2, 50% native data is already sufficient to reveal the presence of either ligand and so predicted data does not prove to be as useful. Although noisy, data predicted using the highly-correlated Apo data (Figure 6.12C) performs better than that predicted using data from ligand 1

	Completeness (%)	**TFZ**	**LLG**	*R* **value**	*R* free **value**
Native data					
Ligand 1	99.67	11.0	158	25.83	30.12
Ligand 2	93.7	9.7	151	26.76	32.32
Ligand 3	94.3	9.0	171	25.50	29.03
Predicted from 100%					
Ligand 1	98.5	11.0	159	25.89	30.45
Ligand 2	99.8	11.1	162	26.75	32.68
Ligand 3	98.3	9.5	167	26.21	30.07
Predicted from 70%					
Ligand 1	91.1	9.4	147	27.29	31.85
Ligand 2	93.9	9.5	149	27.37	33.31
Ligand 3	77.3	9.0	159	27.47	34.24
Predicted from 50%					
Ligand 1	79.9	5.4	122	28.19	34.42
Ligand 2	84.1	4.7	147	27.79	33.43
Ligand 3	70.5	4.9	143	28.42	31.02

Table 6.1: Statistics of structure solution using predicted data

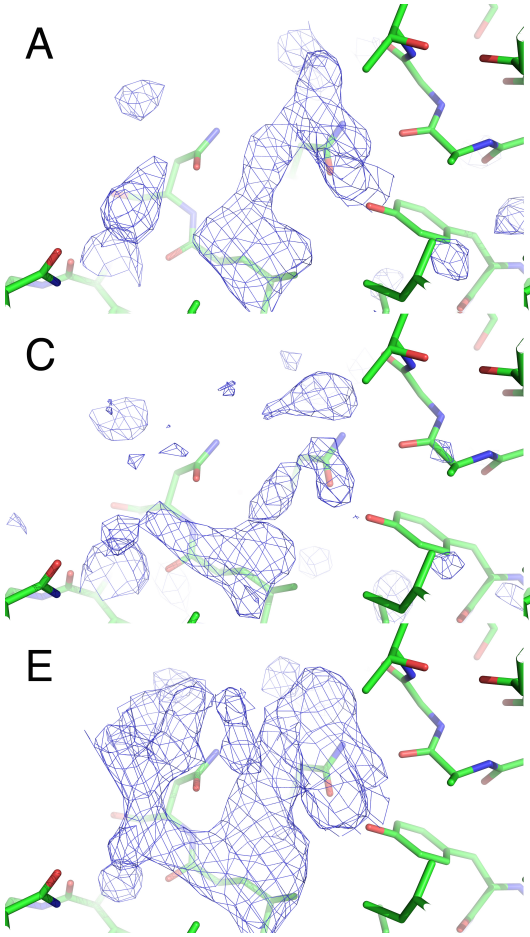
Summary of statistics from PHASER and REFMAC using a selection of file types for structure solution. TFZ = translation function Z-score, LLG = log-likelihood gain.



	Apo	1	2	3
Apo	1.000	0.879	0.927	0.907
1	0.879	1.000	0.857	0.750
2	0.927	0.857	1.000	0.864
3	0.907	0.750	0.864	1.000

Figure 6.11: Visualisation of ligand 1 in difference maps calculated from predicted data

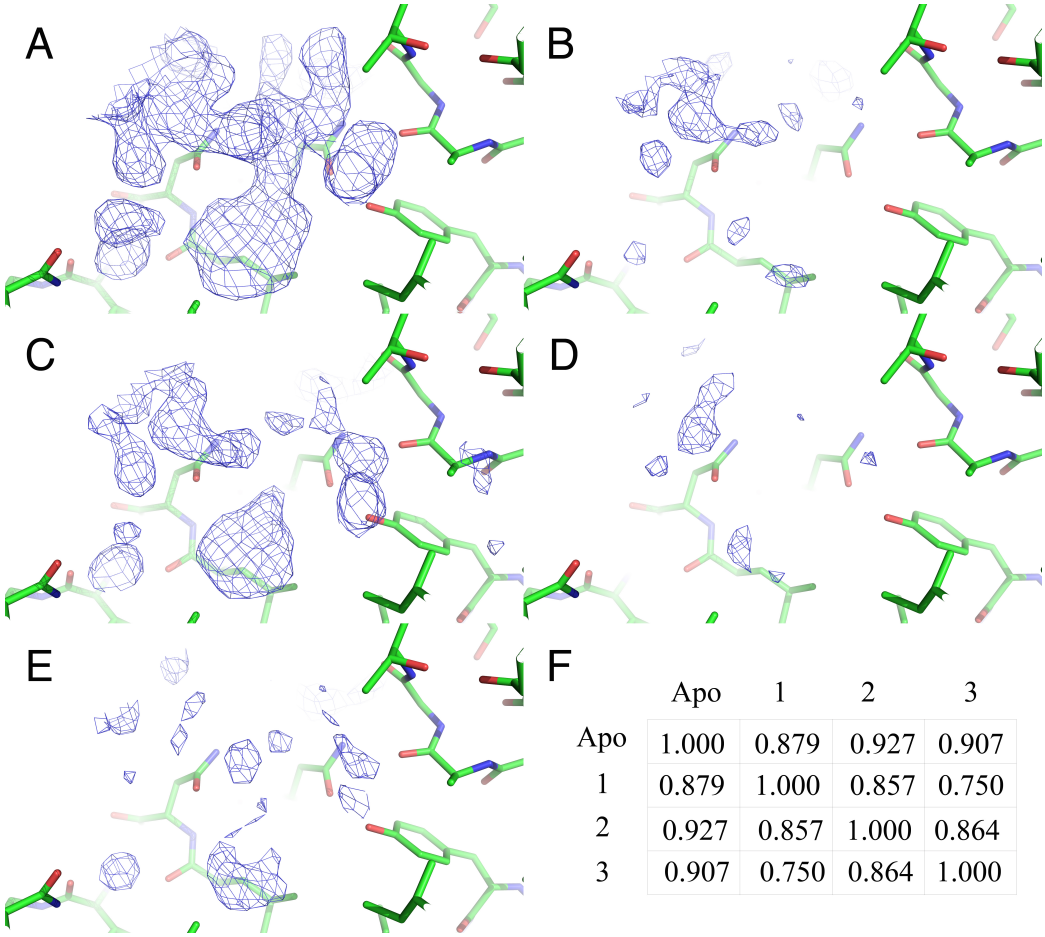
The $F_o - F_c$ maps generated following refinement of the same model using predicted or native data are contoured at 2σ. Data used in refinement were (A) 100% native corresponding to ligand 1, (B) 50% native corresponding to ligand 1, (C) predicted using 50% native and data corresponding to Apo, (D) predicted using 50% native and data corresponding to ligand 2, and (E) predicted using 50% native and data corresponding to ligand 3. (F) Overall correlation between data sets.



	Apo	1	2	3
Apo	1.000	0.879	0.927	0.907
1	0.879	1.000	0.857	0.750
2	0.927	0.857	1.000	0.864
3	0.907	0.750	0.864	1.000

Figure 6.12: Visualisation of ligand 2 in difference maps calculated from predicted data

The $F_o - F_c$ maps generated following refinement of the same model using predicted or native data are contoured at 2σ. Data used in refinement were (A) 100% native corresponding to ligand 2, (B) 50% native corresponding to ligand 2, (C) predicted using 50% native and data corresponding to Apo, (D) predicted using 50% native and data corresponding to ligand 1, and (E) predicted using 50% native and data corresponding to ligand 3. (F) Overall correlation between data sets.



	Apo	1	2	3
Apo	1.000	0.879	0.927	0.907
1	0.879	1.000	0.857	0.750
2	0.927	0.857	1.000	0.864
3	0.907	0.750	0.864	1.000

Figure 6.13: Visualisation of ligand 3 in difference maps calculated from predicted data

The $F_o - F_c$ maps generated following refinement of the same model using predicted or native data are contoured at 2σ. Data used in refinement were (A) 100% native corresponding to ligand 3, (B) 50% native corresponding to ligand 3, (C) predicted using 50% native and data corresponding to Apo, (D) predicted using 50% native and data corresponding to ligand 1, and (E) predicted using 50% native and data corresponding to ligand 2. (F) Overall correlation between data sets.

or ligand 3, where additional features have been introduced to the difference map of the second (Figure 6.12E). This highlights one limitation of data estimation is that local biases from reference data may be wrongly interpreted, which is best demonstrated in the case of ligand 1. Although ligand 1 has overall lowest correlation with ligand 3, it is by taking ligand 3 data into account that predictions appear to perform nearly as well as the native data. Closer inspection reveals that ligands 1 and 3 are very similar in shape and so the fact that this data appears to perform better is probably because of local contributions from the complete data set. A user should be careful to select a reference data set that contains as little bias possible depending on their chosen intention. For example, if attempting to confirm the presence of a ligand it would be wise to use Apo data as a reference in order to avoid false positives. Overall correlation should be used to guide the selection.

The results presented here show that in certain cases large numbers of incomplete data wedges from different crystals could be used in conjunction to 'complete' each data set individually. The approach is very different from conventional scaling that assumes a single underlying crystal structure and combines all data into a single set of reflections, obviously not a sensible thing to do when the crystals contain different ligands or protein conformations. Provided that the correlation between crystals is sufficiently high a different output file could be created for every crystal and populated with a different set of reflections, predicted taking into account observations from all crystals and the relatedness between them.

6.4 Discussion

6.4.1 Implications

The ability to obtain the true covariance between crystals from noisy diffraction data is non-trivial without prior knowledge of the structures. The likelihood method presented here provides a tool for estimation of true covariances from noisy, incomplete data sets that has a wide range of applications. As determination of large and complicated structures becomes more and more routine, it becomes increasingly common to rely on poorly diffracting, radiation-sensitive crystals for structure solution (Fry et al., 1999; Brodersen et al., 2003; Carpenter et al., 2008). In these cases the likelihood method can assist with analysis of multiple data sets by providing an accurate measure of the relatedness between data that a user will attempt to combine. With true covariance at hand the crystallographer is well-placed for selecting portions of data that will be compatible and rejecting those that are not.

The results presented in Section 6.3.1 also demonstrate the ability of the likelihood method to cluster crystals according to underlying structural features. A powerful application of the algorithm is to identify subtle changes between crystals that are separate from general coordinate error or experimental noise. In practice, this would enable a crystallographer to refine different structures accordingly by indicating which portions of data correspond to each of the alternatives to be refined. For cases where unit cell dimensions are similar, true covariance will provide the a way of allocating incomplete data from multiple crystals to a cluster where each batch of images cannot be used alone to solve a structure. This would be particularly useful for time-resolved crystallography (Hajdu et al., 2000; Schotte et al., 2004) where, for example, conformational changes in DNA polymerase have been followed over a reac-

tion period using data from 15 different crystal structures (Freudenthal et al., 2013). As suggested in Section 6.3.1, potential applications extend beyond detection of conformational changes to distinguishing empty crystals from those that have ligands bound. For high-throughput drug discovery (Blundell et al., 2002), batches of images that are insufficient in number could be quickly assessed for the presence of ligand without ever having to process data further.

Whilst true covariance could be used to devise a strategy for a conventional scaling approach, an alternative route would be using the algorithm to generate a set of predicted data for each of the crystals included in the analysis. This would allow the user to refine a structure for each sub-structure without the requirement for a complete native data set. The scenarios described above apply just as well to cases where data from different crystals are incomplete and prediction can be used to identify ligand-bound crystals or conformational changes in a protein. As emphasised several times throughout this Chapter, by accounting for differences between crystals the likelihood approach has an advantage over standard multi-crystal protocols that assume a single underlying crystal structure and therefore do not exploit the full amount of information contained within data. This is exemplified by the results presented in Section 6.3.2, which show that by treating data separately but taking into account true covariance, three different structures can be solved by using data sets that alone are insufficient to solve one.

6.4.2 Future developments

At present the data prediction algorithm performs well in certain cases, but the estimated I's and σ_I's remain only a rough approximation to their true values and the algorithm has not been optimised for using data less than ~50% complete. One way to improve data prediction further would be to

devise a better prior distribution for I and σ_I^2. An alternative option would be to implement an expectation maximisation (EM) algorithm (Dempster et al., 1977) based on the distribution (6.9). EM algorithms specialise in obtaining a maximum likelihood estimate for parameters of a statistical model that depends on unobserved data. An EM algorithm will typically iterate between an estimation step, where expectation values of the likelihood function are used to predict values of missing data, and a maximisation step that maximises the log-likelihood on the basis of the (present and predicted) data. Once new parameter values have been obtained they are used to predict new values for the missing data, and the cycle continues until convergence. In this way missing I^o values could be predicted for incomplete data sets alongside minimisation of (6.9), a procedure that may even result in a more robust estimation of the true covariance.

Perhaps the most obvious extension of this work is to the simultaneous refinement of different structures using unmerged data collected from multiple crystals. An intensity-based likelihood function for refinement can be constructed using a similar approach to that outlined in Section 6.1.1. This would also include refining different 'states' of the same crystal at different time-points during data collection, the user choosing, for example, to refine a different structure corresponding to the state of the crystal each time a diffraction image was taken. The ensemble of structures would then represent changes in the crystal as a function of time on the beam line. It must be made clear that the approach would not be to simply assign data to different structures and refine each independently, but rather account for all data in a probability density function depending on all states of that crystal. For example, a single state could be selected for refinement by integrating out all others, the likelihood function giving the probability of observing the entire data set conditionally

on that one structure.

Constructing a likelihood function that accounts for different states of a crystal would reduce the information loss incurred during conventional scaling, which assumes a single state gave rise to all data. Changes in the crystal can be modelled as a Markov process, with structure factors corresponding to *hidden* variables in state space and intensities as variables in observation space. When accounting for changes in the crystal structure the probability $p(F_i; F_j)$ can be interpreted as a transition function giving the probability for a crystal in state j to go to state $i > j$ within a given time interval. Together with $p(I_i^o; I_i)$ representing the measurement process, these distributions define a hidden Markov process beginning in an initial state F_0 with probability $p(F_0)$. A little manipulation shows the approach outlined in Section 6.1.1 generalises quite naturally to this case. A likelihood function suitable for refinement would take into account intensities I_j^c calculated from the crystal structure j, weighted appropriately by a factor $D_{i,j}$ that describes the expected value of the observed intensity I_i^o from crystal i. Depending on the precise form of expectation values, a Gaussian approximation could be constructed for $p(\{I_i^o\}; \{I_i^c\})$ using the method described above.

Modelling changes in the crystal over time would be particularly well-suited to accounting for radiation damage during refinement. This would potentially solve some of the problems faced when working with radiation-sensitive crystals (Holton, 2009; Garman, 2010). In an ideal scenario there would be no requirement for the user to combine or cut data from different crystals manually, instead passing all data to a pipeline that first clusters image batches on the basis of a true covariance estimation. For clusters that differed dramatically a different atomic model would be required for refinement, but in general a single atomic model could be refined using a likelihood function

that accounts for all possible states populated by the different crystals during data collection. At no point would a single state be assumed to have given rise to all data. With help of the likelihood method, problems encountered when merging data from multiple crystals (Riekel et al., 2005) may be significantly reduced.

Appendix A

Materials and methods

A.1 Preparation of chemicals and reagents

All chemicals were of analytical grade where possible, except where stated and were purchased from Sigma™. Hi-Trap™metal chelating columns and HiPrep™QXL Sepharose™columns were purchased from GE Healthcare. Crystallisation plates used for sitting drop vapour diffusion were MRC plates purchased from Innovadyne Technologies, Inc, tape for sealing the crystallisation trays were obtained from Hampton Research Ltd, California, USA. For cryogenic crystallography crystals were mounted in Litho™loops from Molecular Dimensions Ltd. Prior to data collection, frozen crystals were transferred into pucks, with all magnetic caps and vials being SPINE standard and purchased from Hampton Research Ltd. tRNAPhe and ribosomes from *Thermus thermophilus* harbouring a C-terminal truncation of protein L9 were prepared by Ann Kelley as previously described (Selmer et al., 2006) and the 30S subunit was purified as reported in (Wimberly et al., 2000). mRNA with the sequence 5'-GGCAAGGAGGUAAAAAUGUUCAAAA-3' was purchased from Dharmcon (Thermo Scientific). de-6-MSA-pactamycin was chemically

synthesised in the laboratory of S. Hanessian and transported as a resin, the solubility being 1.24 mg in 10% dimethyl sulfoxide (DMSO).

C-terminally His-tagged EF-G from *T. thermophilus* was over-expressed in *Escherichia coli* strain BL21 DE3 using the T7 vector pET-16 (Novagen). Cells were harvested and frozen prior to protein purification, where cells were thawed in buffer A (100 mM HEPES KOH-pH 7.5, 150 mM NaCl, 20 mM imidazole, 20 mM $MgCl_2$) supplemented with phenylmethanesulfonylfluoride (PMSF). All purification buffers also contained 6 mM β-mercaptoethanol. The suspension was homogenised using an Emulsiflex (Avestin, Ottawa, Canada) and the cell debris removed by centrifugation at 25 000 rpm for 20 min. The supernatant was heated to 65°C to denature endogenous proteins and centrifuged once more at 25 000 rpm for 20 min. The sample was passed through a 12 μm filter to remove debris not cleared by centrifugation and loaded onto a Hi-Trap Nickel Sepharose column (Amersham Biosciences) equilibrated in buffer A. The column was washed in buffer B (100 mM HEPES KOH-pH 7.5, 150 mM NaCl, 50 mM imidazole, 20 mM $MgCl_2$) to remove protein bound non-specifically to the column resin and EF-G was eluted with buffer C (100 mM HEPES KOH-pH 7.5, 150 mM NaCl, 300 mM imidazole, 20 mM $MgCl_2$) on an AKTA Purifier (Amersham Biosciences). Pooled fractions were dialysed overnight against buffer D (100 mM HEPES KOH-pH 7.5, 20 mM NaCl, 20 mM $MgCl_2$), or buffer E (100 mM HEPES KOH-pH 7.5, 150 mM NaCl, 20 mM $MgCl_2$) for the domain IV/V mutant, supplemented with TEV protease prior to ion exchange chromatography. The dialysed sample was re-loaded onto the Hi-Trap Nickel Sepharose column (to remove the His-tag and uncleaved protein), then onto a HiPrep^{TM}QXL Sepharose^{TM}column, both equilibrated in buffer D. EF-G was eluted against a linear gradient of buffer F (100 mM HEPES KOH-pH 7.5, 1 M NaCl, 20 mM $MgCl_2$) and a single peak was collected, con-

centrated, and transferred to buffer G (5 mM HEPES KOH-pH 7.5, 50 mM KCl, 10 mM NH_4Cl) prior to gel filtration on a HiLoad 26/60 Superdex 200 prep-grade column (Amersham Biosciences). A single peak was collected and judged to be >99% pure by SDS-PAGE.

SDS Polyacrylamide Gel Electrophoresis (SDS-PAGE) was carried out using pre-poured 4-12% acrylamide NuPAGE®gels. Samples were prepared prior to electrophoresis by adding sample buffer (2× buffer consists of: 200 mM Tris pH 6.8, 20% *v/v* glycerol, 4% *w/v* SDS, 0.05% *w/v* bromophenol blue, 5% *v/v* β-mercaptoethanol) and heating to 95°C for 10 minutes before loading. Samples were loaded onto the gel against a protein ladder consisting of proteins of known molecular weight and ran at 190 V for approximately 40 min in SDS-PAGE running buffer (10× SDS-running buffer: 3% *w/v* Tris base, 14% *w/v* glycine, 1% *w/v* SDS). Gels were stained using 0.4% *w/v* Coomassie Brilliant Blue R250, 50% *v/v* methanol, 10% *v/v* acetic acid for 5 min after warming in the microwave. Gels were subsequently destained in 10% *v/v* acetic acid for 30 min.

A.2 EF-G binding assay and complex formation

Ribosomes (4.0 μM) and mRNA (8.0 μM) were incubated at 55°C for 6 min in buffer G before addition of tRNA. Either tRNAfMet (16.0 μM) or dH_2O was added for incubation at 55°C for 20 min and then either tRNAPhe (16.0 μM) or dH_2O was included for a further 20 min. Separately, EF-G (20.0 μM) was incubated with GDPCP or GTP (100 μM) for 20 min at 37°C and then combined with the ribosome complex for a final incubation at room temperature for 20 min. The individual samples, each with a final volume of 70 μL, were then layered on top of a 1.1 M sucrose solution in buffer G supplemented with

100 μM GDPCP or GTP accordingly. These were then subjected to ultracentrifugation at 45 000 rpm for 4 hours at 4°C. Following ultracentrifugation, unbound protein contained within the supernatant was removed and the ribosomes contained within the pellet were resuspended in buffer G along with any EF-G that had been retained in complex. Resuspended samples were then analysed by SDS-PAGE.

For crystallisation, ribosomes (4.0 μM) and mRNA (8.0 μM) were incubated at 55°C for 6 min before addition of tRNAPhe (16.0 μM) and a further incubation at 55°C for 20 min. Separately, EF-G (20.0 μM) was incubated with GDPCP (6.0 mM) for 20 min at 37°C and mixed with the ribosome complex for a final incubation at 37°C for 20 min in buffer G. Immediately prior to crystallisation, the detergent HEGA-9 was added (46 mM). All concentrations refer to the final values in the sample. Typical total sample volumes used for crystallisation experiments did not exceed 500 μL.

A.3 Crystallisation, data collection and structure solution

The 30S ribosomal subunit from *T. thermophilus* was crystallised using the method described previously (Wimberly et al., 2000). Crystals of diffraction quality were transferred to a cryo-protectant (100 mM MES KOH-pH 6.5, 200 mM KCl, 75 mM NH_4Cl, 15 mM $MgCl_2$, 26% *v/v* 2-Methyl-2,4-pentanediol) containing a mixture of 1 mM de-6-MSA-pactamycin and 100 μM paromomycin to improve resolution. Diffraction data were collected from a single crystal that diffracted beyond 3.1 Å on the IO4 beam line at the Diamond Light Source, Harwell, England. Diffraction images were integrated and scaled using the XDS package (Kabsch, 2010a) prior to a round of restrained refine-

ment in REFMAC5 (Murshudov et al., 2011) with the empty 30S structure as a starting model. Each initial refinement was followed by alternating cycles of model building in COOT (Emsley et al., 2010) and automated refinements using jelly-body restraints in REFMAC5. At each stage of refinement electron density for ligands could be clearly identified in the unbiased difference maps, but ligand atomic coordinates were not included until the final round of refinement where a de-6-MSA-pactamycin molecule and seven paramomycin molecules were placed with confidence into the electron density map. The final model had an R/Rfree ratio of 18.4/22.7. Coordinates and structure factors have been deposited in the Protein Data Bank with accession code 4KHP and a full summary of data collection and refinement statistics are displayed in Table A.1.

Crystals of the EF-G-mRNA-tRNAPhe-70S complex were grown via streak seeding using tungsten wire and vapour diffusion in sitting drop trays by mixing 3 μL of sample with 3 μL reservoir solution (100mM MES KOH-pH 6.3, 75mM KCl, 6.0-6.5% *w/v* PEG 20K). Crystals of plate morphology grew to full size (∼200 μm by 100 μm by 50 μm) over a period of three weeks and were cryo-protected in a step-wise fashion by sequentially increasing the concentrations of PEG 20K and PEG 400 in the crystallisation buffer to 6.8% and 30% respectively, while maintaining the concentration of other components. Crystals were plunged into liquid nitrogen and stored until data collection. Two independently complete sets of data were collected from single crystals on beam line ID 14-4 at the European Synchrotron Light Source (McCarthy et al., 2009) and on beam line IO4 at the Diamond Light Source, Harwell, UK, respectively. Data were integrated, merged and scaled using XDS (Kabsch, 2010a), and found to be consistent with space group $P2_1$ and unit cell dimensions $a = 201.58$ Å, $b = 241.65$ Å, $c = 305.80$ Å and $\beta = 99.48°$.

Molecular replacement was performed using MOLREP (Vagin and Teplyakov, 2010) in two stages, first with the 50S subunit of the 70S *T. thermophilus* structure (Selmer et al., 2006) as a search model, followed by inclusion of the 30S. The solution showed a single ribosome in the asymmetric unit in the fully rotated conformation. Refinement was carried out in alternating cycles of automated refinements using either PHENIX (Adams et al., 2010) or REFMAC5 (Murshudov et al., 2011), with manual refinement and model building in COOT (Emsley et al., 2010). Coordinates and structure factors have been deposited in the Protein Data Bank with accession codes 4JUW and 4JUX; data collection and refinement statistics are displayed in Table A.2. All figures were generated using PyMOL (Schrödinger, LLC, 2010) or Jalview for sequence alignments (Waterhouse et al., 2009).

	30S-de-6-MSA-pactamycin-paromomycin
Data collection	
Beam line	IO4 (DLS)
Space group	$P4_12_12$
Cell dimensions	
a, b, c (Å)	402.24, 402.24, 177.32
Resolution (Å)	29.7-3.1 (3.2-3.1)
R_{sym} (%)	13.7 (69.1)
I/σ_I	5.9 (1.5)
Completeness (%)	91.4 (95.1)
Refinement	
Resolution (Å)	29.7-3.1
No. unique reflections	222,956
R/Rfree	18.4/22.7
No. atoms	52,042
Mean B value	81.05
Bond length r.m.s.d. (Å)	0.010
Bond angles r.m.s.d. (°)	1.743

Table A.1: Data collection and refinement statistics for the structure of de-6-MSA-pactamycin-paromomycin bound to the 30S ribosomal subunit

Values in parentheses are for outer resolution bin.

	Data set 1	Data set 2	70S-tRNA-EF-G-GDPCP
Data collection			
Beam line	ID14-4 (ESRF)	IO4 (DLS)	
Space group	$P2_1$	$P2_1$	$P2_1$
Cell dimensions			
a, b, c (Å)	203.42, 243.05, 309.97	201.58, 241.65, 305.80	201.58, 241.65, 305.80
α, β, γ (°)	90.00, 99.53, 90.00	90.00, 99.48, 90.00	90.00, 99.48, 90.00
Resolution (Å)	39.6-3.1 (3.2-3.1)	39.6-2.9 (3.0-2.9)	39.6-2.9 (3.0-2.9)
R_{sym} (%)	17.8 (58.7)	24.1 (138.4)	22.4 (137.8)
I/σ_I	8.2 (2.4)	4.9 (1.1)	7.6 (1.0)
Completeness (%)	99.6 (97.1)	97.8 (98.9)	99.8 (98.9)
Refinement			
Resolution (Å)			39.6-2.9
No. unique reflections			635,092
R/Rfree			19.6/24.5
No. atoms			150,122
B values			
RNA			37.8
Protein			51.4
Bond length r.m.s.d. (Å)			0.004
Bond angles r.m.s.d. (°)			1.488

Table A.2: Data collection and refinement statistics for the structure of an intermediate state of translocation

Both sets of data 1 and 2 were scaled together with unit cell parameters from data set 2. Values in parentheses are for outer resolution bin.

Appendix B

Header files

B.1 Class structure

The *C++* header file `intensity_multicryst.h` is included here for reference when reading Section 6.2.2.

```
namespace MULTICRYST
{
class CRYSTAL
{
private:
string filepath;
clipper::HKL_info crystal_info;
clipper::HKL_data<clipper::data32::F_sigF> F_sigF;
clipper::HKL_data<clipper::data32::I_sigI> I_sigI;
clipper::HKL_data<clipper::data32::I_sigI_ano> I_sigI_1;
clipper::HKL_data<clipper::data32::I_sigI_ano> I_sigI_2;
bool F_sigF_exists;
bool I_sigI_exists;
bool I_sigI_ano_exists;
public:
```

```
void init(string &);
clipper::HKL_info& info();
clipper::HKL_data<clipper::data32::F_sigF>& FsigF();
clipper::HKL_data<clipper::data32::I_sigI>& IsigI();
clipper::HKL_data<clipper::data32::I_sigI_ano>& IsigI1();
clipper::HKL_data<clipper::data32::I_sigI_ano>& IsigI2();
bool is_FsigF();
bool is_IsigI();
bool is_IsigIano();
void display();
string get_filepath();
};
class CRYSTALS
{
private:
vector<CRYSTAL> crystals;
public:
void add(string &);
unsigned int size();
inline CRYSTAL& operator[](unsigned int);
};
struct HKL_INSTANCE
{
unsigned int crystal;
unsigned int state;
int index;
};
class HKL_INDEX
{
private:
vector<HKL_INSTANCE> instances;
unsigned int number_of_xtals;
clipper::HKL_info::HKL_reference_index index;
```

```
public:
HKL_INDEX(clipper::HKL_info::HKL_reference_index &,
   CRYSTALS &);
clipper::HKL_info::HKL_reference_index& idx();
int get_label(unsigned int);
int get_inverse_label(unsigned int);
unsigned int size();
inline HKL_INSTANCE& operator[](unsigned int);
};
class ALL_INDICES
{
private:
vector<HKL_INDEX> indices;
CRYSTALS *data;
clipper::HKL_info lattice;
bool I_sigI_exists;
unsigned int number_of_bins;
vector<unsigned int> bin;
vector< Array2D<unsigned int> > common;
vector< Array2D<double> > covariance;
vector< Array2D<double> > correlation;
public:
ALL_INDICES(CRYSTALS &);
void init(CRYSTALS &);
bool is_IsigI();
bool missing_FsigF(unsigned int , unsigned int);
bool missing_IsigI(unsigned int , unsigned int);
unsigned int number_of_crystals();
unsigned int return_number_of_bins();
unsigned int bin_number(unsigned int) const;
vector< Array2D<unsigned int> > common_reflections();
vector< Array2D<double> > sample_covariance();
vector< Array2D<double> > sample_correlation(bool);
```

```
void update_sample_covariance();
double get_I(unsigned int,unsigned int);
double get_sigI(unsigned int,unsigned int);
double get_normI(unsigned int,unsigned int);
double get_normsigI(unsigned int,unsigned int);
unsigned int get_epsilon(unsigned int);
double get_sample_covariance(unsigned int, unsigned int,
   unsigned int);
double get_sample_correlation(unsigned int, unsigned int,
    unsigned int);
double get_bin_sample_covariance(unsigned int, unsigned
   int, unsigned int);
double get_bin_sample_correlation(unsigned int, unsigned
   int, unsigned int);
unsigned int get_bin_common(unsigned int, unsigned int,
   unsigned int);
vector< Array2D<double> > return_sample_correlation();
vector<unsigned int> return_bin() const;
unsigned int size();
inline HKL_INDEX& operator[](unsigned int);
};
class PARAMETERS
{
private:
vector< Array2D<double> > covariance_matrix;
unsigned int number_of_bins;
vector<unsigned int> bin;
public:
PARAMETERS(ALL_INDICES &);
Array2D<double> get_covariance_matrix(unsigned int);
Array2D<double> get_bin_covariance_matrix(unsigned int);
double get_covariance(unsigned int, unsigned int,
   unsigned int);
```

```
double get_bin_covariance(unsigned int, unsigned int,
   unsigned int);
void set_covariance(unsigned int, Array2D<double>);
void renormalise(ALL_INDICES &);
unsigned int covariance_dim1(unsigned int) const;
unsigned int return_number_of_bins();
unsigned int bin_number(unsigned int) const;
Array2D<double> pack_unpack(const Array2D<double> &,
   Array2D<double> &,unsigned int);
Array2D<double> sigma(ALL_INDICES &, unsigned int);
Array2D<double> sigma_for_true(ALL_INDICES &, unsigned
   int);
};
inline CRYSTAL& CRYSTALS::operator[](unsigned int x)
{
return crystals[x];
}
inline HKL_INSTANCE& HKL_INDEX::operator[](unsigned int x
   )
{
return instances[x];
}
inline HKL_INDEX& ALL_INDICES::operator[](unsigned int x)
{
return indices[x];
}
}
```

B.2 Minimisation functions

The *C++* header file `intensity_cluster.h` defines the functions used in minimisation. Matrix and vector operations called by these functions are defined in the header file `vector_operations.h` (not included here).

```
#ifndef CLUSTER_H
#define CLUSTER_H
#include "intensity_multicryst.h"
void ML_estimation(MULTICRYST::ALL_INDICES &, MULTICRYST
   ::PARAMETERS &);
vector< Array2D<double> > all_sigma_inverses(MULTICRYST::
   ALL_INDICES &, MULTICRYST::PARAMETERS &);
double logL(unsigned int, unsigned int, MULTICRYST::
   ALL_INDICES &, MULTICRYST::PARAMETERS &, const vector<
    Array2D<double> >&);
vector<double> dlogL(unsigned int, unsigned int,
   MULTICRYST::ALL_INDICES &, MULTICRYST::PARAMETERS &,
   const vector< Array2D<double> >&);
Array2D<double> EddlogL(unsigned int, unsigned int,
   MULTICRYST::ALL_INDICES &, MULTICRYST::PARAMETERS &,
   const vector< Array2D<double> >&);
Array2D<double> parameters_subtract(unsigned int,
   MULTICRYST::PARAMETERS &, vector<double> &);
void update_sigma_inverses(unsigned int, MULTICRYST::
   ALL_INDICES &, MULTICRYST::PARAMETERS &, vector<
   Array2D<double> > &);
bool subroutine(unsigned int, unsigned int, MULTICRYST::
   ALL_INDICES &, MULTICRYST::PARAMETERS &, vector<
   vector<double> >, vector< Array2D<double> > &);
bool parameters_positive_def(unsigned int, MULTICRYST::
   PARAMETERS &, Array2D<double>);
vector<double> devide_shift(double, vector<double>);
```

```
double g(unsigned int, unsigned int, MULTICRYST::
   PARAMETERS &);
double dg(unsigned int, unsigned int, unsigned int,
   unsigned int, MULTICRYST::PARAMETERS &);
double ddg(unsigned int, unsigned int, unsigned int,
   unsigned int, unsigned int, unsigned int, MULTICRYST::
   PARAMETERS &);
#endif
```

B.3 True value functions

The *C++* header file `intensity_true.h` defines the functions used in estimating the true values of the I's and σ_I's.

```
#ifndef TRUE_H
#define TRUE_H
#include "intensity_multicryst.h"
vector< vector< vector<double> > > True_I_sigI(MULTICRYST
   ::ALL_INDICES &, MULTICRYST::PARAMETERS &);
vector< vector<double> > get_expected_I(MULTICRYST::
   ALL_INDICES &, MULTICRYST::PARAMETERS &);
vector< vector<double> > get_Iobs(MULTICRYST::ALL_INDICES
    &);
vector< Array2D<double> > get_C(MULTICRYST::ALL_INDICES
   &, MULTICRYST::PARAMETERS &);
vector< Array2D<double> > all_sigma_inverses_for_true(
   MULTICRYST::ALL_INDICES &, MULTICRYST::PARAMETERS &);
#endif
```

List of Tables

List of Figures

Bibliography

A. Adamczyk and A. Warshel. Converting structural information into an allosteric-energy-based picture for elongation factor tu activation by the ribosome. *Proc Natl Acad Sci USA*, 108(24):9827–9832, 2011.

P.D. Adams, P.V. Afonine, G. Bunkóczi, V.B. Chen, I.W. Davis, N. Echols, J.J. Headd, L.W. Hung, G.J. Kapral, R.W. Grosse-Kunstleve, A.J. McCoy, N.W. Moriarty, R. Oeffner, R.J. Read, D.C. Richardson, J.S. Richardson, T.C. Terwilliger, and P.H. Zwart. Phenix: a comprehensive python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr*, 66:213–221, 2010.

X. Agirrezabala, J. Lei, J.L. Brunelle, R.F. Ortiz-Meoz, R. Green, and J. Frank. Visualization of the hybrid state of trna binding promoted by spontaneous ratcheting of the ribosome. *Mol Cell*, 32(2):190–197, 2008.

X. Agirrezabala, H.Y. Liao, E. Schreiner, J. Fu, R.F. Ortiz-Meoz, K. Schulten, R. Green, and J. Frank. Head swivel on the ribosome facilitates translocation by means of intra-subunit trna hybrid sites. *Proc Natl Acad Sci USA*, 109(16):6094–6099, 2012.

A. Amunts, A. Brown, X.C. Bai, J.L. Llácer, T. Hussain, P. Emsley, F. Long, G. Murshudov, S.H. Scheres, and V. Ramakrishnan. Structure of the yeast mitochondrial large ribosomal subunit. *Science*, 343(6178):1485–1489, 2014.

A. Antoun, M.Y. Pavlov, M. Lovmar, and M. Ehrenberg. How initiation factors maximize the accuracy of trna selection in initiation of bacterial protein synthesis. *Mol Cell*, 23(2):183–193, 2006.

X.C. Bai, I.S. Fernández, G. Mcmullan, and S.H.W. Scheres. Ribosome structures to near-atomic resolution from thirty thousand cryo-em particles. *Elife*, 2, 2013.

N. Ban, B. Freeborn, P. Nissen, P. Penczek, R.A. Grassucci, R. Sweet, J. Frank, P.B. Moore, and T.A. Steitz. A 9 Å resolution x-ray crystallographic map of the large ribosomal subunit. *Cell*, 93(7):1105–1115, 1998.

N. Ban, P. Nissen, J. Hansen, P.B. Moore, and T.A. Steitz. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, 289(5481):905–920, 2000.

A. Ben-Shem, L. Jenner, G. Yusupova, and M. Yusupov. Crystal structure of the eukaryotic ribosome. *Science*, 330(6008):1203–1209, 2010.

A. Ben-Shem, N. Garreau De Loubresse, S. Melnikov, L. Jenner, G. Yusupova, and M. Yusupov. The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science*, 2011.

M. Beringer, S. Adio, W. Wintermeyer, and M. Rodnina. The g2447a mutation does not affect ionization of a ribosomal group taking part in peptide bond formation. *RNA*, 9(8):919–922, 2003.

B. Bhuyan, A. Dietz, and C. Smith. Pactamycin, a new antitumor antibiotic. i: Discovery and biological properties. *Antimicrob Agents Chemother*, 1961:184–190, 1962.

P. Bieling, M. Beringer, S. Adio, and M.V. Rodnina. Peptide bond formation does not involve acid-base catalysis by ribosomal residues. *Nat Struct Mol Biol*, 13(5):423–428, 2006.

S.C. Blanchard, R.L. Gonzalez, H.D. Kim, S. Chu, and J.D. Puglisi. Trna selection and kinetic proofreading in translation. *Nat Struct Mol Biol*, 11(10):1008–1014, 2004.

D. Blow. *Outline of Crystallography for Biologists*. Oxford university press, 2002.

T.L. Blundell and L.N. Johnson. *Protein Crystallography*. Academic Press, New York, 1976.

T.L. Blundell, H. Jhoti, and C. Abell. High-throughput crystallography for lead discovery in drug design. *Nat Rev Drug Discov*, 1(1):45–54, 2002.

E.V. Bocharov, A.G. Sobol, K.V. Pavlov, D.M. Korzhnev, V.A. Jaravine, A.T. Gudkov, and A.S. Arseniev. From structure and dynamics of protein l7/l12 to molecular switching in ribosome. *J Biol Chem*, 279(17):17697–17706, 2004.

J. Bodley, F. Zieve, and L. Lin. Studies on translocation. iv: The hydrolysis of a single round of guanosine triphosphate in the presence of fusidic acid. *J Biol Chem*, 245(21):5662–5667, 1970a.

J.W. Bodley, L. Lin, M.L. Salas, and M. Tao. Studies on translocation v: Fusidic acid stabilization of a eukaryotic ribosome-translocation factor-gdp complex. *FEBS Lett*, 11(3):153–156, 1970b.

M. Borovinskaya, S. Shoji, J. Holton, K. Fredrick, and J. Cate. A steric block in translation caused by the antibiotic spectinomycin. *ACS Chem Biol*, 2(8):545–552, 2007.

W.S. Bowen, N. Van Dyke, E.J. Murgola, J.S. Lodmell, and W.E. Hill. Interaction of thiostrepton and elongation factor-g with the ribosomal protein l11-binding domain. *J Biol Chem*, 280 (4):2934–2943, 2005.

R.N. Bracewell. *Fourier Transform and Its Applications*. McGraw-Hill Science/Engineering/Math, 1980.

W.L. Bragg. The diffraction of short electromagnetic waves by a crystal. *Proc Camb Phil Soc*, 17:43–57, 1912.

M.S. Bretscher. Translocation in protein synthesis: A hybrid structure model. *Nature*, 218 (142):675–677, 1968.

G. Bricogne. A multisolution method of phase determination by combined maximization of entropy and likelihood. iii. extension to powder diffraction data. *Acta Crystallogr A Found Crystallogr*, 47(6):803–829, 1991.

G. Bricogne. Direct phase determination by entropy maximization and likelihood ranking: Status report and perspectives. *Acta Crystallogr D Biol Crystallogr*, 49(1):37–60, 1993.

A.F. Brilot, A.A. Korostelev, D.N. Ermolenko, and N. Grigorieff. Structure of the ribosome with elongation factor g trapped in the pretranslocation state. *Proc Natl Acad Sci USA*, 110(52): 20994–20999, 2013.

D. Brodersen, W. Clemons, A. Carter, B. Wimberly, and V. Ramakrishnan. Phasing the 30s ribosomal subunit structure. *Acta Crystallogr D Biol Crystallogr*, 59(11):2044–2050, 2003.

D.E. Brodersen, W.M. Clemons, Jr., A.P. Carter, R.J. Morgan-Warren, B.T. Wimberly, and V. Ramakrishnan. The structural basis for the action of the antibiotics tetracycline, pactamycin, and hygromycin b on the 30s ribosomal subunit. *Cell*, 103(7):1143–1154., 2000.

A.T. Brunger. Free r value: A novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, 355:472–475, 1992.

E.P. Carpenter, K. Beis, A.D. Cameron, and S. Iwata. Overcoming the challenges of membrane protein crystallography. *Curr Opin Struct Biol*, 18(5):581–586, 2008.

A.P. Carter, W.M. Clemons, Jr., D.E. Brodersen, R.J. Morgan-Warren, B.T. Wimberly, and V. Ramakrishnan. Functional insights from the structure of the 30s ribosomal subunit and its interactions with antibiotics. *Nature*, 407(6802):340–348, 2000.

A.P. Carter, W.M. Clemons, Jr., D.E. Brodersen, R.J. Morgan-Warren, T. Hartsch, B.T. Wimberly, and V. Ramakrishnan. Crystal structure of an initiation factor bound to the 30s ribosomal subunit. *Science*, 291(5503):498–501., 2001.

T.R. Cech. The ribosome is a ribozyme. *Science*, 289(5481):878–879, 2000.

C. Chen, B. Stevens, J. Kaur, D. Cabral, H. Liu, Y. Wang, H. Zhang, G. Rosenblum, Z. Smilansky, Y. Goldman, and B. Cooperman. Single-molecule fluorescence measurements of ribosomal translocation dynamics. *Mol Cell*, 42(3):367–377, 2011.

Y. Chen, S. Feng, V. Kumar, R. Ero, and Y.G. Gao. Structure of ef-g-ribosome complex in a pretranslocation state. *Nat Struct Mol Biol*, 20(9):1077–1084, 2013.

A. Claude. The constitution of protoplasm. *Science*, 97(2525):451–456, 1943.

W.M. Clemons, Jr., J.L. May, B.T. Wimberly, J.P. Mccutcheon, M.S. Capel, and V. Ramakrishnan. Structure of a bacterial 30s ribosomal subunit at 5.5 Å resolution. *Nature*, 400(6747):833–40., 1999.

J.A. Codelli and S.E. Reisman. Pactamycin made easy. *Science*, 340(6129):152–153, 2013.

S.R. Connell, C. Takemoto, D.N. Wilson, H. Wang, K. Murayama, T. Terada, M. Shirouzu, M. Rost, M. Schuler, J. Giesebrecht, M. Dabrowski, T. Mielke, P. Fucini, S. Yokoyama, and C.M. Spahn. Structural basis for interaction of the ribosome with the switch regions of gtp-bound elongation factors. *Mol Cell*, 25(5):751–764, 2007.

P. Cornish, D. Ermolenko, D. Staple, L. Hoang, R. Hickerson, H. Noller, and T. Ha. Following movement of the l1 stalk between three functional states in single ribosomes. *Proc Natl Acad Sci USA*, 106(8):2571–2576, 2009.

K. Cowtan. 'clipper' code and documentation. January 2010. URL `http://www.ysbl.york.ac.uk/~cowtan/clipper/clipper.html`.

F.H.C. Crick. On protein synthesis. *Symp Soc Exp Biol*, 12:138–161, 1958.

F.H.C. Crick. Codon-anticodon pairing: The wobble hypothesis. *J Mol Biol*, 19(2):548–555, 1966.

F.H.C. Crick, L. Barnett, S. Brenner, and R.J. Watts-Tobin. General nature of the genetic code for proteins. *Nature*, 192:1227–1232, 1961.

C.E. Cunha, R. Belardinelli, F. Peske, W. Holtkamp, W. Wintermeyer, and M.V. Rodnina. Dual use of gtp hydrolysis by elongation factor g on the ribosome. *Translation*, 1(1):e24315, 2013.

J. Czworkowski and P.B. Moore. The conformational properties of elongation factor g and the mechanism of translocation. *Biochemistry*, 36(33):10327–10334, 1997.

J. Czworkowski, J. Wang, T.A. Steitz, and P.B. Moore. The crystal structure of elongation factor g complexed with gdp at 2.7 Å resolution. *EMBO J*, 13(16):3661–3668, 1994.

T. Daviter, H.J. Wieden, and M.V. Rodnina. Essential role of histidine 84 in elongation factor tu for the chemical step of gtp hydrolysis on the ribosome. *J Mol Biol*, 332(3):689–699, 2003.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. B*, 39(1):1–38, 1977.

K. Diederichs and P.A. Karplus. Improved r-factors for diffraction data analysis in macromolecular crystallography. *Nat Struct Biol*, 4(4):269–275, 1997.

V. Dinçbas-Renqvist, A. Engstrom, L. Mora, V. Heurgue-Hamard, R. Buckingham, and M. Ehrenberg. A post-translational modification in the ggq motif of rf2 from *escherichia coli* stimulates termination of translation. *EMBO J*, 19(24):6900–6907, 2000.

G. Dinos, D.N. Wilson, Y. Teraoka, W. Szaflarski, P. Fucini, D. Kalpaxis, and K.H. Nierhaus. Dissecting the ribosomal inhibition mechanisms of edeine and pactamycin: The universally conserved residues g693 and c795 regulate p-site rna binding. *Molecular cell*, 13(1):113–124, 2004.

P.A.M. Dirac. *The Principles of Quantum Mechanics*, volume 27. Oxford university press, 1981.

E. Dodson, M. Moore, A. Ralph, and S. Bailey, editors. *Macromolecular Refinement*, volume Proceedings of the CCP4 Study Weekend, 1996.

S. Dorner, J.L. Brunelle, D. Sharma, and R. Green. The hybrid state of trna binding is an authentic translation elongation intermediate. *Nat Struct Mol Biol*, 13(3):234–241, 2006.

J. Drenth. *Principles of Protein X-Ray Crystallography*. Springer, 2007.

J. Dunkle, L. Wang, M. Feldman, A. Pulk, V. Chen, G. Kapral, J. Noeske, J. Richardson, S. Blanchard, and J. Cate. Structures of the bacterial ribosome in classical and hybrid states of trna binding. *Science*, 332(6032):981–984, 2011.

J. Egebjerg and R.A. Garrett. Binding sites of the antibiotics pactamycin and celesticetin on ribosomal rnas. *Biochimie*, 73(7):1145–1149, 1991.

P. Emsley, B. Lohkamp, W. Scott, and K. Cowtan. Features and development of coot. *Acta Crystallogr D Biol Crystallogr*, 66(4):486–501, 2010.

D. Ermolenko and H. Noller. mrna translocation occurs during the second step of ribosomal intersubunit rotation. *Nat Struct Mol Biol*, 18(4):457–462, 2011.

D.N. Ermolenko, Z.K. Majumdar, R.P. Hickerson, P.C. Spiegel, R.M. Clegg, and H.F. Noller. Observation of intersubunit movement of the ribosome in solution using fret. *J Mol Biol*, 370(3):530–540, 2007.

P. Evans. Scaling and assessment of data quality. *Acta Crystallogr D Biol Crystallogr*, 62(1): 72–82, 2005.

P.R. Evans and G.N. Murshudov. How good are my data and what is the resolution? *Acta Crystallogr D Biol Crystallogr*, 69(7):1204–1214, 2013.

E. Fedorov. The symmetry of regular systems of figures. *Zap Mineralog Obsc*, 28:1–146, 1891.

J. Fei, P. Kosuri, D.D. Macdougall, R.L. Gonzalez, and Jr. Coupling of ribosomal l1 stalk and trna dynamics during translation elongation. *Mol Cell*, 30(3):348–359, 2008.

I.S. Fernández, X.C. Bai, T. Hussain, A.C. Kelley, J.F. Lorsch, V. Ramakrishnan, and S.H.W. Scheres. Molecular architecture of a eukaryotic translational initiation complex. *Science*, 342(6160):1240585, 2013a.

I.S. Fernández, C.L. Ng, A.C. Kelley, G. Wu, Y.T. Yu, and V. Ramakrishnan. Unusual base pairing during the decoding of a stop codon by the ribosome. *Nature*, 500(7460):107–110, 2013b.

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

P. Forman. The discovery of the diffraction of x-rays by crystals; a critique of the myths. *Archive for History of Exact Sciences*, 6(1):38–71, 1969.

J. Frank and R.K. Agrawal. A ratchet-like inter-subunit reorganization of the ribosome during translocation. *Nature*, 406:319–322, 2000.

B.D. Freudenthal, W.A. Beard, D.D. Shock, and S.H. Wilson. Observing a dna polymerase choose right from wrong. *Cell*, 154(1):157–168, 2013.

W. Friedrich, P. Knipping, and M. Laue. Interferenzerscheinungen bei röntgenstrahlen. *Ann Phys*, 346(10):971–988, 1913.

L.Y. Frolova, R.Y. Tsivkovskii, G.F. Sivolobova, N.Y. Oparina, O.I. Serpinsky, V.M. Blinov, S.I. Tatkov, and L.L. Kisselev. Mutations in the highly conserved ggq motif of class 1 polypeptide release factors abolish ability of human erf1 to trigger peptidyl-trna hydrolysis. *RNA*, 5(8): 1014–1020, 1999.

E.E. Fry, J. Grimes, and D.I. Stuart. Virus crystallography. *Mol Biotechnol*, 12(1):13–23, 1999.

E. Gale, E. Cundliffe, P. Reynolds, M. Richmond, and M. Waring. The molecular basis of antibiotic action, 1981. *Wiley, London*, 327:389, 1987.

H. Gao, Z. Zhou, U. Rawat, C. Huang, L. Bouakaz, C. Wang, Z. Cheng, Y. Liu, A. Zavialov, R. Gursky, S. Sanyal, M. Ehrenberg, J. Frank, and H. Song. Rf3 induces ribosomal conformational changes responsible for dissociation of class i release factors. *Cell*, 129(5):929–941, 2007.

Y. Gao, M. Selmer, C. Dunham, A. Weixlbaumer, A.C. Kelley, and V. Ramakrishnan. The structure of the ribosome with elongation factor g trapped in the posttranslocational state. *Science*, 326(5953):694–699, 2009.

E.F. Garman. Radiation damage in macromolecular crystallography: What is it and why should we care? *Acta Crystallogr D Biol Crystallogr*, 66(4):339–351, 2010.

D. Gautheret, S.H. Damberger, and R.R. Gutell. Identification of base-triples in rna using comparative sequence analysis. *J Mol Biol*, 248:27–43, 1995.

R. Giordano, R.M.F. Leal, G.P. Bourenkov, S. McSweeney, and A.N Popov. The application of hierarchical cluster analysis to the selection of isomorphous crystals. *Acta Crystallogr D Biol Crystallogr*, 68:649–658, 2012.

C. Grigoriadou, S. Marzi, S. Kirillov, C.O. Gualerzi, and B.S. Cooperman. A quantitative kinetic scheme for 70s translation initiation complex formation. *J Mol Biol*, 373(3):562–572, 2007a.

C. Grigoriadou, S. Marzi, D. Pan, C.O. Gualerzi, and B.S. Cooperman. The translational fidelity function of if3 during transition from the 30s initiation complex to the 70s initiation complex. *J Mol Biol*, 373(3):551–561, 2007b.

Z. Guo and H.F. Noller. Rotation of the head of the 30s ribosomal subunit during mrna translocation. *Proc Natl Acad Sci USA*, 109(50):20391–20394, 2012.

R.R. Gutell, A. Power, G.Z. Hertz, E.J. Putz, and G.D. Stormo. Identifying constraints on the higher-order structure of rna: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res*, 20(21):5785–5795, 1992.

T. Hahn, U. Shmueli, A.J.C. Wilson, and E. Prince. *International Tables for Crystallography*. D. Reidel Publishing Company, 2005.

J. Hajdu, R. Neutze, T. Sjögren, K. Edman, A. Szöke, R.C. Wilmouth, and C.M. Wilmot. Analyzing protein functions in four dimensions. *Nat Struct Biol*, 7(11):1006–1012, 2000.

W.C. Hamilton, J.S. Rollett, and R.A. Sparks. On the relative scaling of x-ray photographs. *Acta Crystallogr*, 18(1):129–130, 1965.

S. Hanessian, R.R. Vakiti, S. Dorich, S. Banerjee, F. Lecomte, J.R. Delvalle, J. Zhang, and B. Deschênes-Simard. Total synthesis of pactamycin. *Angew Chem Int Ed*, 50(15):3497–3500, 2011.

S. Hanessian, R.R. Vakiti, A.K. Chattopadhyay, S. Dorich, and C. Lavallée. Probing functional diversity in pactamycin toward antibiotic, antitumor, and antiprotozoal activity. *Bioorg Med Chem*, 21(7):1775–1786, 2013.

J.L. Hansen, J.A. Ippolito, N. Ban, P. Nissen, P.B. Moore, and T.A. Steitz. The structures of four macrolide antibiotics bound to the large ribosomal subunit. *Mol Cell*, 10(1):117–128, 2002a.

J.L. Hansen, T.M. Schmeing, P.B. Moore, and T.A. Steitz. Structural insights into peptide bond formation. *Proc Natl Acad Sci USA*, 99(18):11670–11675, 2002b.

M.A. Hanson, C.B. Roth, E. Jo, M.T. Griffith, F.L. Scott, G. Reinhart, H. Desale, B. Clemons, S.M. Cahalan, and S.C. Schuerer. Crystal structure of a lipid g protein-coupled receptor. *Science*, 335(6070):851–855, 2012.

S. Hansson, R. Singh, A.T. Gudkov, A. Liljas, and D.T. Logan. Crystal structure of a mutant elongation factor g trapped with a gtp analogue. *FEBS Lett*, 579(20):4492–4497, 2005.

H.A. Hauptman. The phase problem of x-ray crystallography. *Proceedings of the Indian Academy of Sciences-Chemical Sciences*, 92(4-5):291–321, 1983.

H.A. Hauptman. The phase problem of x-ray crystallography. *Rep Prog Phys*, 54(11):1427–1454, 1991.

H.A. Hauptman. The phase problem of x-ray crystallography. In *Twentieth Century Harmonic Analysis- A Celebration*, pages 163–171. Springer, 2001.

H.A. Hauptman. The phase problem of x-ray crystallography. *Physics Today*, 42(11):24–29, 2008.

J.W.B. Hershey and W.C. Merrick. Pathway and mechanism of initiation of protein synthesis. *Cold Spring Harbor Monograph Series*, 39:33–88, 2000.

A. Hirashima and A. Kaji. Role of elongation factor g and a protein factor on the release of ribosomes from messenger ribonucleic acid. *J Biol Chem*, 248(21):7580–7587, 1973.

M.B. Hoagland, M.L. Stephenson, H.F. Scott, L.I. Hecht, and P.C. Zamecnik. A soluble ribonucleic acid intermediate in protein synthesis. *J Biol Chem*, 231:241–257, 1958.

G.H. Hogeboom, W.C. Schneider, and G.E. Pallade. Cytochemical studies of mammalian tissues i. isolation of intact mitochondria from rat liver; some biochemical properties of mitochondria and submicroscopic particulate material. *J Biol Chem*, 172(2):619–635, 1948.

J.M. Holton. A beginner's guide to radiation damage. *J Synchrotron Radiat*, 16(2):133–142, 2009.

J.M. Holton and K.A. Frankel. The minimum crystal size needed for a complete diffraction data set. *Acta Crystallogr D Biol Crystallogr*, 66(4):393–408, 2010.

H. Hope, F. Frolow, K. Von Böhlen, I. Makowski, C. Kratky, Y. Halfon, H. Danz, P. Webster, K.S. Bartels, H.G. Wittmann, and A. Yonath. Cryocrystallography of ribosomal particles. *Acta Crystallogr*, B 45:190–199, 1989.

P. Howell and G. Smith. Identification of heavy-atom derivatives by normal probability methods. *J App Crystallogr*, 25(1):81–86, 1992.

N. Inoue-Yokosawa, C. Ishikawa, and Y. Kaziro. The role of guanosine triphosphate in translocation reaction catalyzed by elongation factor g. *J Biol Chem*, 249(13):4321–4323, 1974.

K. Ito, M. Uno, and Y. Nakamura. A tripeptide 'anticodon' deciphers stop codons in messenger rna. *Nature*, 403(6770):680–684, 2000.

T. Ito, N. Roongsawang, N. Shirasaka, W. Lu, P.M. Flatt, N. Kasanah, C. Miranda, and T. Mahmud. Deciphering pactamycin biosynthesis and engineered production of new pactamycin analogues. *ChemBioChem*, 10(13):2253–2265, 2009.

R. Jackson, C. Hellen, and T. Pestova. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev Mol Cell Biol*, 11(2):113–127, 2010.

L. Jenner, S. Melnikov, N.G. De Loubresse, A. Ben-Shem, M. Iskakova, A. Urzhumtsev, A. Meskauskas, J. Dinman, G. Yusupova, and M. Yusupov. Crystal structure of the 80s yeast ribosome. *Curr Opin Struct Biol*, 22(6):759–767, 2012.

H. Jin, A. Kelley, and V. Ramakrishnan. Crystal structure of the hybrid state of ribosome in complex with the guanosine triphosphatase release factor 3. *Proc Natl Acad Sci USA*, 108 (38):15798–15803, 2011.

P. Julian, A. Konevega, S. Scheres, M. Lazaro, D. Gil, W. Wintermeyer, M. Rodnina, and M. Valle. Structure of ratcheted ribosomes with trnas in hybrid states. *Proc Natl Acad Sci USA*, 105(44):16924–16927, 2008.

W. Kabsch. Evaluation of single-crystal x-ray diffraction data from a position-sensitive detector. *J Appl Crystallogr*, 21(6):916–924, 1988.

W. Kabsch. Xds. *Acta Crystallogr D Biol Crystallogr*, 66(2):125–132, 2010a.

W. Kabsch. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr D Biol Crystallogr*, 66(2):133–144, 2010b.

R. Karimi and M. Ehrenberg. Dissociation rate of cognate peptidyl-trna from the a-site of hyper-accurate and error-prone ribosomes. *Eur J Biochem*, 226(2):355–360, 1994.

R. Karimi, M.Y. Pavlov, R.H. Buckingham, and M. Ehrenberg. Novel roles for classical factors at the interface between translation termination and initiation. *Mol Cell*, 3(5):601–609, 1999.

P.A. Karplus and K. Diederichs. Linking crystallographic model and data quality. *Science*, 336 (6084):1030–1033, 2012.

E.B. Keller, P.C. Zamecnik, and R.B. Loftfield. The role of microsomes in the incorporation of amino acids into proteins. *J HistochemCytochem*, 2(5):378–386, 1954.

G.J. Kleywegt and T.A. Jones. Good model-building and refinement practice. *Methods Enzymol*, 277:208–230, 1997.

S. Klinge, F. Voigts-Hoffmann, M. Leibundgut, S. Arpagaus, and N. Ban. Crystal structure of the eukaryotic 60s ribosomal subunit in complex with initiation factor 6. *Science*, 334 (6058):941–948, 2011.

A. Korostelev, S. Trakhanov, M. Laurberg, and H.F. Noller. Crystal structure of a 70s ribosome-trna complex reveals functional interactions and rearrangements. *Cell*, 126(6):1065–1077, 2006.

W. Kühlbrandt. Biochemistry. the resolution revolution. *Science*, 343(6178):1443–1444, 2014.

C.G. Kurland. Molecular characterization of ribonucleic acid from *escherichia coli* ribosomes. *J Mol Biol*, 2:83–91, 1960.

E.E. Lattman and P.J. Loll. *Protein Crystallography: A Concise Guide*. JHU Press, 2008.

M. Laurberg, H. Asahara, A. Korostelev, J. Zhu, S. Trakhanov, and H.F. Noller. Structural basis for translation termination on the 70s ribosome. *Nature*, 454(7206):852–857, 2008.

B.S. Laursen, H.P. Sø rensen, K.K. Mortensen, and H.U. Sperling-Petersen. Initiation of protein synthesis in bacteria. *Microbiol Mol Biol Rev*, 69(1):101–123, 2005.

A. Liljas, M. Ehrenberg, and J. Aqvist. Comment on "the mechanism for activation of gtp hydrolysis on the ribosome". *Science*, 333(6038):37; author reply 37, 2011.

M. Lindahl, L.A. Svensson, A. Liljas, S.E. Sedelnikova, I.A. Eliseikina, N.P. Fomenkova, N. Nevskaya, S.V. Nikonov, M.B. Garber, T.A. Muranova, A.I. Rykonova, and R. Amons. Crystal structure of the ribosomal protein s6 from *thermus thermophilus*. *EMBO J*, 13(6): 1249–1254, 1994.

F. Lipmann. Polypeptide chain elongation in protein biosynthesis. *Science*, 164(883):1024–1031, 1969.

B.A. Lippmann and J. Schwinger. Variational principles for scattering processes. i. *Phys Rev*, 79(3):469, 1950.

J.W. Littlefield, E.B. Keller, J. Gross, and P.C. Zamecnik. Studies on cytoplasmic ribonucleoprotein particles from the liver of the rat. *J Biol Chem*, 217(1):111–124, 1955.

Q. Liu, T. Dahmane, Z. Zhang, Z. Assur, J. Brasch, L. Shapiro, F. Mancia, and W.A. Hendrickson. Structures from anomalous diffraction of native biological macromolecules. *Science*, 336 (6084):1033–1037, 2012.

W. Lu, N. Roongsawang, and T. Mahmud. Biosynthetic studies and genetic engineering of pactamycin analogs with improved selectivity toward malarial parasites. *Chem Biol*, 18(4): 425–431, 2011.

V. Luzzati. Traitement statistique des erreurs dans la determination des structures cristallines. *Acta Crystallogr*, 5(6):802–810, 1952.

J.T. Malinowski, R.J. Sharpe, and J.S. Johnson. Enantioselective synthesis of pactamycin, a complex antitumor antibiotic. *Science*, 340(6129):180–182, 2013.

A.S. Mankin. Pactamycin resistance mutations in functional sites of 16s rrna. *J Mol Biol*, 274 (1):8–15, 1997.

K. Martemyanov and A. Gudkov. Domain iii of elongation factor g from *thermus thermophilus* is essential for induction of gtp hydrolysis on the ribosome. *J Biol Chem*, 275(46):35820–35824, 2000.

H. Matthews, M. Usman-Idris, F. Khan, M. Read, and N. Nirmalan. Drug repositioning as a route to anti-malarial drug discovery: Preliminary investigation of the in vitro anti-malarial efficacy of emetine dihydrochloride hydrate. *Malar J*, 12(1):359, 2013.

A.A. McCarthy, S. Brockhauser, D. Nurizzo, P. Theveneau, T Mairs, D. Spruce, M. Guijarro, M. Lesourd, RB. Ravelli, and McSweeney S. A decade of user operation on the macromolecular crystallography mad beamline id14-4 at the esrf. *J Synchrotron Radiat*, 16(6): 803–812, 2009.

A.J. McCoy, R.W. Grosse-Kunstleve, P.D. Adams, M.D. Winn, L.C. Storoni, and R.J. Read. Phaser crystallographic software. *J Appl Crystallogr*, 40(4):658–674, 2007.

A. McPherson. *Introduction to Macromolecular Crystallography*. John Wiley and Sons, 2011.

P. Milon, A.L. Konevega, C.O. Gualerzi, and M.V. Rodnina. Kinetic checkpoint at a late step in translation initiation. *Mol Cell*, 30(6):712–720, 2008.

P. Milon, M. Carotti, A. Konevega, W. Wintermeyer, M.V. Rodnina, and C. Gualerzi. The ribosome-bound initiation factor 2 recruits initiator trna to the 30s initiation complex. *EMBO Rep*, 11(4):312–316, 2010.

D. Moazed and H.F. Noller. Interaction of antibiotics with functional sites in 16s ribosomal rna. *Nature*, 327(6121):389–94., 1987.

D. Moazed and H.F. Noller. Intermediate states in the movement of transfer rna in the ribosome. *Nature*, 342(6246):142–148, 1989.

D. Moazed, J.M. Robertson, and H.F. Noller. Interaction of elongation factors ef-g and ef-tu with a conserved loop in 23s rna. *Nature*, 334(6180):362–364, 1988.

E.H. Moore. On the reciprocal of the general algebraic matrix. *Bull. Amer. Math. Soc.*, 26(9): 394–395, 1920.

L. Mora, V. Heurgue-Hamard, S. Champ, M. Ehrenberg, L.L. Kisselev, and R.H. Buckingham. The essential role of the invariant ggq motif in the function and stability in vivo of bacterial release factors rf1 and rf2. *Mol Microbiol*, 47(1):267–275, 2003.

J. Munro, R. Altman, C. Tung, J. Cate, K. Sanbonmatsu, and S. Blanchard. Spontaneous formation of the unlocked state of the ribosome is a multistep process. *Proc Natl Acad Sci USA*, 107(2):709–714, 2010.

G. Murshudov, P. Skubak, A. Lebedev, N. Pannu, R. Steiner, R. Nicholls, M. Winn, F. Long, and A. Vagin. Refmac5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr*, 67(4):355–367, 2011.

G.N. Murshudov, A.A. Vagin, and E.J. Dodson. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr*, 53(3):240–255, 1997.

G.W. Muth, L. Ortoleva-Donnelly, and S.A. Strobel. A single adenosine with a neutral pk(a) in the ribosomal peptidyl transferase center. *Science*, 289(5481):947–950, 2000.

C. Neubauer, R. Gillet, A. Kelley, and V. Ramakrishnan. Decoding in the absence of a codon by tmrna and smpb in the ribosome. *Science*, 335(6074):1366–1369, 2012.

M. Nirenberg, P. Leder, M. Bernfield, R. Brimacombe, J. Trupin, F. Rottman, and C. O'Neal. Rna codewords and protein synthesis, vii. on the general nature of the rna code. *Proc Natl Acad Sci USA*, 53(5):1161–1168, 1965.

P. Nissen, M. Kjeldgaard, S. Thirup, G. Polekhina, L. Reshetnikova, B.F. Clark, and J. Nyborg. Crystal structure of the ternary complex of phe-trnaphe, ef-tu, and a gtp analog. *Science*, 270(5241):1464–1472, 1995.

P. Nissen, J. Hansen, N. Ban, P.B. Moore, and T.A. Steitz. The structural basis of ribosome activity in peptide bond synthesis. *Science*, 289(5481):920–930, 2000.

J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer-Verlag, 1999.

J.M. Ogle, D.E. Brodersen, W.M. Clemons, Jr., M.J. Tarry, A.P. Carter, and V. Ramakrishnan. Recognition of cognate transfer rna by the 30s ribosomal subunit. *Science*, 292(5518): 897–902., 2001.

J.M. Ogle, F.V. Murphy, M.J. Tarry, and V. Ramakrishnan. Selection of trna by the ribosome requires a transition from an open to a closed form. *Cell*, 111(5):721–732, 2002.

K. Otoguro, M. Iwatsuki, A. Ishiyama, M. Namatame, A. Nishihara-Tukashima, S. Shibahara, S. Kondo, H. Yamada, and S. Omura. Promising lead compounds for novel antiprotozoals. *J Antibiot*, 63:381–384, 2010.

G.E. Palade. A small particulate component of the cytoplasm. *J Biophys Biochem Cytol*, 1(1): 59–67, 1955.

G.E. Palade and P. Siekevitz. Liver microsomes; an integrated morphological and biochemical study. *J Biophys Biochem Cytol*, 2(2):171–200, 1956.

N.S. Pannu and R.J. Read. Improved structure refinement through maximum likelihood. *Acta Crystallogr A Found Crystallogr*, 52(5):659–668, 1996.

T. Pape, W. Wintermeyer, and M.V. Rodnina. Conformational switch in the decoding region of 16s rrna during aminoacyl-trna selection on the ribosome. *Nat Struct Biol*, 7(2):104–107, 2000.

R. Penrose. A generalized inverse for matrices. *Proc. Cambridge Philos. Soc.*, 51:406–413, 1955.

F. Peske, M.V. Rodnina, and W. Wintermeyer. Sequence of steps in ribosome recycling as defined by kinetic analysis. *Mol Cell*, 18(4):403–412, 2005.

M.L. Petermann and M.G. Hamilton. An ultracentrifugal analysis of macromolecular particles of normal and leukemic mouse spleen. *Cancer Res*, 12:373–378, 1952.

M.L. Petermann, N.A. Mizen, and M.G. Hamilton. The macromolecular particles of normal and regenerating rat liver. *Cancer Res*, 13(4-5):372–375, 1953.

M. Pioletti, F. Schlunzen, J. Harms, R. Zarivach, M. Gluhmann, H. Avila, A. Bashan, H. Bartels, T. Auerbach, C. Jacobi, T. Hartsch, A. Yonath, and F. Franceschi. Crystal structures of complexes of the small ribosomal subunit with tetracycline, edeine and if3. *EMBO J*, 20 (8):1829–139., 2001.

N.V. Plotnikov, J. Lameira, and A. Warshel. Quantitative exploration of the molecular origin of the activation of gtpase. *Proc Natl Acad Sci USA*, 110(51):20509–20514, 2013.

J. Poehlsgaard and S. Douthwaite. The bacterial ribosome as a target for antibiotics. *Nat Rev Microbiol*, 3(11):870–881, 2005.

N. Polacek, M. Gaynor, A. Yassin, and A.S. Mankin. Ribosomal peptidyl transferase can withstand mutations at the putative catalytic nucleotide. *Nature*, 411(6836):498–501, 2001.

A. Pulk and J.H.D. Cate. Control of ribosomal subunit rotation by elongation factor g. *Science*, 340(6140), 2013.

J. Rabl, M. Leibundgut, S. Ataide, A. Haag, and N. Ban. Crystal structure of the eukaryotic 40s ribosomal subunit in complex with initiation factor 1. *Science*, 331(6018):730–736, 2011.

D.J.F. Ramrath, L. Lancaster, T. Sprink, T. Mielke, J. Loerke, H.F. Noller, and C.M.T. Spahn. Visualization of two transfer rnas trapped in transit during elongation factor g-mediated translocation. *Proc Natl Acad Sci USA*, 110(52):20964–20969, 2013.

A. et al. Ratje. Head swivel on the ribosome facilitates translocation by means of intra-subunit trna hybrid sites. *Nature*, 468(7324):713–716, 2010.

R. Read. Structure-factor probabilities for related structures. *Acta Crystallogr A Found Crystallogr*, 46(11):900–912, 1990.

G. Rhodes. *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models*. Academic press, 2010.

C. Riekel, M. Burghammer, and G. Schertler. Protein crystallography microdiffraction. *Curr Opin Struct Biol*, 15(5):556–562, 2005.

K.L. Rinehart Jr, D.D. Weller, and C.J. Pearce. Recent biosynthetic studies on antibiotics. *J Nat Prod*, 43(1):1–20, 1980.

R B Roberts, editor. *Microsomal Particles and Protein Synthesis*. Pergamon Press, New York, 1958.

M.V. Rodnina and W. Wintermeyer. Fidelity of aminoacyl-trna selection on the ribosome: Kinetic and structural mechanisms. *Annu Rev Biochem*, 70:415–435, 2001.

M.V. Rodnina, T. Pape, R. Fricke, L. Kuhn, and W. Wintermeyer. Initial binding of the elongation factor tu-gtp-aminoacyl-trna complex preceding codon recognition on the ribosome. *J Biol Chem*, 271(2):646–652, 1996.

M.V. Rodnina, A. Savelsbergh, V.I. Katunin, and W. Wintermeyer. Hydrolysis of gtp by elongation factor g drives trna movement on the ribosome. *Nature*, 385(6611):37–41, 1997.

R. Rosset and R. Monier. A propos de la présence d'acide ribonucléique de faible poids moléculaire dans les ribosomes d'*escherichia coli*. *Biochimica Biophysica Acta*, 68:653–656, 1963.

M.G. Rossmann. *The Molecular Replacement Method: A Collection of Papers on the Use of Non-Crystallographic Symmetry*, volume 13. Routledge, 1972.

M.G. Rossmann and D.M. Blow. The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallogr*, 15(1):24–31, 1962.

B. Rupp. *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*. Garland Publishing: New York, NY, USA, 2010.

A. Savelsbergh, N.B. Matassova, M.V. Rodnina, and W. Wintermeyer. Role of domains 4 and 5 in elongation factor g functions on the ribosome. *J Mol Biol*, 300(4):951–961, 2000.

A. Savelsbergh, V.I. Katunin, D. Mohr, F. Peske, M.V. Rodnina, and W. Wintermeyer. An elongation factor g-induced ribosome rearrangement precedes trna-mrna translocation. *Mol Cell*, 11(6):1517–1523, 2003.

D. Sayre. The squaring method: A new method for phase determination. *Acta Crystallogr*, 5 (1):60–65, 1952.

F. Schluenzen, A. Tocilj, R. Zarivach, J. Harms, M. Gluehmann, D. Janell, A. Bashan, H. Bartels, I. Agmon, F. Franceschi, and A. Yonath. Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. *Cell*, 102(5):615–623, 2000.

F. Schlünzen, R. Zarivach, R. Harms, A. Bashan, A. Tocilj, R. Albrecht, A. Yonath, and F. Franceschi. Structural basis for the interaction of antibiotics with the peptidyl transferase centre in eubacteria. *Nature*, 413(6858):814–821, 2001.

T. Schmeing, R. Voorhees, A. Kelley, and V. Ramakrishnan. How mutations in trna distant from the anticodon affect the fidelity of decoding. *Nat Struct Mol Biol*, 18(4):432–436, 2011.

T.M. Schmeing and V. Ramakrishnan. What recent ribosome structures have revealed about the mechanism of translation. *Nature*, 461(7268):1234–1242, 2009.

T.M. Schmeing, K.S. Huang, D.E. Kitchen, S.A. Strobel, and T.A. Steitz. Structural insights into the roles of water and the 2' hydroxyl of the p site trna in the peptidyl transferase reaction. *Mol Cell*, 20(3):437–448, 2005a.

T.M. Schmeing, K.S. Huang, S.A. Strobel, and T.A. Steitz. An induced-fit mechanism to promote peptide bond formation and exclude hydrolysis of peptidyl-trna. *Nature*, 438(7067): 520–524, 2005b.

T.M. Schmeing, R. Voorhees, A. Kelley, Y. Gao, F.T. Murphy, J. Weir, and V. Ramakrishnan. The crystal structure of the ribosome bound to ef-tu and aminoacyl-trna. *Science*, 326(5953): 688–694, 2009.

F. Schotte, J. Soman, J.S. Olson, M. Wulff, and P.A. Anfinrud. Picosecond time-resolved x-ray crystallography: Probing protein function in real time. *J Struct Biol*, 147(3):235–246, 2004.

E. Schrödinger. An undulatory theory of the mechanics of atoms and molecules. *Phys Rev*, 28 (6):1049–1070, 1926.

Schrödinger, LLC. The PyMOL molecular graphics system, version 1.3r1. August 2010.

B.S. Schuwirth, M.A. Borovinskaya, C.W. Hau, W. Zhang, A. Vila-Sanjurjo, J.M. Holton, and J.H. Cate. Structures of the bacterial ribosome at 3.5 Å resolution. *Science*, 310(5749): 827–834, 2005.

M. Selmer, C.M. Dunham, F.V.T. Murphy, A. Weixlbaumer, S. Petry, A.C. Kelley, J.R. Weir, and V. Ramakrishnan. Structure of the 70s ribosome complexed with mrna and trna. *Science*, 313(5795):1935–1942, 2006.

M. Selmer, Y.-G. Gao, A. Weixlbaumer, and V. Ramakrishnan. Ribosome engineering to promote new crystal forms. *Acta Crystallogr D Biol Crystallogr*, 68(5):578–583, 2012.

R.J. Sharpe, J.T. Malinowski, and J.S. Johnson. Asymmetric synthesis of the aminocyclitol pactamycin, a universal translocation inhibitor. *J Am Chem Soc*, 135(47):17990–17998, 2013.

J.J. Shaw and R. Green. Two distinct components of release factor function uncovered by nucleophile partitioning analysis. *Mol Cell*, 28(3):458–467, 2007.

G. Sheldrick, H. Hauptman, C. Weeks, R. Miller, and I. Usón. *Crystallography of Biological Molecules*, volume International Tables for Crystallography, Vol. F. IUCr, 2001.

G.M. Sheldrick, Z. Dauter, K. Wilson, H. Hope, and L. Sieker. The application of direct methods and patterson interpretation to high-resolution native protein data. *Acta Crystallogr D Biol Crystallogr*, 49(1):18–23, 1993.

F. Sherman, J.W. Stewart, and S. Tsunasawa. Methionine or not methionine at the beginning of a protein. *BioEssays*, 3(1):27–31, 1985.

X. Shi, P. Khade, K. Sanbonmatsu, and S. Joseph. Functional role of the sarcin-ricin loop of the 23s rrna in the elongation cycle of protein synthesis. *J Mol Biol*, 419(3-4):125–138, 2012.

J. Shine and L. Dalgarno. The 3'-terminal sequence of *escherichia coli* 16s ribosomal rna: Complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci USA*, 71(4):1342–1346, 1974.

A. Sievers, M. Beringer, M.V. Rodnina, and R. Wolfenden. The ribosome as an entropy trap. *Proc Natl Acad Sci USA*, 101(21):7897–7901, 2004.

N. Singh, G. Das, A. Seshadri, R. Sangeetha, and U. Varshney. Evidence for a role of initiation factor 3 in recycling of ribosomal complexes stalled on mrnas in *escherichia coli*. *Nucleic Acids Res*, 33(17):5591–5601, 2005.

H. Song, P. Mugnier, A.K. Das, H.M. Webb, D.R. Evans, M.F. Tuite, B.A. Hemmings, and D. Barford. The crystal structure of human eukaryotic release factor erf1- mechanism of stop codon recognition and peptidyl-trna hydrolysis. *Cell*, 100(3):311–321, 2000.

P. Spiegel, D. Ermolenko, and H. Noller. Elongation factor g stabilizes the hybrid-state conformation of the 70s ribosome. *RNA*, 13(9):1473–1482, 2007.

A. Spirin. Temperature effect and macromolecular structure of high-polymer ribonucleic acid of different origin. *Biokhimiya*, 26(3):454–463, 1961.

N. Stein and C. Ballard. Intensity to amplitude conversion using ctruncate. *Acta Crystallogr A Found Crystallogr*, 65:s161, 2009.

G.L. Taylor. Introduction to phasing. *Acta Crystallogr D Biol Crystallogr*, 66(4):325–338, 2010.

J.M. Thomas. Centenary: The birth of x-ray crystallography. *Nature*, 491(7423):186–187, 2012.

A. Tissières and J.D. Watson. Ribonucleoprotein particles from *escherichia coli*. *Nature*, 182: 778–780, 1958.

J. Tomsic, L.A. Vitali, T. Daviter, A. Savelsbergh, R. Spurio, P. Striebeck, W. Wintermeyer, M.V. Rodnina, and C.O. Gualerzi. Late events of translation initiation in bacteria: A kinetic analysis. *EMBO J*, 19(9):2127–2136, 2000.

D.S. Tourigny, I.S. Fernández, A.C. Kelley, and V. Ramakrishnan. Elongation factor g bound to the ribosome in an intermediate state of translocation. *Science*, 340(6140):1235490, 2013a.

D.S. Tourigny, I.S. Fernández, A.C. Kelley, R.R. Vakiti, A.K. Chattopadhyay, S. Dorich, S. Hanessian, and V. Ramakrishnan. Crystal structure of a bioactive pactamycin analogue bound to the 30s ribosomal subunit. *J Mol Biol*, 425(20):3907–3910, 2013b.

S. Trobro and J. Aqvist. Mechanism of peptide bond synthesis on the ribosome. *Proc Natl Acad Sci USA*, 102(35):12395–12400, 2005.

A. Vagin and A. Teplyakov. Molecular replacement with molrep. *Acta Crystallogr D Biol Crystallogr*, 66:22–25, 2010.

M. Valle, A. Zavialov, J. Sengupta, U. Rawat, M. Ehrenberg, and J. Frank. Locking and unlocking of ribosomal motions. *Cell*, 114(1):123–134, 2003.

R. Voorhees, A. Weixlbaumer, D. Loakes, A. Kelley, and V. Ramakrishnan. Insights into substrate stabilization from snapshots of the peptidyl transferase center of the intact 70s ribosome. *Nat Struct Mol Biol*, 16(5):528–533, 2009.

R. Voorhees, T.M. Schmeing, A.C. Kelley, and V. Ramakrishnan. The mechanism for activation of gtp hydrolysis on the ribosome. *Science*, 330(6005):835–838, 2010.

R.M. Voorhees and V. Ramakrishnan. Structural basis of the translational elongation cycle. *Ann Rev Biochem*, 82:203–236, 2013.

R.M. Voorhees, T.M. Schmeing, A.C. Kelley, and V. Ramakrishnan. Response to comment on "the mechanism for activation of gtp hydrolysis on the ribosome". *Science*, 333:37–b, 2011.

J.P. Waller. Fractionation of the ribosomal protein from *escherichia coli*. *J Mol Biol*, 10:319–336, 1964.

J.P. Waller and J.I. Harris. Studies on the composition of the protein from *escherichia coli* ribosomes. *Proc Natl Acad Sci USA*, 47:18–23, 1961.

A.M. Waterhouse, J.B. Procter, D.M. Martin, M. Clamp, and G.J. Barton. Jalview version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191, 2009.

J. Watson. The synthesis of proteins upon ribosomes. *Bull Soc Chim Biol Fr*, 46:1399, 1964.

J. Watson. Genes, girls, and gamow. after. *The Double Helix. Alfred A. Knopf, New York. AVRION MITCHISON*, 55, 2001.

M. Weiss and R. Hilgenfeld. On the use of the merging r factor as a quality indicator for x-ray data. *J App Crystallogr*, 30(2):203–205, 1997.

M.S. Weiss. Global indicators of x-ray data quality. *J App Crystallogr*, 34(2):130–135, 2001.

A. Weixlbaumer, H. Jin, C. Neubauer, R.M. Voorhees, S. Petry, A.C. Kelley, and V. Ramakrishnan. Insights into translational termination from the structure of rf2 bound to the ribosome. *Science*, 322(5903):953–956, 2008.

P.F. Wiley, H.K. Jahnke, F.A. Mackellar, R.B. Kelly, and A.D. Argoudelis. Structure of pactamycin. *J Org Chem*, 35(5):1420–1425, 1970.

A.J.C. Wilson. The probability distribution of x-ray intensities. *Acta Crystallogr*, 2:318–321, 1949.

B.T. Wimberly, D.E. Brodersen, W.M. Clemons, Jr., R.J. Morgan-Warren, A.P. Carter, C. Vonrhein, T. Hartsch, and V. Ramakrishnan. Structure of the 30s ribosomal subunit. *Nature*, 407(6802):327–339, 2000.

H. Wolf, G. Chinali, and A. Parmeggiani. Kirromycin, an inhibitor of protein biosynthesis that acts on elongation factor tu. *Proc Natl Acad Sci USA*, 71(12):4910–4914, 1974.

W. Wong, X.C. Bai, A. Brown, I.S. Fernández, E. Hanssen, M. Condron, Y.H. Tan, J. Baum, and S.H.W. Scheres. Cryo-em structure of the plasmodium falciparum 80s ribosome bound to the anti-protozoan drug emetine. *Elife*, page e03080, 2014.

J. Woodcock, D. Moazed, M. Cannon, J. Davies, and H.F. Noller. Interaction of antibiotics with a- and p-site-specific bases in 16s ribosomal rna. *EMBO J*, 10(10):3099–3103, 1991.

T.Y. Wu and T. Ohmura. *Quantum Theory of Scattering*. Courier Dover Publications, 2011.

S.W. Wukovitz and T.O. Yeates. Why protein crystals favour some space-groups over others. *Nat Struct Mol Biol*, 2(12):1062–1067, 1995.

A. Yonath. Antibiotics targeting ribosomes: Resistance, selectivity, synergism, and cellular regulation. *Annu. Rev. Biochem.*, 74:649–679, 2005.

A. Yonath, J. Mussig, B. Tesche, S. Lorenz, V.A. Erdmann, and H.G. Wittmann. Crystallization of the large ribosomal subunits from *bacillus stearothermophilus*. *Biochem Int*, 1:428–435, 1980.

A. Yonath, J. Piefke, J. Mussig, H.S. Gewitz, and H.G. Wittmann. A compact three-dimensional crystal form of the large ribosomal subunit from *bacillus stearothermophilus*. *FEBS Lett*, 163 (1):69–72, 1983a.

A. Yonath, B. Tesche, S. Lorenz, J. Mussig, V.A. Erdmann, and H.G. Wittmann. Several crystal forms of the bacillus stearothermophilus 50s ribosomal particles. *FEBS Lett*, 154(1):15–20, 1983b.

A. Yonath, H.D. Bartunik, K.S. Bartels, and H.G. Wittmann. Some x-ray diffraction patterns from single crystals of the large ribosomal subunit from *bacillus stearothermophilus*. *J Mol Biol*, 177(1):201–206, 1984.

E.M. Youngman, J.L. Brunelle, A.B. Kochaniak, and R. Green. The active site of the ribosome is composed of two layers of conserved nucleotides with distinct roles in peptide bond formation and peptide release. *Cell*, 117(5):589–599, 2004.

H. Zaher, J. Shaw, S. Strobel, and R. Green. The 2'-oh group of the peptidyl-trna stabilizes an active conformation of the ribosomal ptc. *EMBO J*, 30(12):2445–2453, 2011.

P.C. Zamecnik and E.B. Keller. Relation between phosphate energy donors and incorporation of labeled amino acids into proteins. *J Biol Chem*, 209:337–353, 1954.

A.V. Zavialov, L. Mora, R.H. Buckingham, and M. Ehrenberg. Release of peptide promoted by the ggq motif of class 1 release factors regulates the gtpase activity of rf3. *Mol Cell*, 10(4): 789–798, 2002.

W. Zhang, J. Dunkle, and J. Cate. Structures of the ribosome in intermediate states of ratcheting. *Science*, 325(5943):1014–1017, 2009.

J. Zhou, L. Lancaster, S. Trakhanov, and H. Noller. Crystal structure of release factor rf3 trapped in the gtp state on a rotated conformation of the ribosome. *RNA*, 18(2):230–240, 2012.

J. Zhou, L. Lancaster, J.P. Donohue, and H.F. Noller. Crystal structures of ef-g-ribosome complexes trapped in intermediate states of translocation. *Science*, 340(6140), 2013.