High-resolution structural models of Ribosome Nascent chain Complexes restrained by experimental NMR data

Thesis submitted by
Luke S. Goodsell

For the degree of
Doctor of Philosophy in Biochemistry and Structural Biology

Research Department of Structural and Molecular Biology
University College London
Gower Street, London, WC1E 6BT

27th May 2015

I, Luke S. Goodsell, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

As understanding of the ways in which the complex cellular environment affects the in vivo folding of proteins improves, improved methods for their study are required. It is possible to produce limited quantities of ribosome-nascent chain complexes (RNCs) and techniques for gathering data about them are improving, but no single technique provides all the information required to understand folding of nascent proteins on the ribosome and there are still significant data that cannot be obtained experimentally. In particular, while NMR chemical shift and residual dipolar couplings may be recorded, the samples are of too low concentration and stability to conduct the most informative NOESY experiments that are traditionally used for revealing atomic-resolution structure.

Recently, the ability to use chemical shifts to reveal structural details and dynamic properties of small proteins has been developed. By simulating multiple molecules and predicting the average chemical shift of the ensemble, the simulation may be restrained to conform to the experimentally measured data, making testable predictions about the atomic-resolution dynamic properties of the molecule. By adapting these methods to the macromolecular RNC structures it is theorized that the limited chemical shift data available may be used to provide structural details of the protein as it emerges from a ribosome. This, however, is faced by many challenges, including the ability to simulate such large number of atoms in a suitable timescale and applying the restraints to the nascent chain alone.

The thesis presented describes the development of computational techniques to characterize RNCs, including the concepts and challenges faced, the chemical-shift restrained simulation of nascent chains alone, the development of techniques to perform chemical-shift restrained molecular dynamics simulations of the RNCs and the application of these techniques to a model system.

Acknowledgments

I would like to first express my profound gratitude towards the Wellcome Trust for providing the funding that allowed me to conduct my PhD studies and to Gabriel Waksman as the head of the Department of Structural and Molecular Biology for providing access to such an inspiring and prestigious location to conduct my work, as well as Dr Alethea Tabor for coordinating the PhD programme.

I will forever be grateful for the supervision and mentoring provided by Prof John Christodoulou. He continually provided ideas, opportunities and challenges that inspired me and made me want to achieve greater accomplishments, as well as providing the support and understanding through difficult times I encountered. I am also sincerely grateful to the support, resources, and ideas of my second supervisor, Prof Michele Vendruscolo, who frequently provided the ideas and connections around which avenues of research could be formed.

I am grateful to my thesis committee chair, Prof Paul Driscoll, for providing adept and experienced oversight.

I would like to thank the members of my host groups from whom I learned so much. In particular, I am grateful for the training and support of Dr Lisa Cabrita, Dr Carlo Camilloni, Dr Chris Waudby, Dr John Kirkpatrick and Dr Hélene Launay. I would like to thank other students with whom I worked, Dr Maria-Evangelia Karyadi, Dr Xiaolin Wang, Toshitaka Tajima, Dr Aleksandr Sahakyan, Anne Wentink, Annika Weise and Géraldine Levy.

Finally, I have many family members without whom I would never have been able to achieve this work. I cannot sufficiently express how eternally grateful I am to my wife, Yan Chen, for her support, encouragement and motivation throughout my research. I am sincerely grateful for parents for providing me with the desire and ability to pursue my interests. I am grateful to my two older brothers for providing the challenges and goalposts that encouraged me to always try harder. Finally, I am grateful to my Γιαγιά and Παππού for instilling a family attitude of constant learning and self-improvement to which I aspire.

Contents

Table of Figures

Abbreviations

αSyn	α-synuclein
AFM	Atomic force microscopy
ATP	Adenosine triphosphate
CFTR	Cystic Fibrosis transmembrane conductance regulator
Cryo-EM	Cryo-electron microscopy
CS-MD	CamShift Mmolecular dynamics
ddFLN	*Dictyostelium discoideum* filamin gelation factor
Da	Daltons (unit)
DNA	Deoxyribonucleic acid
E. coli	*Escherichia coli*
EF-G	Elongation factor G
FID	Free induction decay
FRET	Förster resonance energy transfer
GelFac	*Dictyostelium discoideum* filamin gelation factor
GROMACS	Gröningen machine for chemical simulations
HMQC	Heteronuclear multiple quantum coherence
HSQC	Heteronuclear single quantum coherence
IPTG	Isopropyl β-D-1-thiogalactopyranoside
MD	Molecular dynamics
mRNA	Messenger RNA
MW	Molecular Weight
NAC	Non-Aβ component
NMR	Nuclear magnetic resonance
NOESY	Nuclear Overhauser Effect Spectroscopy
PADC	Probe-accessible distance calculator
ppm	Parts per million
PTC	Peptidyl transferase centre
RDC	Residual dipolar coupling
R_g	Gyroscopic radius
R_H	Hydrodynamic radius

RMS	Root mean square
RMSD	RMS deviation
RMSF	RMS fluctuation
RMSΔδ	RMS delta-delta
RNA	Ribonucleic acid
RNC	Ribosome-Nascent chain Complex
rRNA	Ribosomal RNA
SecM	Secretion monitor
TF	Trigger Factor
tRNA	Transfer RNA
UAS1/2	Universal adapter Site 1/2
vdW	van der Waals
wt	Wild-type
XRC	X-ray crystallography

1 Introduction

1.1 The Ribosome

All starting life scientists are taught that much biological activity is mediated by polymers of amino acids called proteins and that the function of proteins is determined by the three-dimensional arrangement of the amino acid residues. Furthermore, as has become cliché to cite, since Christian Anfinsen's investigation into thermodynamic properties of ribonuclease A it has been dogma that the conformation of a protein is determined by the sequence of amino acid residues within it (Anfinsen, 1973). However, it is becoming increasingly evident that there are processes during the biosynthesis of proteins that may determine which of a number of possible conformations a nascent protein may adopt (Cabrita et al., 2010).

The ribosome is the molecular machinery common to all living organisms that translates the genetic sequence – in the form of messenger ribonucleic acid polymers (mRNAs) – to the appropriate amino acid sequence. In the last twelve years, x-ray crystallography (XRC) and electron cryo-microscopy (cryo-EM) have enabled the determination of high resolution structures of the ribosome in various states, including the separate subunits (Ban et al., 2000), the computationally recombined subunits (Yusupov et al., 2001) the entire vacant ribosome (Schuwirth et al., 2005) and during translation (Mitra et al., 2005). These and other studies have illuminated much of the structural basis of translation. The significance and importance of high-resolution structural detail of ribosomes is such that it has led to the creation of a database specifically for aligned ribosome structures (Jarasch et al., 2011).

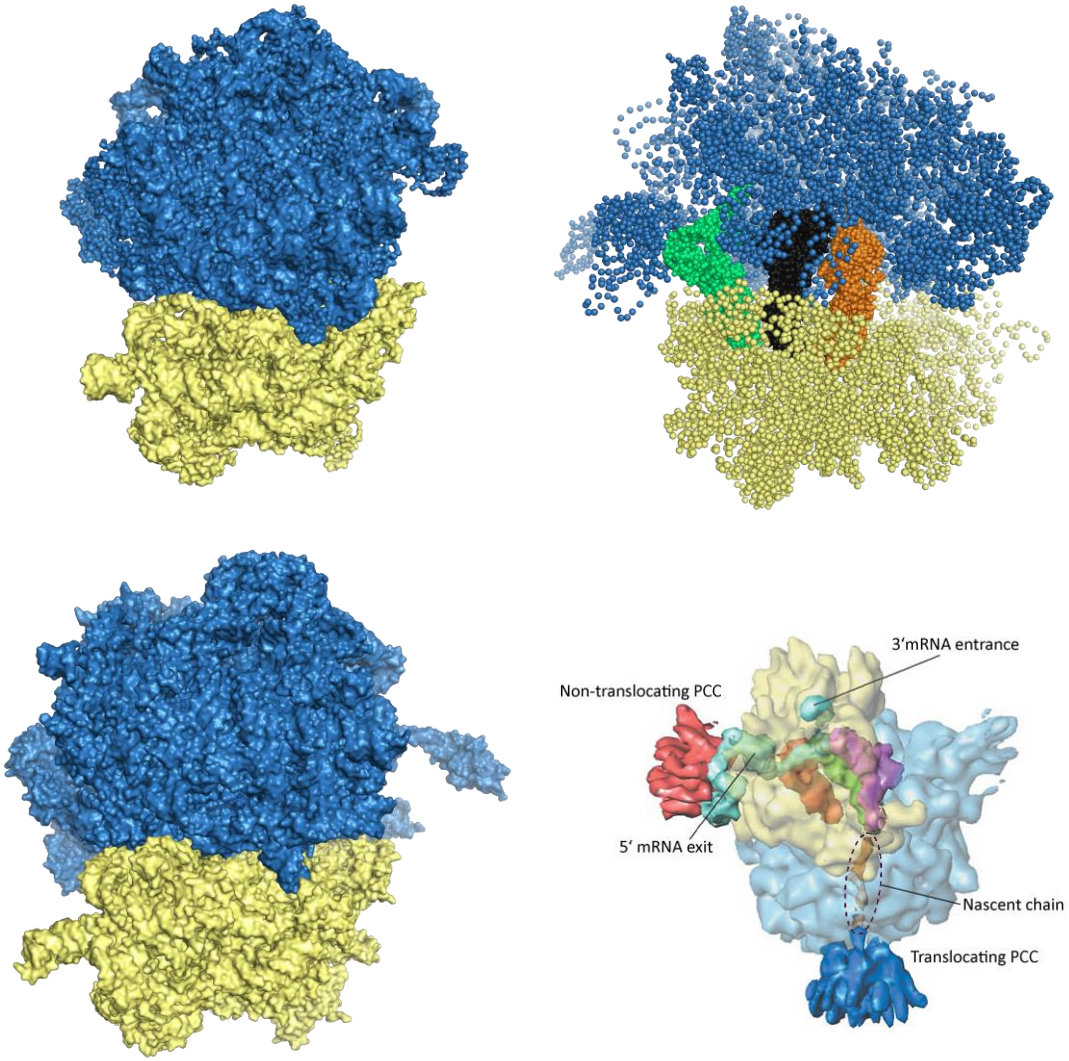

Figure 1.1 Solvent-accessible surface models showing the progressive improvements in ribosome structures. Blue: the large, 50S subunit, yellow: the small, 30S subunit. A) The separate subunits were initially solved with a small proportion (~60% by mass) of the atoms observed. Here, the separate subunits (Ban et al., 2000) have been docked using information derived from later whole-ribosome solutions. B) Low resolution but more-complete structure (Yusupov et al., 2001). C) High resolution, high-coverage, vacant ribosome (Schuwirth et al., 2005). By comparison to the (A), which is oriented the same as (C) and based on earlier data, the far greater completeness of the structure is apparent. D) Cryo-EM density map revealing nascent chain within a translating ribosome bound to a protein conducting channel (PCC), adapted from Mitra et al., 2005.

1.1.1 Structure

Typical prokaryotic ribosomes are 2.5MDa in mass, while eukaryotic ribosomes are approximately 3.3MDa, but the precise composition of ribosomes varies between species and over time. For example, in *Arabidopsis thaliana* leaf cells, expression of specific ribosomal proteins is altered in response to phytohormone signalling, which is suggested to cause differential mRNA preference, thereby providing an additional method of regulating mRNA translation (Schippers and Mueller-Roeber, 2010). Moreover, the structure of ribosomes is dynamic, changing the relative orientation of its subunits during translation to facilitate the passage of mRNA and tRNA, and the orientations of protruding ribosomal protein domains alter to allow recognition of correct tRNAs. For these reasons, high-resolution models of ribosomes are very difficult to obtain.

High-resolution structures are now available for both prokaryotic and eukaryotic ribosomes in key stages of function (Ben-Shem et al., 2010, Julián et al., 2008, Selmer et al., 2006), revealing the atomic-level details of *in-vivo* protein synthesis, though the details of structure of the nascent peptides have still to be discovered.

The eukaryotic ribosome is considerably (~32%) larger than the prokaryotic ribosome, though they contain the same ribonucleic acid catalytic core and there are homologous prokaryotic ribosomal proteins for most eukaryotic ribosomal proteins. The added complexity of recording data for such a large macromolecule, as well as the increased difficulty in handling eukaryotic systems has made data for the eukaryotic ribosome less forthcoming. For this reason, the remainder of this thesis will refer exclusively to the prokaryotic ribosome.

A translating ribosome consists of a large (50S) and a small (30S) subunit that, clasped either side of a mRNA molecule, is broadly reminiscent of a spherical, 200Å-diameter bead on a string (Figure 1.2A). At the core of this – at the centre of the interface between the two subunits and the mRNA – is the peptidyl transferase centre (PTC), at which amino acid residues are peptide-bound to eventually form the protein. The nascent chain egresses through a narrow tunnel that runs approximately 100Å through the 50S subunit roughly orthogonal to the

path of the mRNA between the subunits. All proteins must pass through this exit tunnel before entering the cytosol.

The PTC is formed of RNA, leading to the ribosome being considered a ribozyme. Indeed, approximately 1.1MDa of the mass of the ribosome is formed of RNA. The PTC is highly conserved across species both in sequence and structure. The ribosome incorporates three RNA molecules – one in the small subunit (16S rRNA) and two in the large subunit (5S and 23S rRNAs) – and 55 different proteins – 22 in the small subunit (S1 to S22) and 34 in the large subunit (L1 to L36), with proteins S20 and L26 being two copies of the same protein. Proteins L7 and L12 are identical except that acetylation of serine-2 converts L12 to L7. L7 and L12 form a homodimer, 2 copies of which associate with L10 to form complex called the ribosomal stalk but which historically has been assigned the label L8. The ribosomal stalk protrudes from the ribosome and has a particularly dynamic structure, which allows it to interact with GTP-bound translation factors. It has, therefore, proven difficult to obtain high-resolution structures of the stalk attached to the ribosome.

Within the cleft between the 50S subunit and the 30S subunit are three discrete spaces that accommodate transfer RNA (tRNA) – molecules that provide pairing between a triplet of specific nucleotides on the mRNA and a specific amino acid, to which it may be bound. The three sites – acceptor (A), peptidyl transfer (P) and exit (E) (Figure 1.2B) – accommodate the tRNA molecules as they are recognised, allowed to add their amino acid cargo to the nascent chain and released respectively. A notation has been derived to denote that tRNA molecules may be loaded with a specific amino acid (aminoacylated tRNA) and recognise another specific codon on the mRNA. A tRNA that is loaded with *N*-formlymethionine (fMet) but that recognises the methionine codon (Met), for example, would be denoted as fMet-tRNA^{Met}.

The exit tunnel in the 50S subunit is narrow (10-20Å wide, Figure 1.2A) – accommodating only a linear nascent chain through the first 80Å – and contains a sharp kink approximately 30Å from the PTC, which has been implicated to have a role in regulation of translation. The final 20Å, termed the exit port, broadens significantly before reaching the outer surface of the ribosome. The solvent

accessible surface of the exit tunnel is formed principally by the 23S rRNA and proteins L22 and L4. The exit port, however, is formed by ribosomal proteins which provide the binding sites for molecules that act co-translationally, including the Universal Adapter Site 1 (UAS1), formed by L25 and L35, and Universal Adapter Site 2 (UAS2), formed by L31 and L17 (Pech et al., 2010). Binding partners include the Signal Recognition Particle – which is involved with translocation of nascent peptides to the endoplasmic reticulum – and peptide deformylase – which modifies the N-terminus of nascent peptides before it may be buried by the folding of the protein.

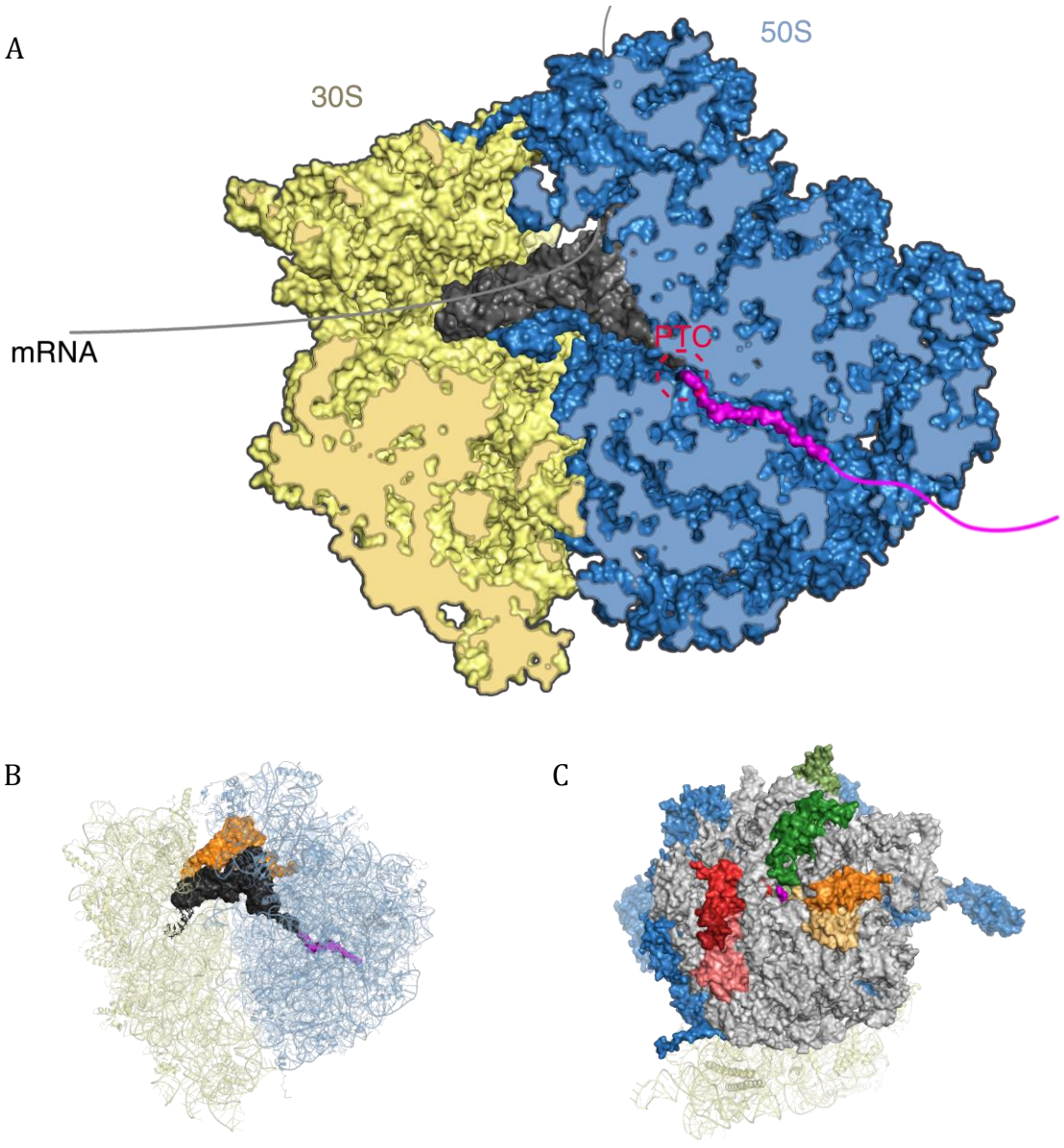

Figure 1.2 Schematic of translating ribosome. A) Cross section through a translating ribosome to showing the P-site tRNA (black), with the nascent chain (magenta)

egressing through the exit tunnel of the 50S subunit. B) View from above the ribosome (relative to A) in which the 50S and 30S subunits are translucent and the P-site (orange) and A-site(green) tRNAs are visible, as well as a portion of the mRNA (black) and the egressing nascent chain. C) Exterior view facing the exit port. The 30S subunit is shown as a yellow cartoon below and behind the 50S subunit, shown as the solvent-accessible surface. The 23S and 5S rRNAs are shown in grey with the 50S ribosomal proteins coloured. In the centre of the exit port, a nascent chain can be seen emerging from the exit tunnel in magenta. On the right side of the exit port in shades of orange is the Universal Adapter Site 1 (UAS1) formed by L29 (above) and L23 (below). On the left side of the exit port in shades of red is UAS2, formed by L22, L32 and L17 from top to bottom respectively. UAS1 and UAS2 form binding sites for cofactors that interact with nascent chains under different conditions. L24 (dark green), however, forms part of the exit port and may interact with nascent chains directly. Also visible within the tunnel and on the top exterior of the 50S subunit is L4 (pale green), which forms part of the exit tunnel wall.

1.1.2 Function

The principle role of ribosomes is, of course, to translate with high fidelity a messenger RNA sequence into a polypeptide sequence that will become a functional protein molecule. As already alluded to, however, the ribosome has ancillary roles in regulation of protein expression, translocation of nascent chains across membranes and facilitating post-translational modification, but is also increasingly becoming understood to play a role in affecting the structural properties of the nascent chain as will be discussed in section 1.2.1. Many of the functional dynamics of the ribosome have been elucidated using computational techniques (Munro et al., 2009, Whitford et al., 2010).

Translation of an mRNA molecule by a ribosome is commonly divided into three stages: initiation, elongation and termination (Ramakrishnan, 2002). Subsequent recycling of the ribosome is variably considered either a separate, fourth stage, or part of the termination stage.

Initiation of translation involves the 30S subunit, three protein initiation factors (IF1-3) and an initiator tRNA (fMet-tRNA^{fMet}) (Figure 1.3, top left). A specific Shine-Delgarno sequence on the mRNA is recognised by the 16S rRNA in the 30S subunit. The complimentary base pairing between the two causes the initiation codon, AUG, of the mRNA to be placed adjacent to the A-site of the 30S subunit.

IF1 binds to the A site and induces a conformational change in the 30S subunit akin to a transition state between the 50S-associated and unassociated forms. IF2 is a GTPase that binds to the A site and hydrolyses GTP to allow the 30S and 50S subunits to associate. IF3 occupies the E site and destabilises the binding of any tRNA at the P-site but the initiator tRNA. At this stage the 50S subunit becomes associated and elongation may begin.

A ternary complex of Elongation Factor Thermo unstable (EF-Tu), GTP and aminoacylated tRNA (EF-Tu· GTP· tRNA) binds to the A site of the ribosome (Figure 1.3, top centre). The correct codon being present is recognised by the 23S rRNA, which, upon binding of non-cognate tRNA, alters the ribosome geometry, preventing elongation from proceeding (Ogle et al., 2001). Once cognate tRNA has been bound to the A-site, EF-Tu hydrolyses the bound GTP, releasing both EF-Tu and GDP, which permits the subsequent accommodation of the tRNA into the PTC. The P-site tRNA is then spontaneously deacylated and the peptide chain is transferred to the A-site tRNA. GTP-bound elongation factor G (EF-G· GTP) binds to the A-site tRNA, which shifts it into an intermediate A/P postion, and the P-site tRNA into an intermediate P/E position. EF-G hydrolyses the GTP, which results in the A/P-site tRNA to move fully into the P-site, the P-site tRNA to move into the E site and the mRNA to move through the ribosome by one codon. Following this translocation, the EF-G· GDP dissociates, leaving deacylated tRNA in the E-site, the nascent-chain-bound tRNA in the P site and the A site vacant, ready for another round of elongation.

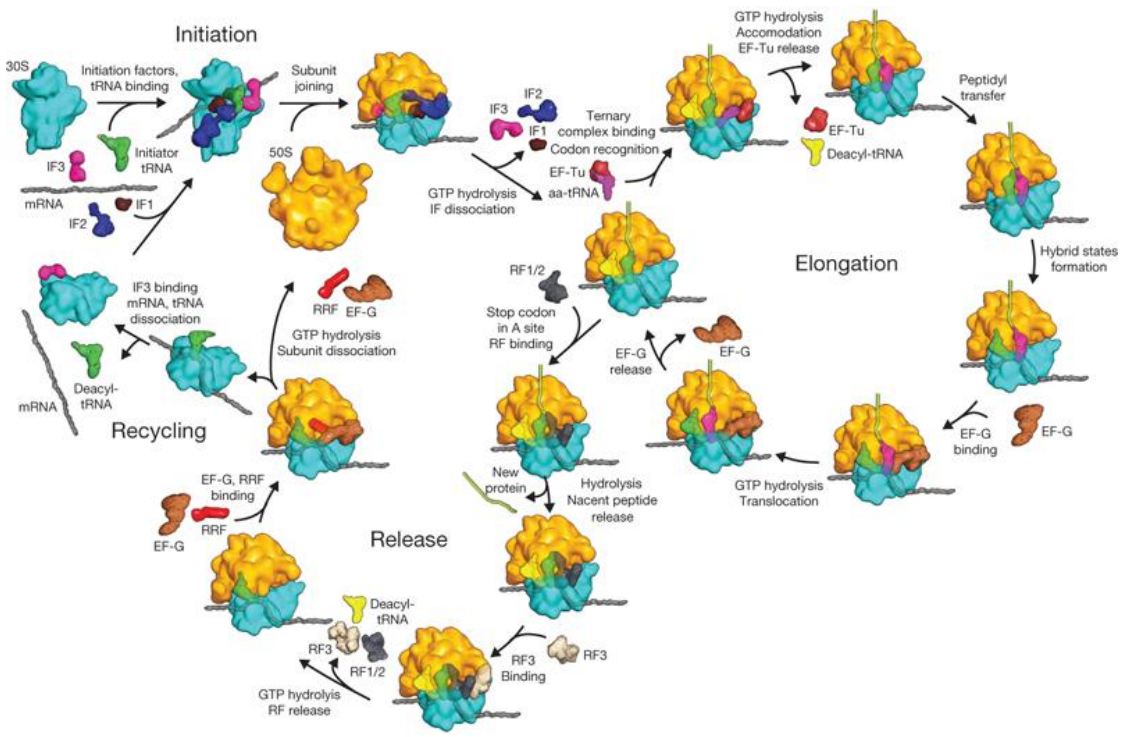

Figure 1.3 Schematic showing the key stages in ribosomal translation. Adapted from (Schmeing and Ramakrishnan, 2009).

When a stop codon is present on the mRNA at the A site of the translating ribosome, one of two release factors, RF-1 and RF-2, bind to the A-site (Figure 1.3, centre). Both release factors recognise the codon UAA, but only RF-1 recognises UAG and only RF-2 recognises UGA. RF-1/2 hydrolyses the P-site tRNA, resulting in the release of the nascent peptide chain. After this hydrolysis, RF-3 may bind to the ribosome and promote the rapid dissociation of RF1/2 (Zavialov et al., 2001).

Ribosome release factor (RRF), a structural mimetic for tRNA, binds to the A-site of the ribosome following termination of translation and, with the assistance of EF-G, hydrolyses GTP to dissociate the 50S and 30S subunits. IF-3 is then needed to dissociate the deacylated tRNA from the P site of the 30S subunit and the subsequent release of mRNA. The ribosome components and factors are now available to be recycled.

1.2 The Ribosome-Nascent chain Complex

Recently, research has focused on the role of the ribosome in the folding pathway of proteins. It has been shown, for example, that the dynamic structure of the ribosome is essential in allowing the proper folding of some proteins (Bashan and Yonath, 2005, Zhang et al., 2009), with the dynamic structure of the protein exit tunnel allowing and possibly guiding the folding of proteins prior to termination of translation (Osapay and Case, 1991, Etchells and Hartl, 2004). Prior computational investigations have suggested that the exit tunnel has a carefully tuned interaction with the nascent chain to allow the nascent chain to efficiently emerge as if the exit tunnel were Teflon-coated (Fulle and Gohlke, 2009, Petrone et al., 2008). This knowledge, combined with the known interactions between nascent peptide chains and chaperones (Kaiser et al., 2006) has led to more sophisticated models of protein folding and aggregation during biosynthesis (Clark, 2004). Ribosome-bound nascent chains may occupy conformational space not explored by the full-length protein in the absence of a ribosome, as is discussed in section 1.2.1.

1.2.1 Co-translational folding

Studies of protein folding have traditionally involved folding and denaturation of full proteins in isolation (Hartl and Hayer-Hartl, 2009). Such studies suggest that the isolated proteins adopt only a narrow range of conformations on a pathway to the native structure (Fersht, 1995, Englander and Mayne, 1992, Baldwin, 1993, Dobson et al., 1994). It has become generally acknowledged, however, that proteins may undertake multiple routes to their final structure, and that their native structure often exists as a range of similar conformations in stable equilibria (Villali and Kern, 2010, Sosnick and Barrick, 2011).

While efforts have been made to consider the effects of macromolecular crowding, salt concentrations and other properties of the cytosol (Rudolph and Lilie, 1996, van den Berg et al., 1999), until relatively recently they have neglected the effects that the ribosome may play as the protein progressively emerges from the exit tunnel. Indeed, there are two potential sources for investigation: the effect of additional C-terminal amino acid residues becoming

folding-competent on timescales equal to or greater than the rate of folding; and the effects that the ribosome exit port may play as a catalyst in folding of the nascent protein via its steric, electrostatic or dynamic properties.

The folding rate of proteins in isolation varies considerably from microseconds (Qiu et al., 2002) to hours (Goldberg et al., 1990) and has spawned an entire field for predicting the folding rate of proteins from the amino acid sequence, the best approaches of which use structure prediction to predict the contact order, which is the most significant factor affecting folding rate. Measurements of co-translational folding rate are considerably more difficult, however a study of the large Cystic Fibrosis Transmembrane conductance Regulator (CFTR) protein — a model multidomain protein used extensively in studies of co-translation folding as its large size necessitates folding before ribosomal release (Neal et al., 2003) — in which folding ability was monitored by FRET as the nascent chain was artificially released from the PTC showed that 75% of the population of the folding-competent residues 389–500 of NBD1 were folded within 2 minutes (Khushoo et al., 2011).

Similarly, the rate of peptide synthesis can vary, though is typically on the order of $10–10^2$ peptide bonds min^{-1} in prokaryotes (Zhang et al., 2009). The rate of synthesis, which matches the example rate of *de novo* folding given above, is most commonly the factor limiting the rate of protein folding *in vivo*. Moreover, there is evidence that by controlling the rate of peptide synthesis, organisms may control the folding pathway of proteins; synonymous single nucleotide polymorphisms that involve an abundant codon being swapped for a rare form in the Multidrug Resistance 1 (MDR1) protein leads to a change in the folding path energy landscape that favours an similar structure that has altered substrate specificities (Kimchi-Sarfaty et al., 2007, Tsai et al., 2008).

Although multiple nucleotide codons may result in the incorporation of the same amino acid, the tRNAs that correspond to each of those codons are not present in the same concentration with some being abundant while others may be much more rare. For example, in E. coli, the tRNA that recognises arginine codon CGG is present at approximately 640 copies per cell, whereas the tRNA that recognises arginine codons CGU/CGA, CGC is present at approximately 4700 copies per cell

(Dong et al., 1996). When a rare codon is encountered, the rate at which the correct tRNA is inserted into the A site (and therefore the next amino acid is incorporated) will substantially slow the rate of protein synthesis. This effect may be used for post-transcriptional protein regulation (Kuhar et al., 2001). However, it may also be used to interfere with the co-translational folding dynamics of nascent chains. A similar effect may be achieved by use of mRNA secondary structure to impede translocation, mRNA stability (Ivanov et al., 1997, Chamary et al., 2006) or a long stretch of synonymous codons to temporarily deplete the tRNA availability (Purvis et al., 1987).

While there is evidence in specific cases that due to the slow rate of peptide synthesis relative to the rate of folding that co-translational folding occurs (Komar, 2009) and is even necessary in, for example, allowing the stable folding of Suf1 (Zhang et al., 2009), the role – if any – that the ribosome plays in directing the folding pathway at the surface is still being debated.

1.2.2 Methods of preparation

Studies of co-translational folding typically require significant periods of time – much longer than the rate of translation would allow – rendering the real-time study of co-translational folding impractical. A single NMR dataset might take from hours to days to acquire, for example (Christensen et al., 2011). For this reason, structural studies of co-translational folding typically involve pausing ("stalling") translation at multiple specific points in the translation of a particular protein so as to assemble a series of models of the protein as it emerges from the ribosome. Since the nascent chain is ejected from the ribosome as soon as a stop codon enters the A site of the ribosome, preparation of RNC samples revolves around various methods of preventing a stop codon entering the ribosome.

Early methods involved removal of the stop codon from the mRNA, either by engineering the DNA prior to transcription or by enzymatic degredation of the mRNA during translation (Gilbert et al., 2004, Hsu et al., 2007). These methods are both substantially flawed by the transfer messenger RNA (tmRNA) surveillance mechanism, which stimulations the release of stalled nascent chains by trans-translation as well as the decay of causative mRNAs (Dulebohn et al., 2007).

Precise protein expression has for many years been conducted using *in vitro* expression systems in which the precise constituents can be controlled, a mixture consisting of ribosomes, mRNA and the other necessary molecules for protein synthesis are present (Lakshmipathy et al., 2007, Matsuura et al., 2007, Takahashi et al., 2009). It is possible to modify this approach to allow RNC preparation by using a mRNA transcript with a single codon for a specific amino acid – usually cysteine – and ensuring that no Cys-tRNA^{Cys} is present in the reaction mixture. When this codon is encountered by a ribosome, translation is stalled and a RNC is formed. This approach has been used, for example, to monitor force–extension profiles of T4 lysozyme RNC by optical tweezers (Kaiser et al., 2011) although samples produced in this manner have lifetimes too short for high-resolution study.

An alternative approach is to use the translation arrest motifs – short sequences of amino acids that lead to translation arrest – from the tryptophan operon (tnaC) or secretion monitoring protein (SecM). The SecM stalling motif (commonly referred to as simply SecM) is a 17 amino acid sequence (FxxxxWIxxxxGIRAGP) that causes translation to stall in the pre-translocation stage (Figure 1.3) with Gly-tRNA^{165} in the P site and Pro-tRNA^{166} in the P-site (and therefore not peptide bound to the nascent chain) (Bhushan et al., 2011, Nakatogawa and Ito, 2002). As seen in Figure 1.4, cryo-EM maps of the SecM sequence stalled within the ribosome exit tunnel have shown that the specific positioning of Arg163 (residue 14 of the SecM motif) leads to its interaction with A2062 of the 23S rRNA. Previous studies have suggested that A2062 acts as a nascent peptide sensor and that its interaction with S2503 triggers inactivation of the PTC (Vázquez-Laslop et al., 2010). The cryo-EM data suggest that this is caused by repositioning of the P-site tRNA A76 such that peptide bond formation between the A-site Pro-tRNA^{Pro} and P-site Gly-tRNA^{Gly} is inhibited.

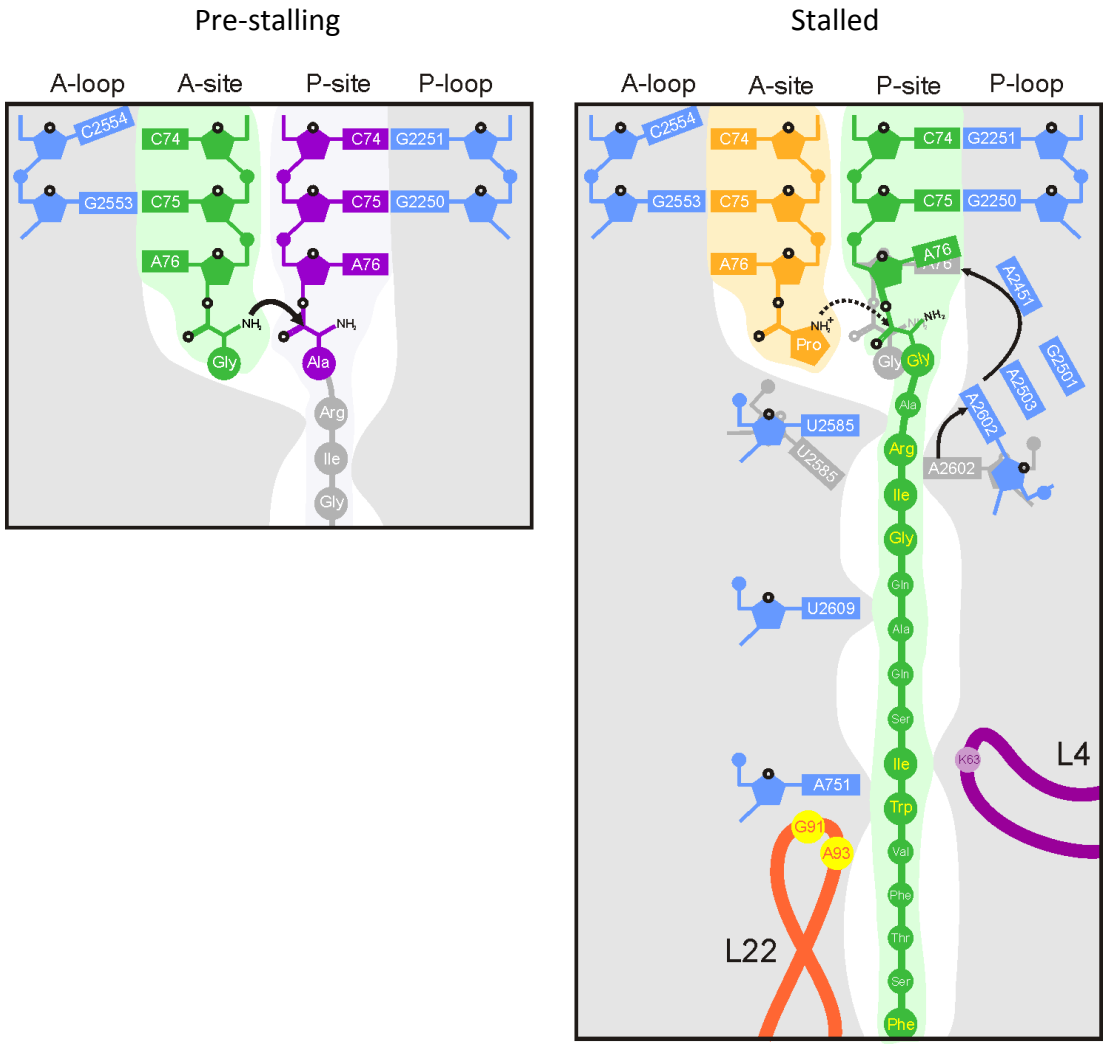

Figure 1.4 Schematic of ribosome PTC and exit tunnel prior to and during SecM stalling. Left: residue 16 of the SecM stalling sequence is present in the A site (Gly-tRNA^{165} of the original SecM sequence). The P-site tRNA is positioned so as to allow nucleophilic attack by the amino group of the A-tRNA (black arrow). Right: interactions between the SecM motif and the exit tunnel cause the P-site SecM-tRNA^{Gly} to be repositioned such that the nucleophilic attack (dotted line) is inhibited. Adapted from (Bhushan et al., 2011).

Recently, advances have been made in use of NMR to study the nascent protein chain while it is still bound to the PTC (Christensen et al., 2011, Hsu et al., 2007, Hsu et al., 2009b, Hsu et al., 2009c). Production of RNCs *in vivo* in E. coli cells harnesses the fact that 25% of the cellular mass is ribosomes. This involves introducing stalling and purification motifs to a cloned protein-coding sequence that stalls translation of the protein at a specific residue and allows subsequent

isolation, affinity purification and study by NMR. The SecM motif (Evans et al., 2005, Nakatogawa and Ito, 2002) is used to efficiently stall translation *in vivo* while not impeding cell viability, while multiple histidine residues (typically hexa-His) are inserted at the N-terminus to allow efficient purification of the RNC (Figure 1.5). Once the RNC has been studied, the nascent chain can sometimes be released, allowing confirmation that that the nascent chain was ribosome-bound by observation of a change in conformation. Unfortunately, puromycin – an aminoacyl-tRNA analog that is commonly used for prematurely releasing nascent chains – is ineffective against SecM (Muto et al., 2006).

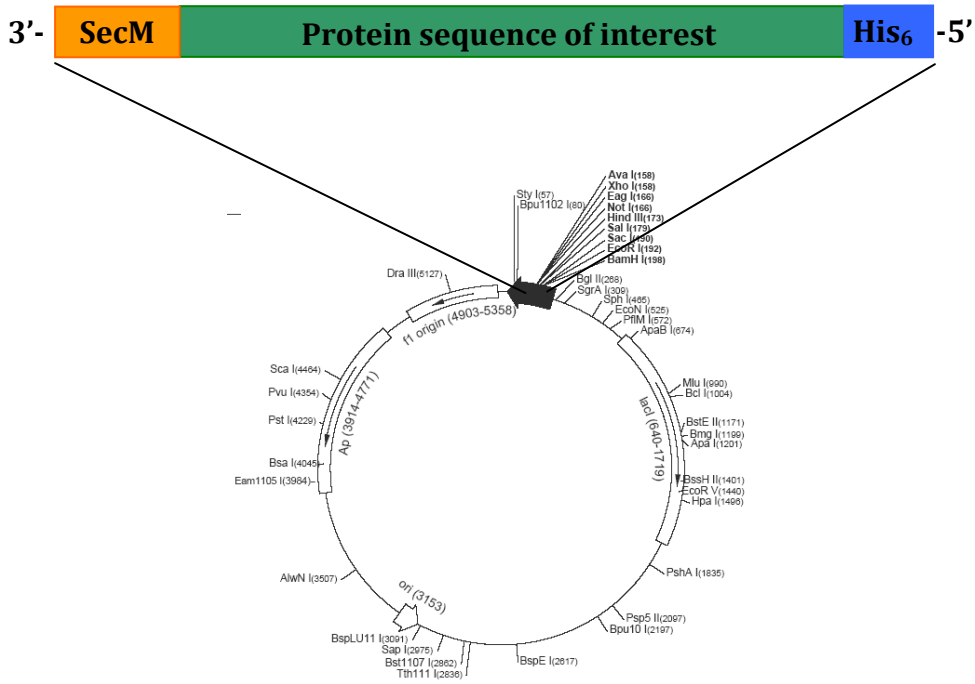

Figure 1.5 Schematic of the pLDC17 plasmid incorporating the sequence of an arbitrary protein of interest. The SecM stalling sequence and approximately 13 residues of the protein of interest remain within the exit tunnel of the ribosome. Adapted from (Oldfield, 2002).

1.2.3 Techniques for studying RNCs

While vast improvements in the techniques for preparations of RNCs, samples still suffer impediments to their study, including the innate stability of the ribosome limiting concentrations to approximately 10μM and the system complexity leading to low sample stability and a commensurate short lifespan of

hours to days before either the nascent chain is released or the ribosome degrades (Waudby et al., 2009).

Recently, cryo-EM studies of stalled ribosomes have been used to study nascent chains (Becker et al., 2009, Seidelt et al., 2009). By recording electron density for a large number of both empty and stalled ribosomes and subtracting the former from the latter before incorporating the signal from all the copies, electron density for nascent chains has been observed. Due to the confined space that may be occupied by the nascent chain within the exit tunnel, the nascent chain in each copy reliably stays within a narrow range of conformations, which allows data from multiple copies to be combined to give greater signal. However, outside of the exit tunnel, a more diverse range of conformations – including less compact linker regions – may be adopted, leading to little signal.

Förster Resonance Energy Transfer (FRET) has been used to investigate proximity between specific regions on the nascent chain and other regions on the ribosome or other regions within the nascent chain (Osapay and Case, 1991). While this allows probing of compaction and folding within the ribosomal exit tunnel, it provides only small numbers of distance parameters for each experiment, thus requiring a large number of experiments to provide atomic-resolution models. It also suffers the need to incorporate a relatively large fluorescent tag if no intrinsic fluorophore is within the target system. Furthermore, fluorescence anisotropy can be used to measure the nanosecond-timescale local dynamics of ribosome-bound nascent chains both within and outside the exit tunnel (Pople, 1958). Such methods have been used to detect independent nascent chain structure on the ribosome (Ellis et al., 2008).

Other common biochemical techniques have also been used to study nascent chain structure on the ribosome, each typically giving small amounts of high-resolution data or large amounts of low-resolution data that are insufficient to provide atomic-resolution models of nascent chains. Such techniques include binding of conformation-sensitive antibodies (Tsalkova et al., 1998, Clark and King, 2001), atomic force microscopy (Loksztejn et al., 2012), single molecule force–extension measurements (Kaiser et al., 2011, Katranidis et al., 2011), limited proteolysis of translating systems (Neal et al., 2003) and acquisition of

native function (Fedorov and Baldwin, 1995, Nicola et al., 1999, Kelkar et al., 2012).

NMR experiments can overcome many of the problems with the techniques outlined above. NMR of protein systems are – in part – highly useful due to the abundance of potential probes intrinsically available; it is possible to obtain NMR signal from most nuclei within a protein system or to selectively silence the signal from most nuclei by careful incorporation of isotopes that yield an NMR signal. It is, therefore, possible to produce RNC samples in which the nascent chain yields an NMR signal while the ribosome does not. Furthermore, the conformational heterogeneity of the nascent chain outside of the ribosomal exit tunnel that renders cryo-EM and XRC ineffective allows for the nascent chain to tumble isotropically independently of the ribosome, resulting in sharp NMR signals.

While the achievable RNC sample concentration and lifetime are too low for the kinds of NMR experiments typically used for acquisition of high-resolution structural data (namely NOE-spectroscopy), it is still possible to perform many other experiments to ascertain structural details including recording the chemical shifts alone, multidimensional spectra, selective excitation and residual dipolar couplings. These can be used for monitoring selective broadening (Hsu et al., 2007), side-chain dynamics (Hsu et al., 2009b), the progressive appearance of native-like crosspeaks (Christensen et al., 2011, Eichmann et al., 2010, Rutkowska et al., 2009). While the these advances give unprecedented structural detail of the cotranslational properties of nascent chains, they still do not make complete use of all the information contained within the data they record.

1.3 The NMR chemical shift

NMR spectroscopy is an exquisitely informative tool with a broad range of applications, including molecular structure determination. In NMR spectroscopy, a sample for study is subjected to a powerful magnetic field – typically between 11.7T and 21.1T for the study of biomolecules – which causes all nuclei within the sample that have an intrinsic magnetic moment to become partially aligned with the external magnetic field.

The chemical shift is an extremely sensitive and precise probe of the electronic and magnetic environment of a nucleus and consequently the molecular structure. The factors that affect the precise chemical shift include the properties of neighbouring nuclei, the angles and lengths of bonds and both through-bond and through space effects, which give information about electric fields arising from nearby aromatic ring currents and polar or charged groups. Although the chemical shift contains all this information, it is in part because of the large number of factors that affect the chemical shift that it has proven difficult to use the chemical shift to directly infer structural characteristics.

1.3.1 The physical basis of NMR chemical shifts

As well as mass, electric charge and magnetism, atomic nuclei possess an important property termed spin. Spin is an abstract concept that is a form of angular momentum, although it is not acquired or lost through rotation of the nucleus – unlike rotational angular momentum – but instead is an intrinsic property of the particle derived from the composition of elementary particles. The total spin angular momentum of a particle with spin, L_{tot}, is given by the expression

$$L_{tot} = \left[S(S+1)\right]^{1/2}\hbar \tag{1.1}$$

where S is the spin quantum number of the particle and $\hbar$ (h-bar) is the Dirac constant. Spin is quantized to discrete values that give rise to a stable particle. S may be an integer (0, 1, 2, …) or a half integer (denoted 1/2, 3/2, 5/2, …). Bosons (such as the photon) have integer spin, while fermions (such as the electron, neutron and proton) have half-integer spin. While S reveals the total angular momentum, it does not reveal the direction of that spin. A particle with spin S may a direction of spin at one of the discrete values from $-S$ to $+S$ at integer values, i.e. it has $2S + 1$ sublevels. In the absence of a magnetic field, these sublevels are degenerate. However, when a magnetic field is applied, each level gives rise to a slightly different energy of the particle, a phenomenon called the Zeeman effect.

Within a nucleus, the spin of individual nucleons (protons and neutrons) may be combined to produce the total spin angular momentum of the nucleus. The spin

of each nucleon may either be additive ("spin up") or subtractive ("spin down") giving rise to a total nuclear spin quantum number, I, and each nucleus has a lowest energy state configuration called the ground state nuclear spin. The difference in energy between possible nucleon spin configurations greatly exceeds the energies experienced chemical reactions or NMR experiments and so may be ignored in this context. However, as with the simpler scenario above, a nucleus with total nuclear spin quantum number I may have $(2I + 1)$ sublevels from $-I$ to $+I$ at integer intervals and in the presence of a magnetic field, these sublevels are at different energy levels, resulting in Zeeman splitting. Nuclei with spin $I = 0$ are called "zero-spin" nuclei and have only one energy level, resulting in no Zeeman splitting. Since these nuclei have no sublevels, they experience no Zeeman splitting. These nuclei include three isotopes commonly found in biological molecules: ^{12}C, ^{16}O and ^{32}S. Fortunately, isotopes that are NMR-visible may often be incorporated instead, such as ^{13}C and ^{17}O, which have spin-1/2 and spin-5/2 respectively.

Although only a small number of substances are *ferromagnetic* – the macroscopic magnetism in which objects are attracted to magnets – all matter is inherently *magnetic*, that is it is capable of interacting with magnetic fields. This is expressed as a *magnetic moment*, μ. The interaction between the magnetic moment and a magnetic field results in a magnetic energy being associated with matter:

$$E_{mag} = -\mu \cdot B \tag{1.2}$$

where E_{mag} is the magnetic energy and B is the magnetic field, which is a vector, having both a strength and a direction. Note that, due to the use of a dot product, E_{mag} is dependent on the relative directions of the magnetic moment and the magnetic field: when the two are parallel, E_{mag} is lowest; when the two are antiparallel, E_{mag} is highest.

The angular momentum of a particle with spin is a vector property with a direction that indicates the axis of rotation that may point in any direction. Similarly, due to the properties of atomic structure, the nucleus also has a vector magnetic moment due to its magnetism. Due to the quantum mechanical origins

of both nuclear spin and magnetism, the magnetic moment, is proportional to spin angular momentum:

$$\hat{\mu} = \gamma \hat{S} \tag{1.3}$$

where $\hat{\mu}$ is the magnetic moment and γ is the gyromagnetic ratio (or magnetogyric ratio), which may be either positive or negative depending on the nucleus. Most atomic nuclei have a positive gyromagnetic ratio and so have a spin angular momentum parallel to the magnetic moment. The circumflex diacritic indicates that the parameter is a quantum mechanical operator that describes a wavefunction rather than a discrete particle.

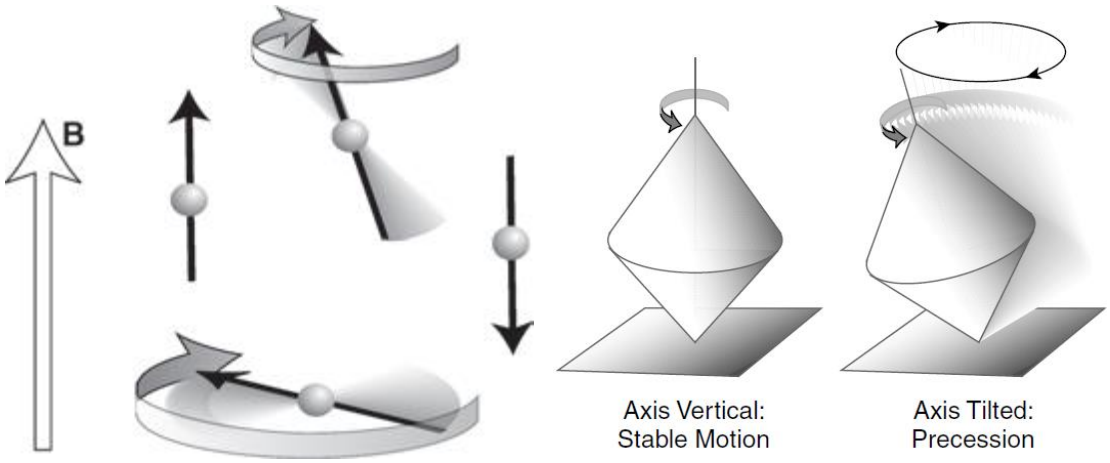

Figure 1.6 Diagram illustrating forms of precession. Left: precession of the nuclear spin about the magnetic field, B, depends on the initial alignment of the spin axis prior to the presence of the field. Right: the spinning top toy is exhibits similar properties in classical physics, wherein the axis of spin will precess about the gravitational field as the angular momentum and weight of the object interact. Adapted from (Levitt, 2008).

In a sample that is in equilibrium in the absence of a magnetic field, there is no net alignment of the individual magnetic moments of the constituent nuclei and the sample is said to be isotropic. Upon application of a magnetic field to the sample, the magnetic moment would be drawn towards aligning with the field lines. However, as seen in Figure 1.6, due to the angular momentum the spin of each nucleus instead precesses around the field, maintaining a fixed angle between the spin angular momentum and the direction of the magnetic field that is dependent on the initial spin polarization. If the initial spin angle was perpendicular to the field, it

will precess about a flat disc; if it was initially parallel, it will stay there. Most nuclei will be somewhere between these extremes.

The frequency of the precession, ω^0, is related to the gyromagnetic ratio and magnetic field strength, B^0, as follows:

$$\omega^0 = -\gamma B^0 \tag{1.4}$$

The frequency of precession of nuclear spins is called the Larmor frequency and is usually measured as the number of cycles per second, Hz. However, for equations describing the properties of nuclei, it is simpler to use Larmor frequencies measured in radians per second, which is the frequency in hertz multiplied by 2π. The equations presented here, including equation (1.3), are given for units of radians per second. The sign of the Larmor frequency indicates the direction of spin precession with a negative Larmor frequency corresponding to clockwise precession as observed in the opposite direction to the magnetic field.

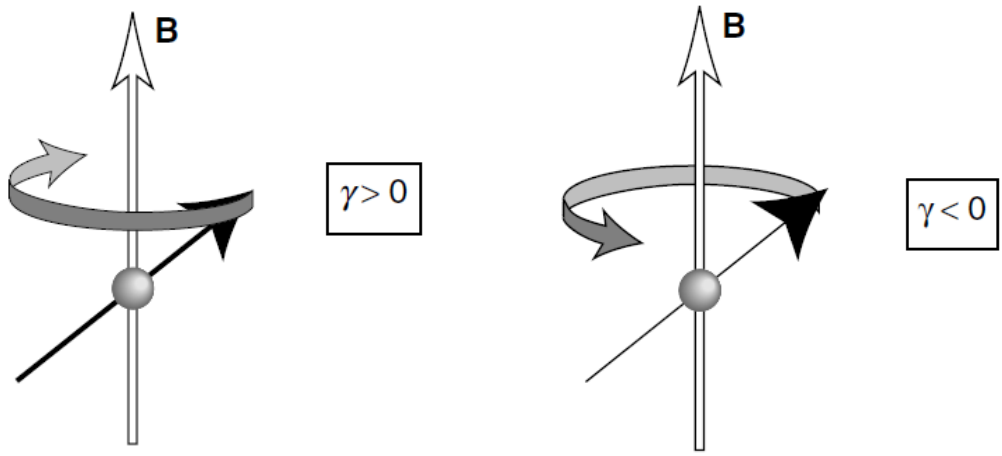

Figure 1.7 Diagramatic representation of precession for nuclei with positive (left) and negative (right) gyromagnetic ratios, γ. Note that due to the relationship described in equation 0, a positive gyromagnetic ratio corresponds to a negative Larmor frequency. Adapted from (Levitt, 2008).

Within a sample in a magnetic field, the nuclei experience thermal interactions with neighbouring particles, slight changes in magnetic field direction and

magnetic shielding from neighbouring particles. These all act to cause the cone angle of precession to change over time. Due to the effect of the applied magnetic field, those orientations parallel to the magnetic field are slightly more energetically favourable than those antiparallel. This gives rise to anisotropic distribution of nuclear spin polarisations. Individual nuclear spins will continue to wander, but after a period of time, the net alignment will not significantly increase and the sample is said to be at thermal equilibrium. The acquisition of net longitudinal nuclear spin polarization is approximately exponential, governed by the spin-lattice relaxation time constant, T_1, which depends on the isotope and sample conditions, including viscosity and temperature.

The longitudinal magnetization is of such low strength relative to the applied magnetic field that it is almost undetectable and so not feasible for study. However, by applying a suitable radiofrequency, rf, pulse, it is possible to rotate the average axis about which the nuclei are precessing – and so rotate the net nuclear magnetization – by π/2 radians such that the net magnetization is now pointing in a single direction perpendicular to the applied magnetic field. This is called transverse magnetization. By convention, the axis of magnetization is the *z* axis, while the plane perpendicular to this is composed of the *x* and *y* axes. A rotation by π/2 about the *x* axis leads to net magnetization along the –*y* axis, as shown in figure Figure 1.8.

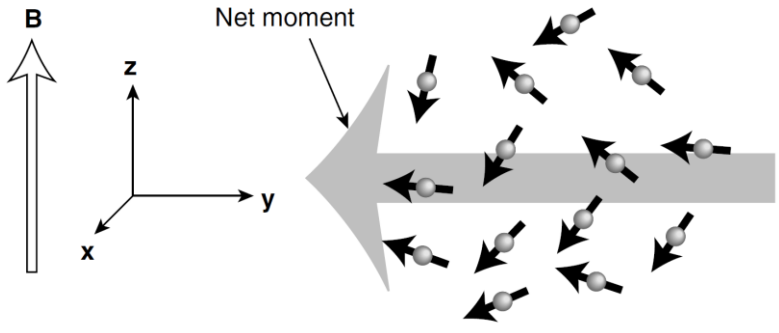

Figure 1.8 Diagram depicting the effect of a rotation of nuclear magnetization by π/2 radians about the *x* axis. The dark arrows indicate the axes about which the individual nuclear spins are precessing. Adapted from (Levitt, 2008).

Once the rf pulse is discontinued, the net moment begins precessing about the applied magnetic field through the *xy*-plane, perpendicular to the *z*-axis. As

previously described by equation (1.3), the frequency of this precession is determined by the Larmor frequency of the individual nuclei. The transverse magnetization precession relaxes over time, giving rise to a decaying magnetization signal governed by the transverse relaxation time constant T_2. For small molecules in liquids, both T_1 and T_2 are on the order of seconds. For large molecules or solids, T_2 may be of the order of milliseconds.

The precession of transverse magnetization can be used to induce an electric field in a coil near to the sample. The signal is small, but because it is orthogonal to the applied magnetic field, it is detectable with a very sensitive rf detector. It also occurs at a very specific frequency range so that instrumentation can be highly tuned to detect it. This signal is the NMR signal of free-induction decay (FID). This FID signal can contain an enormous wealth of data concerning the atomic-resolution properties of the sample.

The NMR spectrometer performs three principle tasks:

1) Induce spin polarization with a strong magnetic field,
2) Rotate the spin polarization carefully with precise rf pulses,
3) Detect and record the very small rf signal given by the transverse spin magnetization precessing.

The recorded FID data are subsequently processed – an entire field of study alone – to reveal a plethora of structural details. Essentially, the oscillating FID signal undergoes a Fourier transform, which calculates the component frequencies that are required to give rise to the FID waveform. This yields a plot of component frequencies versus amplitude. The component frequencies correspond to the Larmor frequencies of the nuclei within the NMR spectrometer sample.

The Larmor frequency for a nucleus within a NMR experiment, Ω^0, are typically expressed as their recorded value, ω^0, relative to that of a reference frequency, ω_{ref}:

$$\Omega^0 = \omega^0 - \omega_{ref} \tag{1.5}$$

Due to technical limitations, in order that accurate and sensitive readings of NMR spectra are recorded, each detector of rf signal is tuned to a narrow frequency range of approximately 1MHz, referred to as a '*channel*'. A spectrometer usually has multiple channels tuned to isotopes of interest, such as ^{1}H or ^{13}C, and the central frequency of each channel is the reference frequency, ω_{ref}.

As the Larmor frequency depends on the strength of the magnetic field, variations in the magnetic field will lead to variations in the Larmor frequency. Furthermore, since matter has a microscopic effect on the local magnetic field, the field strength experienced by individual atoms will vary depending on the nature of the matter around it. It is for this reason that two nuclei of identical composition may give rise to different resonant frequencies. The two important factors affecting microscopic magnetic field variations, and therefore the Larmor frequency of the corresponding nuclei are: the local electronic environment (the '*chemical shift*') and the presence of other magnetic nuclear spins ('*spin-spin coupling*').

The size of the chemical shift depends on the magnetic field: the greater the magnetic field, the greater the size of the shift. However, since both the Larmor frequency and chemical shift are linearly proportional to the magnetic field (to a very good approximation), the field-independent chemical shift, δ, can be calculated:

$$\delta = \frac{\omega^0 - \omega^0_{TMS}}{\omega^0_{TMS}} \tag{1.6}$$

where ω^0_{TMS} is the is a Larmor frequency of the same isotope as the nucleus of interest in a reference compound (in this case tera-methyl silane, TMS) exposed to the same magnetic field. TMS consists of four methyl groups bound to a silicon atom. It is frequently used as a reference compound as it contains multiple carbon and hydrogen atoms in a very similar chemical environment, giving rise to a strong, consistent and precise NMR signal. It can therefore be reliably used to compare both ^{13}C and ^{1}H chemical shifts. By definition, therefore, TMS spins have chemical shift δ = 0. Since δ values are very small (typically less than 2×10^{-4} in ^{13}C spectra, or 10^{-5} in ^{1}H spectra), values are quoted as parts per million, ppm, where 1ppm = 10^{-6}.

Spin-spin coupling can occur in different ways and will affect both the peak position and shape. Direct dipole-dipole coupling between neighbouring nuclei is the strongest, but is also removed by the tumbling of molecules. Indirect dipole-dipole coupling, commonly called *'J-coupling'*, occurs when a nucleus weakly magnetizes the molecular electrons, which affects the magnetic field experienced by other neighbouring nuclei.

J-coupling causes peaks to be split due to interactions between nearby nuclei, which causes peak intensities to be diminished, and leads to crowded spectra. It can be eliminated by *heteronuclear decoupling*, wherein the FID is recorded on one channel, whilst an rf pulse tuned to the Larmor frequency of a different nucleus is applied to the sample. Nuclei for the latter frequency will not cause splitting in the spectrum of the former.

1.3.2 Factors affecting chemical shift

As discussed in section 1.3.1, magnetic matter in a magnetic field has an energy associated with it that depends on the magnetic field strength. Also discussed was the fact that the strength of the magnetic field experienced by individual nuclei is affected by the surrounding matter. Therefore, while the sample under examination is placed in a NMR spectrometer with a sophisticated, macroscopically-uniform magnetic field, B_0, the individual nuclei actually experience a slightly-different effective local magnetic field, B_{eff}. This should then be substituted into equation 0 to determine the actual frequency of precession:

$$\omega = -\gamma B_{eff} \tag{1.7}$$

B_{eff} is caused by an additional microscopic magnetic field experienced by nuclei as a result of an induced magnetic field from neighbouring electrons, B_{ind}, which can either act enhance or subtract from the applied magnetic field:

$$B_{eff} = B^0 + B_{ind} \tag{1.8}$$

The nature and effect of the magnetic shielding or deshielding experienced by each nucleus can be described by a *nuclear shielding tensor*, σ:

$$B_{eff} = (1-\sigma)B^0 \tag{1.9}$$

It is this phenomenon that yields the chemical shift. As mentioned in the opening of section 1.3, the chemical shift is affected by many factors, including local through-bond effects (e.g. bond geometries, the proximity of functional groups, the identities of neighbouring atoms, bond hybridisations) and through-space effects (e.g. proximity to aromatic ring currents or charged and polar groups). The various contributions are sufficiently independent that a formula for the chemical shift can be given as

$$\Delta\delta = \delta_{total} - \delta_{rc} = \delta_{tor} + \delta_{ring} + \delta_{HB} + \delta_{e} + \delta_{side} + \delta_{misc} \tag{1.10}$$

where δ_{rc} is the random coil ("intrinsic") chemical shift for the specific amino acid residue, δ_{total} is the measurable chemical shift for the residue, δ_{rc} is the component resulting from the specific amino acid's intrinsic random coil contribution, δ_{tor} is the contribution of the backbone torsion, δ_{ring} is the ring current contribution, δ_{HB} results from hydrogen bonding, δ_{e} from the local charge (electric field), δ_{side} from the side chain torsion, and δ_{misc} from other experimental factors such as covalent bonding, temperature and solvent. A meta-analysis of literature has previously been performed to establish the approximate relative contribution of each factor to the final chemical shift and is reproduced in Table 1.1.

Attribute	Approximate contribution to chemical shift (%)					
	^{1}HN	^{15}N	$^{1}H_{\alpha}$	$^{13}C_{\alpha}$	$^{13}C_{\beta}$	^{13}CO
Random coil	0	50	25	50	75	25
Torsions (φ/ψ)	0	0	50	25	10	50
Torsions (φ/ψ_{i-1})	25	25	0	0	0	0
Side chain (χ)	5	0	0	5	5	5
Side chain (χ_{i-1})	5	5	0	0	0	0
Hydrogen bonds	25	5	5	5	0	5
Ring currents	10	0	10	0	5	5
Local charges	10	0	0	0	0	0
Miscellaneous	20	15	10	5	5	10

Table 1.1 Contribution of factors affecting the chemical shifts of amino acid residue backbone nuclei. Reproduced from (Wishart and Case, 2002).

The random coil, δ_{rc}, contribution refers to the chemical shift associated with the nucleus when the molecule of which it is a component is unfolded, that is, it is free to sample all sterically feasible conformations (Bundi and Wüthrich, 1979). For simplicity, is often considered to be the chemical shift associated with the nucleus in a specific amino acid, as the covalent structure is the most significant factor affecting the random coil contributions and the through-space effects are assumed to average to zero. Deviations from the random coil value are referred to as secondary chemical shifts, $\delta_{secondary}$, as they are considered to arise from secondary structure of the polypeptide, which is defined as

$$\delta_{secondary} = \delta_{\exp} - \delta_{rc} \tag{1.11}$$

where δ_{exp} is the specific experimentally recorded chemical shift.

Efforts have been made to define values of δ_{rc} for each nucleus of each amino acid using mimic peptides, analyses of chemical shift databases and consideration of neighbouring residues.

Ring current contributions arise from the nearby presence of aromatic rings in which the circulation of electrons in the presence of a magnetic field are modelled as 'ring currents', which induce local magnetic fields that de-shield nearby nuclei in the plane of the aromatic ring and enhance the shielding of

nearby nuclei outside of the plane of the ring. The ring current contribution, δ_{ring}, is defined as

$$\delta_{ring} = iBG(r) \tag{1.12}$$

where *i* is a scaling factor related to the size of the aromatic ring, *B* is a measure of the magnetic susceptibility of the atom and *G(r)* is a function that describes the relative spatial locations of the atom and ring (Haigh and Mallion, 1979).

Empirically measured values of i and B have been recorded in different chemical environments (Osapay and Case, 1991, Case, 1995), providing reference data for the parameterisation of models (Haigh and Mallion, 1979, Johnson and Bovey, 1958, Pople, 1958). While the three models give results with broadly-similar accuracy, the Haigh and Mallion system has previously been found to be more accurate (Moyna et al., 1998), although recent efforts to 'definitively' determine the most reliable model suggest that Pople's simpler point-dipole model performs just as well as the more complex alternatives (Christensen et al., 2011).

The classical point-dipole model considers the contribution to arise as a function of the distance between an atom and the ring plane, and the angle between the ring plane centre and the centre of the atom:

$$G(r) = \frac{1-3\cos^2(\theta)}{r^3} \tag{1.13}$$

Where *r* is the distance between the ring plane centre and the centre of the atom of interest, and θ is the angle between a line connecting the two.

The semi-classical approach of Johnson and Bovey considers two ring loops – one above the ring plane and one below – as well as the radius of the ring, α, radial and vertical cylindrical co-ordinates, ρ and z, a ring current I and elliptic integrals K and E.

$$G(r) = \frac{I}{\alpha} \cdot \frac{2}{\left[(1+\rho)^2 + z^2\right]^{\frac{1}{2}}} \cdot \left[K + \frac{1-\rho^2-z^2}{(1-\rho)^2 + z^2} \cdot E \right] \tag{1.14}$$

with the modulus, *k*, of the complete elliptic integrals given by

$$k = \frac{4\rho}{(1+\rho)^2 + z^2} \tag{1.15}$$

The Haigh and Mallion model considers the sum of areas formed by connecting each triplet of adjacent ring atoms *i* and *j* with the point at which the atom of interest orthogonally projects onto the plane, r_0

$$G(r) = \sum_{ij}\left[S_{ij}\left(\frac{1}{r_i^3} + \frac{1}{r_j^3} \right)\right] \tag{1.16}$$

The local magnetic shielding of a nucleus can also be weakened or strengthened by the polarisation of proximal bonds by polar and charged moieties.

1.3.3 Transfer of magnetisation

Since the NMR signal is proportional to the current in the receiving coil, and the current is proportional to the rate of change of the magnetic moment, and magnetic moment is proportional to γ (the gyromagnetic ratio), and the Larmor frequency is also proportional to γ, it therefore follows that the signal intensity of a nucleus in an NMR experiment is proportional to the Larmor frequency. As a consequence, the signal-to-noise ratio arising from a ^{1}H nucleus is approximately 300 times larger than that of a ^{15}N nucleus. By transferring magnetisation from a neighbouring ^{1}H nucleus to a ^{15}N nucleus, the signal of the latter can greatly amplified.

The most common way to transfer magnetisation is a process called Insensitive Nuclei Enhanced by Polarization Transfer (INEPT). By appropriate pulse timing for a specific J-coupling frequency, it is possible to transfer magnetisation from a high-γ species to a low-γ species.

1.3.4 Acquisition of NMR spectra

NMR experiments typically consist of four phases:

1. *Initialisation*. The apparatus is prepared for the experiment.
2. *Excitation*. The desired rf signal is applied to the sample. This can consist of a complicated *pulse sequence* designed to cause a very specific transverse magnetisation profile within the sample.

3. *Detection*. A brief period after the pulse sequence has finished, the precessing transverse magnetisation induces a current in the rf receiver, which is recorded as a digital waveform in the computer.
4. *Processing*. The digital representation of the induced current waveform is processed by various mathematical operations as desired by the operator. Most significant of these is the Fourier transform, which identifies the many component frequencies present in the waveform and the amplitude of each.

As previously stated, the NMR signal is extremely weak. The instrumentation are very sensitive, such that they detect the background rf *noise* from unwanted sources such as thermal motions of electrons in the receiver and of charged particles in the sample. Since – especially in biological NMR samples – the desired *signal* is indiscernible from the noise (ie there is signal/noise ≤ 1), methods must be used to detect the information-carrying NMR signal. After careful design of apparatus, the most common technique is to record multiple scans for the same experiment. This relies on the fact that the NMR signal should be reproducible each time, but the noise is uncorrelated with the experiment and so randomly distributed on each scan. By summing each experiment, the NMR signal, s_{NMR}, will increase linearly with the number of scans, while the noise signal, s_{noise}, increases by a factor of $\sqrt{\mathfrak{n}}$ where $\mathfrak{n}$, the blackletter 'n', is the the number of scans.

As stated in section 1.3.1, J-coupling – caused by heteronuclear polarization through shared electrons – can be largely eliminated by an appropriate rf pulse at the Larmor frequency of those nuclear species not currently being detected. By extending this technique, it is possible to exploit the indirect spin-spin coupling by selectively transferring magnetization between specific nuclear spin-spin systems so that an rf signal is emitted only for nuclei in specific bonding environments. This process is called *heteronuclear polarization transfer* and is routinely used to identify the spin-spin interaction network – and, more broadly, the proximity of nuclei – within a given chemical environment.

A technique used routinely in biomolecular NMR is *multi-dimensional spectroscopy*. At the core of this, is the collection of a series of FIDs in which one or more parameters – such as the time between two rf pulses – is varied by small

increments and the resulting data are plotted as a multi-dimensional grid, as demonstrated in Figure 1.9.

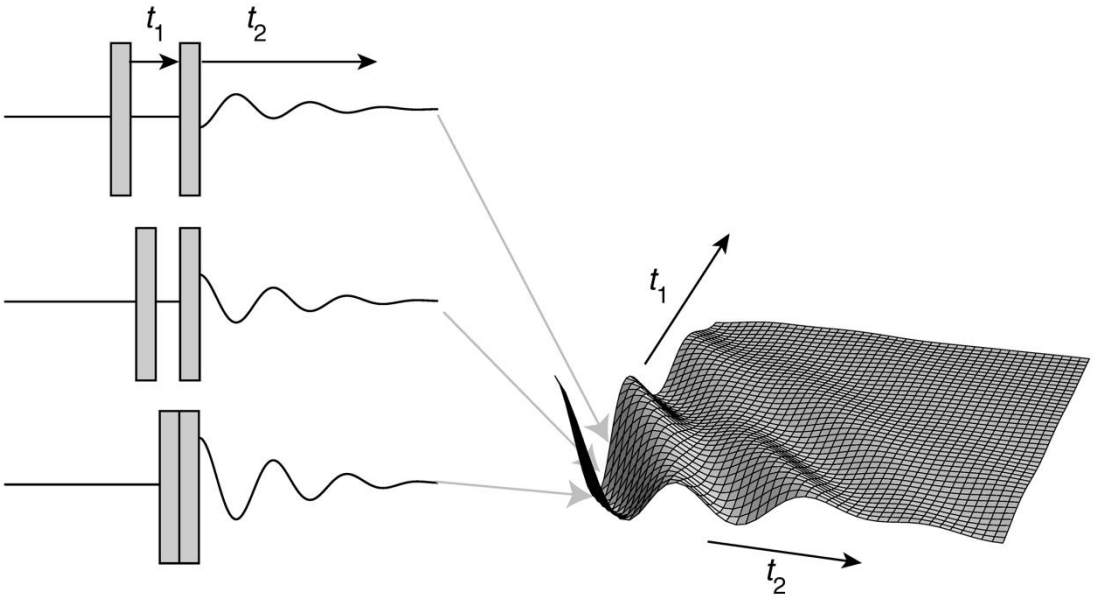

Figure 1.9 Diagram depicting how the FIDs from multiple pulse sequences are combined to produce a two-dimensional data matrix. Here, t_1 describes the time between two rf pulses and t_2 describes the time coordinate of the recorded FID. Adapted from (Levitt, 2008).

This technique, called *arrayed signal acquisition*, may be extended in to further dimensions by varying additional parameters and conducting the pulse sequence and collecting the FID for each combination of values for each of the varied parameters.

The two-dimensional FID, as seen in Figure 1.9, typically undergoes a processing that includes Fourier transformation. Just as a one-dimensional FID is processed to plot the signal intensity at each relative Larmor frequency, Ω, a two-dimensional spectrum is processed to give a plot of intensity at each contributing Fourier-transformed signal, $S(\Omega_1, \Omega_2)$, that result from their corresponding signal contributions, $s(t_1, t_2)$. This can be plotted either as a surface plot or a contour plot.

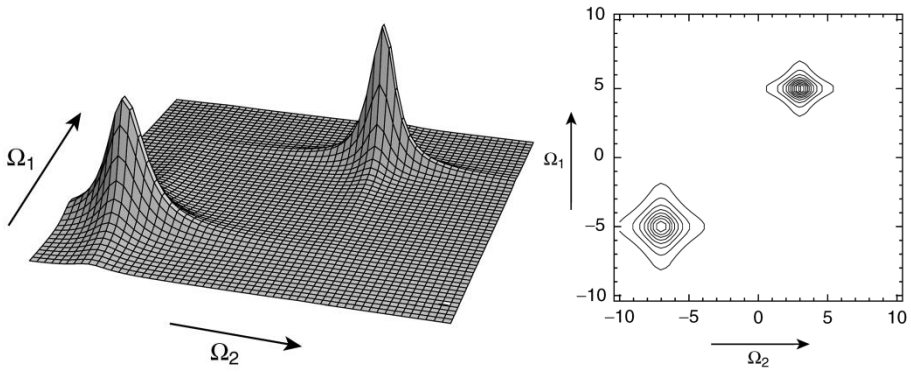

Figure 1.10 Illustrative representations of a two-dimensional spectrum as both a surface plot (left) and a contour plot (right). Adapted from (Levitt, 2008).

While there are many experiments for detecting different chemical relationships between nuclei, the classes of experiment most frequently used in the biomolecular NMR are heteronuclear single quantum coherence (HSQC), heteronuclear multiple quantum coherence (HMQC) and heteronuclear multiple-bond coherence (HMBC).

In HSQC, high-γ nuclei, S (the ^{1}H nucleus), are magnetised. The magnetisation is transferred (by INEPT, see section 1.3.3) to a chosen nearby low-γ species, I (either ^{15}N or ^{13}C), before being transferred back to the original proton. A spin-echo pulse sequence is used to decouple the signal so that multiplets are collapsed, giving rise to a single peak per unique proton chemical environment. This still includes protons uncoupled to an I nucleus, so the experiment is re-run with the phase of the second INPET transfer inverted, so that only uncoupled protons produce a signal. By subtracting the spectrum of the latter from the former, a spectrum showing the chemical environments of only I-coupled protons is produced. Consequently, HSQC spectra can be used to selectively reveal the chemical environments of ^{1}H-^{15}N and ^{1}H-^{13}C systems within a protein.

1.4 Molecular dynamics simulations

Molecular dynamics (MD) simulations are the computational application of models of atomic behaviour to predict the motions and interactions of molecules from a given starting environment. The desired result may be an equilibrium conformation or the properties of a dynamic system (such as the range of conformations sampled and their relative preponderances), but it can only ever

be as accurate as the model of motion on which it is based. Consequently, the results of MD simulations must be validated by the generation and testing of predictions about the system under investigation.

MD simulations are a subset of molecular modelling, in which chemical systems are represented by models suitable for tackling a particular problem. In MD simulations, the forces acting on each part of a molecular model are calculated according to a given interaction model. These forces are used to calculate the trajectory (including direction and acceleration) of each particle. The displacement of each particle is then calculated after a discrete time interval in a process called integration. This process is repeated many thousands of times to produce a system trajectory on the picosecond to microsecond timescale.

For any given starting molecular structure, there are many different MD simulations than can be performed, with many interaction models, integration strategies and other system properties to consider.

1.4.1 Fundamental principals

Molecular dynamics simulations can be summarised by two simple equations (Bungartz et al., 2014). Firstly, Newton's second law of motion:

$$\vec{F}_i = m_i \vec{a}_i \tag{1.17}$$

This describes how the vector sum of the forces, F, on a particle i is equal to that particle's mass, m_i, multiplied by the vector acceleration, a_i. Given a situation where the mass of and net force acting on a particle are known, the acceleration can therefore be calculated.

Secondly, the force acting on a particle, F_i, due to a vector potential, V(r) = ($r_{1,x}$, $r_{1,y}$, $r_{1,z}$ … $r_{N,x}$, $r_{N,y}$, $r_{N,z}$) between particle i at co-ordinates $r_{x,y,z}$ and each of the N particles in a system can be calculated as:

$$F_i = -\nabla_i V(r) \tag{1.18}$$

This is generated by calculating the force between pairs of non-bonded atoms *i* and *j*:

$$F_i = \sum_j F_{ij} \tag{1.19}$$

MD simulations proceed by repeatedly solving each of these two equations for each particle in a system over a series of very small (typically 1-10fs) time steps, Δt. The nature of the particles – atoms, amino acids, moieties, for example – is not binding, so long as the properties relevant to the potentials and the masses of the particles are known.

Due to its reliance on Newtonian physics, MD applies only to systems that can be described by classical mechanics. Systems in which quantum properties are significant – such as the behaviour of single protons (hydrogen nuclei) or low-temperature liquid helium – cannot be accurately modelled by MD simulations. As a consequence, quantum properties are incorporated in MD simulations either as harmonic oscillators or as constraints on the motion of particles. These are implemented as modifications of *V*, with different models considering different properties of the system.

The model used for calculating *V*, and consequently F, is encapsulated in a force field. Many force fields have been developed to tackle different types of molecular system. None of these models is inherently better than all others in all use cases, and it is often necessary to repeat simulations using different force fields to assist with assessing whether a result is real or an artefact of the chosen force field.

1.4.2 Integrators

The act of calculating the displacement of particles within a MD system is performed by an integrator – named after the act of integrating the equations of motion for discrete time steps.

Each time step follows this process:

1. Take as input the co-ordinates and velocities of all particles in the system at time *t*.

2. Calculate the potentials and subsequently the forces and accelerations acting on all particles at *t*.
3. Using this information, calculate the co-ordinates and velocities of all particles at time $t + \delta t$.

Rearrangement of equation (1.17) yields

$$\vec{a} = \frac{\vec{F}}{m} \tag{1.20}$$

Acceleration of a particle, being the rate in change in displacement, can be calculated as the second temporal derivative of the particle's co-ordinates:

$$a = \ddot{r} \tag{1.21}$$

A naïve approach to calculating the new co-ordinates and velocity would therefore be

$$\mathrm{r}(t + \delta t) = \mathrm{r}(t) + \delta t.\mathrm{v}(t) \tag{1.22}$$

$$\mathrm{v}(t + \delta t) = \mathrm{v}(t) + \delta t.\mathrm{a}(t) \tag{1.23}$$

This technique, called Euler's Method, suffers many problems that make it unsuitable for MD simulations, including numerical instability and truncation.

As with all MD processes, there are a many algorithms that have been developed to accomplish this task, including Störmer–Verlet integration (Verlet, 1967), Beeman's algorithm (Schofield, 1973, Beeman, 1976), Runge–Kutta methods (Press et al., 2007), Symplectic integration (Ruth, 1983). They all attempt to tackle the problems associated with time discretisation in slightly different ways, often choosing a different compromise between accuracy and computational efficiency as well as providing the ability to ignore or consider artifices of MD, such as the effects of volume, pressure and temperature of a system.

1.4.2.1 Velocity-Störmer-Verlet integration

One frequently-used algorithm is a variant of Störmer-Verlet called the Velocity-Störmer-Verlet integrator (Swope et al., 1982). By considering the sub-*t* acceleration of a particle, we derive

$$\mathrm{r}(t+\delta t) = \mathrm{r}(t) + \delta t.\,\mathrm{v}(t) + \frac{\delta t^2}{2}\mathrm{a}(t) \tag{1.24}$$

While the starting co-ordinates and velocity are provided, the acceleration (both at t and $t+\delta t$) are calculated from the energy potential for the co-ordinates.

To calculate the velocity, the velocity at time $(t+\frac{\delta t}{2})$ is first calculated by substitution in equation (1.23):

$$\mathrm{v}(t+\frac{\delta t}{2}) = \mathrm{v}(t) + \frac{\delta t}{2}\mathrm{a}(t) \tag{1.25}$$

This is then used to calculate the velocity at time $(t+\delta t)$ using the acceleration at time $(t+\delta t)$ and the velocity at time $(t+\frac{\delta t}{2})$:

$$\mathrm{v}(t+\delta t) = \mathrm{v}(t+\frac{\delta t}{2}) + \frac{\delta t}{2}\mathrm{a}(t+\delta t) \tag{1.26}$$

Combining equations (1.25) and (1.26) we can calculate the velocity at time $(t+\delta t)$ as:

$$\mathrm{v}(t+\delta t) = \mathrm{v}(t) + \frac{\delta t}{2}\left[\mathrm{a}(t) + \mathrm{a}(t+\delta t)\right] \tag{1.27}$$

Equations (1.24) and (1.27) are the calculations followed by the Velocity-Störmer-Verlet integrator. While this can provide accurate results, one large draw-back is the necessity to calculate the acceleration for each particle twice for each time step. One commonly-used integrator that reduces this burden is the Leapfrog integrator (Hockney et al., 1974, Hut et al., 1995, Pronk et al., 2013).

1.4.2.2 Leapfrog integration

The leapfrog integrator is very similar to the Velocity-Störmer-Verlet integrator, except that it considers the velocities only at $(t+\frac{\delta t}{2})$. The velocities are calculated as in equation (1.25), and co-ordinates are calculated as

$$\mathrm{r}(t+\delta t) = \mathrm{r}(t) + \mathrm{v}(t+\frac{\delta t}{2})\delta t \tag{1.28}$$

This integrator considerably reduces the computational complexity compared to the Velocity-Störmer-Verlet integrator, while retaining similar levels of accuracy for many applications with constant error with respect to time.

1.4.3 Energy minimisation

The initial starting co-ordinates for a MD system are typically unrealistic, which leads to unrealistically-high forces acting on the particles and consequently the system partially or fully pulls itself apart; it explodes. To mitigate this problem, the initial structures undergo a process of energy minimisation, in which this excess free energy is removed by artificially re-arranging the structure into a more favourable conformation while retaining the initial inter-particle connectivity. The purpose is not to find the single most energetically favourable configuration for the structure (the global minimum) – such an endeavour would be computationally infeasible – but instead to find the nearest local energy minimum – the first minimum encountered when moving down the steepest neighbouring gradient of the energy landscape.

Since the initial conformation is often within a local minimum with respect to the feasible structures, it is typically sensible to try several rounds of increasing the energy followed by minimisation in a process called simulated annealing.

1.4.4 Force fields

Calculating the potential between particles in an MD system is accomplished by the force field. This consists of a set of equations (*potential functions*) that define how the inter-particle potential (and the derived force) should be calculated; and a corresponding set of parameters – often derived empirically – that define

scaling and weighting of these equations for different particles (Pronk et al., 2013).

While it may appear desirable to simulate the effect of each of the four fundamental forces of nature (gravitational, electromagnetic, strong nuclear and weak nuclear), doing so would be computationally infeasible for systems with more than a very small number of subatomic particles. Instead, the effects of these forces are modelled as atom-, moiety- or molecule-scale interactions.

There are many different force fields in common use each with many different versions tailored to modelling specific properties of specific molecule types. One such model that has been popular for many decades due to its computational simplicity is the Lennard-Jones (LJ) potential (Frenkel and Smit, 2002). This models the effects of the van der Waal's (vdW) interaction and exchange interaction.

1.4.4.1 Lennard-Jones potential

The *vdW interaction* is commonly called the vdW force, though this is disfavoured in scientific literature to avoid confusion with the fundamental forces. The vdW interaction arises from the attraction between proximal atoms in which one or more temporary dipoles are induced by the atoms' proximity. Since an increasing proximity causes an increased charge polarisation, the attraction increases exponentially with decreasing distance. The vdW potential, V_{vdW}, between two atoms, i and j, with separation r_{ij} is calculated as

$$V_{vdW}\left(r_{ij}\right) = -4\varepsilon_{ij}\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \tag{1.29}$$

where ε_{ij} and σ_{ij} are empirically derived parameters that respectively describe the depth of the potential well and distance at which the interatomic potential is zero for the particles. Note that the negative sign indicates an attractive potential.

The *exchange interaction* includes the more famous *Pauli repulsion*, but is applicable to more particles. The exchange interaction refers to the quantum mechanical interaction between two atoms that become so close that their probability clouds (their *wave functions*) begin to overlap. For fermions (a class of particle including electrons and many nuclei), this interaction is strongly

repulsive at short distances, and increases exponentially with decreasing distance. The exchange potential, V_e, is calculated as

$$V_e\left(r_{ij}\right)=4\varepsilon_{ij}\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} \tag{1.30}$$

Note that the potential is positive, and the non-constant term is simply the square of that for the vdW potential, making calculation of these two potentials computationally efficient.

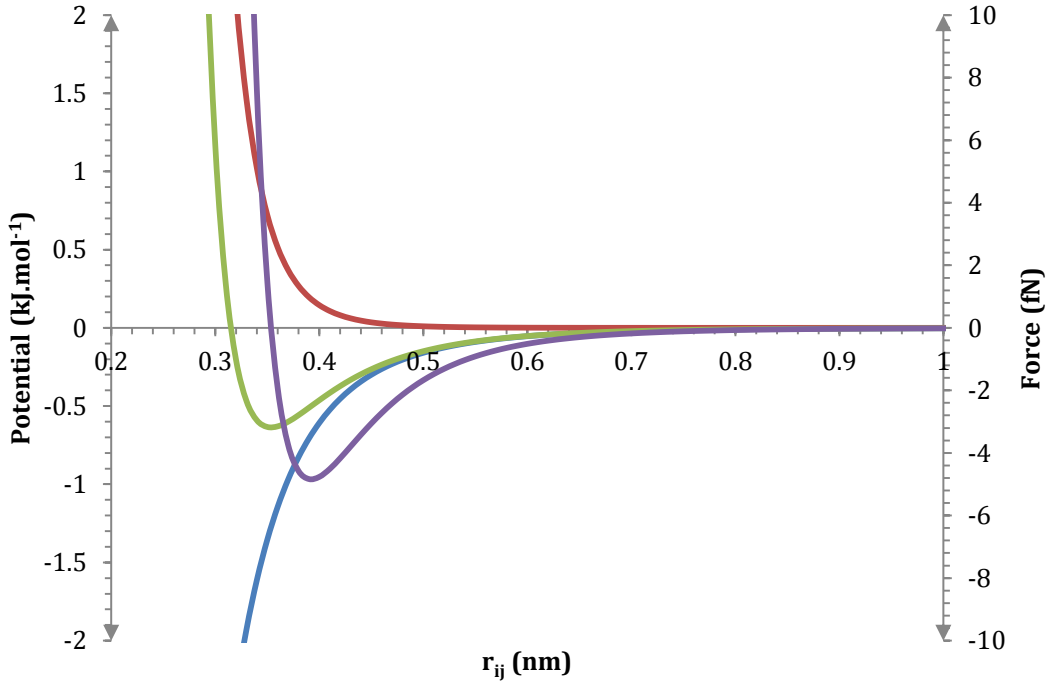

Figure 1.11 Graph showing the relationship between vdW potential (blue), exchange potential (red), LJ potential (green) and LJ force (purple). A negative force indicates attraction, while a positive force indicates repulsion. Equilibrium LJ force is achieved at minimum LJ potential. The LJ potential crosses the x axis at $x=\sigma$ and has minimum value $y=-\varepsilon$. The values used for σ and ε are those of the TIP3P water model.

The Lennard-Jones potential combines these two interactions as

$$V_{LJ}\left(r_{ij}\right)=4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}-\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{1.31}$$

The force arising from the LJ potential, F_{LJ}, is the negative derivative of the potential

$$F_{LJ}\left(r_{ij}\right)=24\varepsilon\sigma_{ij}^{6}\left[\left(\frac{1}{r_{ij}}\right)^{7}-2\sigma_{ij}^{6}\left(\frac{1}{r_{ij}}\right)^{13}\right] \tag{1.32}$$

As seen in Figure 1.11, the LJ potential induces attraction at longer distances and strong repulsion at shorter distances. The distance at which no force is exerted, r_{eq}, is

$$r_{\min,ij}=-\sigma_{ij}\sqrt[6]{2} \tag{1.33}$$

Thus, the LJ potential can be expressed as

$$V_{LJ}\left(r_{ij}\right)=\varepsilon_{ij}\left[\left(\frac{r_{\min,ij}}{r_{ij}}\right)^{12}-2\left(\frac{r_{\min,ij}}{r_{ij}}\right)^{6}\right] \tag{1.34}$$

While computationally efficient and useful for simple systems, the LJ potential fails to account for many of the interactions within protein-based systems, making it unsuitable for protein MD.

1.4.4.2 Protein MD force fields

Commonly used families force fields for modelling proteins include Amber CHARMM, GROMOS and OPLS/AA. These force fields calculate the total energy of the system, E_{total} as the sum of the energy arising from bonded interactions, E_{bonded}, non-bonded interactions, $E_{non\text{-}bonded}$ and other force field-specific terms, E_{other}:

$$E_{total}=E_{bonded}+E_{non\text{–}bonded}+E_{other} \tag{1.35}$$

(Guvench and MacKerell, 2008)

Bonded interactions have energies arising from bond stretching, E_{bonds}, bond angles, E_{angles}, and dihedral angles, $E_{dihredrals}$:

$$E_{bonded}=E_{bonds}+E_{angles}+E_{dihedrals} \tag{1.36}$$

While there are many approaches to calculating these terms, typical forms are as follows. The energy arising from an inter-particle bond of length b is governed by a stiffness parameter, K_b, and the equilibrium length, b_0, such that

$$E_{bonds} = \sum_{bonds} K_b (b - b_0) \tag{1.37}$$

Similarly, the energy arising from a bond angle, θ, is governed by a stiffness parameter, K_θ, and equilibrium angle, θ_0:

$$E_{angles} = \sum_{angles} K_\theta (\theta - \theta_0) \tag{1.38}$$

Dihedral angle energies are parameterised by weighting value, K_χ, dihedral angle, χ, a periodicity value (number of peaks/troughs in energy per full rotation), n, and phase value, σ:

$$E_{dihedrals} = \sum_{dihedrals} K_\chi \left[1 + \cos(n\chi - \sigma)\right] \tag{1.39}$$

Non-bonded interactions are accounted for by a combination of the Lennard Jones potential and Coulomb potential. The Coulomb potential, V_c, between two particles, i and j, with charges, q_i and q_j, separated by distance r_{ij} and with relative dielectric constant ε_r is given by

$$V_c(r_{ij}) = \frac{q_i q_j}{\varepsilon_r r_{ij}} \tag{1.40}$$

Thus, the non-bonded potential can be derived by combining equations (1.35) and (1.40):

$$E_{non\text{–}bonded} = \sum_{\text{non–bonded pairs } ij} \left(\varepsilon_{ij} \left[\left(\frac{r_{\min,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{\min,ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\varepsilon_r r_{ij}} \right) \tag{1.41}$$

1.4.5 Periodic Boundary Conditions

The space being simulated within an MD system is finite. Attention must therefore be paid to how interactions with the boundary of the system space are treated. The traditional way to handle such interactions is to treat the system as a box that is repeated in each direction; particles at one boundary thus interact

with other particles within the same system at another boundary. This is called the periodic boundary condition (PBC). While the effects of artificial boundaries are ameliorated, this approach instead applies periodic condition artefacts. By applying different periodic conditions and system sizes, it is possible to compare the effects of those parameters and disentangle such effects from the properties under investigation. However, such effects are considered to be less consequential than the effect of placing the system within a vacuum.

Many possible box shapes are possible. In general, a shape that minimises the amount of solvent being simulated is preferred, since the dynamics of solvent particles typically are not of interest.

As a result of the PBC, the number of interactions that could be considered increases enormously; not only could particles now interact with those in a neighbouring image, but also with those in the image next to that, and so on. The *minimum image convention* stipulates that interactions are only considered with adjacent images. This is typically imposed by considering only the interactions between particles within a finite cutoff distance, r_c, that is equal to or less than half the simulation box length.

Restrictions on the interactions to be considered require the choice of both a suitable cutoff distance and a means by which to identify the neighbouring particles. Fortunately, the neighbouring particles changes infrequently with respect to Δt, so the neighbour searching need not be performed at each time step, though the choice of algorithm can still have a large effect on the simulation rate. A commonly-used algorithm is the Verlet Neighbour List. This defines two spherical regions about each particle: the sphere within r_c and a sphere within r_m, where $r_\mathrm{c} > r_\mathrm{m}$. The distance between each pair of particles, i and j is first calculated and lists of atoms within each sphere (where $r_\mathrm{ij} \leqslant r_\mathrm{c}$ or $r_\mathrm{ij} \leqslant r_\mathrm{m}$) is generated. The list of atoms within the outer sphere requires updating every $\mathrm{N_m}$ steps, where

$$r_m - r_c > N_m \langle v_{\max} \rangle \Delta t \tag{1.42}$$

where $\langle v_{\max} \rangle$ is the highest mean velocity. The inner list may be updated regularly, but need only consider the particles within r_m. This essentially means

that the global neighbour list need only be updated when it is at all possible that a particle may have entered into the r_c.

This, however, often fails to accurately incorporate long-range electrostatic effects.

1.4.6 Electrostatic cut-offs

Long-range interactions can be modelled by one of several algorithms, including the Ewald Summation, Particle Mesh Ewald (PME) and Particle-Particle-Particle-Mesh (P^3M).

Since the charge potential is inversely proportional to the distance between two charged particles i and j with charges q_i and q_j, the electrostatic potential, V, arising from a simulation box image removed from the origin by n images can be calculated by considering the box length, L:

$$V = \frac{1}{2}\sum_{1}^{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j}{4\pi\varepsilon_0\left|r_{ij}+\left(n_x L, n_y L, n_z L\right)\right|} \tag{1.43}$$

This equation is conditionally convergent (the sum of the absolute value of the operand is not finite), but computationally slow to calculate.

We therefore have a set of computational tools for simulating the properties and dynamics of biological systems using empirically-derived parameters to describe average properties of common particles. The ability to accurately incorporate NMR data in MD would offer significantly more refined MD simulations.

1.5 Use of chemical shifts in protein structure determination

Based on efforts to rationalise the relationship between NMR chemical shifts and protein structure, many tools have been developed to predict chemical shifts for a given structure with varying accuracy and computational efficiency. These tools, including CamShift (Kohlhoff et al., 2009), *Che*Shift (Vila et al., 2009), PROSHIFT (Meiler, 2003), SHIFTS (Xu and Case, 2001), SHIFTX (Neal et al., 2003), SHIFTX2 (Han et al., 2011), and SPARTA+ (Shen and Bax, 2010), have subsequently enabled the development of tools to predict protein structure from chemical shifts. Following the success of CASP to provide objective assessment of tools for predicting protein structure, the Critical Assesment of Structure

Determination by NMR (CASD-NMR) has been established and the two primary tools for solving structure from chemical shifts directly, CHESHIRE (Cavalli et al., 2007) and CS-Rosetta (Shen et al., 2009), have been tested for their ability to blindly determine structures, with CHESHIRE providing greatest accuracy (Rosato et al., 2012). Another similar tool, CS23D (Wishart et al., 2008), was not assessed and so cannot easily be compared for accuracy.

1.5.1 Structure determination by molecular fragment replacement

Attempting to re-fold the structure from an entirely disordered state would likely be error-prone and time consuming. The CHESHIRE (chemical shift restraints) pipeline, however, allows *ab initio* protein structure determination by use of molecular fragment replacement in conjunction with 3PRED secondary structure prediction from chemical shifts and TOPOS database searching for fragments with similar sequence, secondary structure and chemical shifts.

The CHESHIRE pipeline involves production of a large number of low-resolution structures, selection and refinement of the best-scoring structures, characterisation of the refined ensemble and, finally, selection of a very small number of the best structures for use either as a starting point for CS-MD, or for comparison to the ensembles produced by CS-MD. So far, approximately 4000 low-resolution structures have been produced for each of the folded state datasets and refinement is currently taking place. Refinement and characterisation is expected to take approximately 2 weeks to complete.

1.5.2 Structure determination by restrained molecular simulations

In addition to the possibility of calculating structures from chemical shifts using Monte Carlo refinement methods such as CS-Rosetta and CHESHIRE, the same NMR data can be used to guide molecular dynamics simulations, providing even greater information about the behaviour of the protein in solution (Camilloni et al., 2012b, Robustelli et al., 2010). To this end, the CamShift chemical shift predictor was developed as a fast alternative for which the chemical shifts are readily differentiable with respect to the atomic co-ordinates, making feasible its use as a term within the molecular dynamics energy function (Kohlhoff et al., 2009). Such chemical shift restrained molecular dynamics (CS-MD) (Robustelli et

al., 2010) can allow structure determination, but has also been used to understand the motions of dynamic proteins, such as the multiple native states of RNase A (Kelkar et al., 2012). CS-MD therefore has the potential to reveal the free energy landscapes of stalled nascent chains, which typically have a paucity of high-resolution data available, as explained in previous reports.

1.6 Model proteins of interest for co-translational properties

In order to study the ability to obtain structural information about co-translational properties of nascent chains, two model protein systems were considered in the study presented here: Gelation Factor domain 5 and α-synuclein.

1.6.1 Gelation Factor domain 5

Gelation factor – also known as Actin Binding Protein 120 (ABP-120), Ig2 or ddFLN – is an 857-residue F-actin cross-linking protein of the slime mold *Dictyostelium discoideum*. It consists of an amino-terminal actin-binding domain coupled to 6 immunoglobulin-like filamin domains (Fucini et al., 1997, Noegel et al., 1989). Gelation factor is of interest as the large number of stable independently folding domains provide much opportunity for cotranslational folding to occur and the study thereof.

The presence of multiple folded domains indeed presents the possibility for studying the folding pathways of each, which may be affected by the translation rate of downstream sequence. Previous Atomic Force Microscopy (AFM) studies have shown that filamin domain 4 folds via an obligate slow-folding intermediate before folding the remaining residues at 3-times the rate while the other domains form more stable folded structures with no obvious intermediate (Schwaiger et al., 2004, Schwaiger et al., 2005). Being able to rationalise such folding pathways with events that occur on the ribosome during translation may reveal important aspects about *in vivo* co-translational protein folding.

RNCs of Gelation factor-derived constructs have been studied by NMR and domains shown to fold while paused on the ribosome, as well as putative sites for interaction with the ribosome suggested (Hsu et al., 2007). Assignments for the native and denatured ^{1}H, ^{15}N and ^{13}C NMR spectra of domain 5 are available

(Hsu et al., 2009a). As more data become available, the desire to identify the structure and dynamics of Gelation factor domains during translation becomes more opportune.

1.6.2 αSynuclein

The *in vivo* misfolding of αSynuclein (αSyn) results in the formation of aggregates of amyloid fibrillar deposits (Lewy bodies), is strongly featured in the aetiology of PD and has been implicated in 60% of familial and sporadic AD cases (Yokota et al., 2002). PD and AD are the most common forms of neurodegenerative disease and are both chronic and progressive. Pathophysiology of PD is marked by loss of dopaminergic neurons resulting in disorganisation of the basal ganglia connectivity and subsequent hypokinesia, bradykinesia and other disturbances to motor organisation (Bartels and Leenders, 2009).

αSyn is a 140 residue, intrinsically disordered protein in solution. There is latent helicity (approximately 10%) in the N-terminal 100 amino acids (observed by chemical shift measurements) and indeed it has been shown that the amino-terminal forms two α-helices in lipid membranes (Eliezer et al., 2001). αSyn is the only member of the three-member family in humans to be implicated in disease (Spillantini et al., 1998). As seen in Figure 1.12, αSynuclein can be broadly divided into three domains consisting of the amino-terminal, carboxy-terminal and a central fibril-associated core containing a hydrophobic region thought to be the site at which aggregation is mostly likely initiated: the non-Aβ component (NAC). Mutants of the αSyn gene found in certain demographics have been found to cause early-onset of disease: A53T (Polymeropoulos et al., 1997), E46K (Zarranz et al., 2004) and A30P (Kruger et al., 1998). That these mutants are outside the NAC suggests that the regions in which they are located play a role in stabilising the αSyn structure to prevent aggregation.

Physiologically, αSyn is abundant in nerve terminals and is often localised with synaptic vesicles (Clayton and George, 1999). Experiments based on a mouse model have suggested a role for αSyn in conjunction with cystein-string protein-α (CSPα) and SNARE proteins in the protection of the presynaptic membrane (Chandra et al., 2005). αSyn has been shown to selectively inhibit phospholipase

D2 – involved in endocytosis – suggesting it may be involved with the regulation of membrane transport (Goedert, 2001).

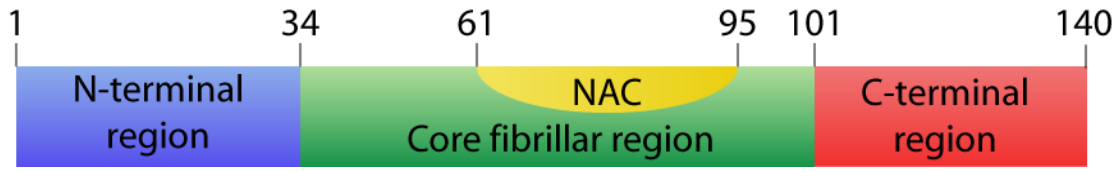

Figure 1.12. Schematic representation of the principal αSyn regions along the 140 amino acid residues, as described previously (Der-Sarkissian et al., 2003). The residue numbers at the region boundaries are given.

The mechanism of aggregation of αSyn is not fully established, but a proposed disease pathway is thought to result from damaged αSyn being improperly processed by the proteasome and lysosome, allowing oligomers to form (Uversky, 2007). The oligomers form ordered aggregates (Lewy bodies) which can interfere with cellular machinery in several ways, including blocking endoplasmic reticulum-Golgi traffic (Cooper et al., 2006) and interference with the proteasome, inhibiting the ability of the cell to withstand stationary phase ageing (Chen et al., 2005).

The removal of the αSyn acidic C-terminal 40 residue tail results in a dramatic increase in the rate of aggregation. Despite being intrinsically disordered, αSyn does possess an overall topology favouring the acidic C-terminal tail being in contact with the NAC region (Dedmon et al., 2004). The study of fragments of αSyn has also been of considerable interest as a means of comparison with homologues not normally observed to aggregate *in vivo*. Recent work has shown that the NAC region alone is not the key determinant for the aggregation of αSyn; re-addition of the hydrophobic core of the NAC region to β-syn (a close homologue) did not result in a significant aggregation rate enhancement (Rivers et al., 2008).

A long-term ambition of a large number of laboratories worldwide is to develop methods and strategies that prevent the misfolding and subsequent aggregation of αSyn before the onset of pathological consequences. However, little is yet known about how the intrinsically disordered protein is protected during its vectorial synthesis. As with all disordered proteins, there is an absence of

understanding of the structural properties that protect αSyn during its synthesis, how immediate degradation by the proteasome is prevented, and how the native state is maintained.

2 Materials and Methods

2.1 Structure completion

Conversion of coarse-grained structures to all-atom models (known as structure completion) is inherently imperfect, as there are both multiple feasible conformations that fit to a given set of coarse-grained co-ordinates, and selection of the most favoured conformation is not necessarily obvious. While molecular dynamics suites, including GROMACS, include the capacity to place some missing atoms, a more accurate approach lies with use of SCWRL (Krivov et al., 2009). This uses a library of favoured backbone-dependent rotamer conformations and a simple steric-clash force field to predict likely sidechain conformations. Backbone atoms can therefore be placed using favoured φ/ψ angles as used in Ramachandran Plot analysis (Lovell et al., 2003), and the sidechain atoms placed by SCWRL4. The MMTSB Tool Set provides such functionality with the `complete.pl` script (Feig et al., 2009). Following placement of missing atoms, `complete.pl` refines the structure by finding the nearest local minima with the CHARMM22 force field.

2.2 Hydrodynamic radius calculation

Hydrodynamic radii, R_H, values were calculated with the Hydropro software (Ortega et al., 2011). Where large ensembles of structures required R_H calculation, batch jobs were submitted to the UCL Legion computer cluster wherein each run of the job would extract the co-ordinates, complete the structure (if necessary) following the protocol above, and perform the Hydropro calculation.

A sample Hydropro configuration file is shown in Figure 2.1.

```
2.9-asyn_nores                !Name of molecule
aSyn                          !Name for output file
this_structure.pdb            !Strucutural (PBD) file
1,                            !INDMODE
2.9,                          !AER, radius of the atomic elements
8,                            !NSIG
1.6,                          !Minimum radius of beads in the shell (SIGMIN)
2.6,                          !Maximum radius of beads in the shell (SIGMAX)
20,                           !T (temperature, C)
0.01,                         !ETA (Viscosity of the solvent in poises)
15854.4240,                   !RM (Molecular weigth)
0.702,                        !Partial specific volume, cm3/g
1.0,                          !Solvent density, g/cm3
21                            !Number of values of H
2.e+7,                        !HMAX
30,                           !Number of intervals for the distance distribution
-1.,                          !RMAX
1000,                         !Number of trials for MC calculation of covolume
1                             !IDIF=1 (yes) for full diffusion tensors
*                             !End of file
```

Figure 2.1 Sample configuration file for Hydropro.

2.3 Energy minimisation molecular dynamics simulations

All the molecular dynamics simulations conducted in the preparation of the presented work were conducted using the GROMACS molecular simulation toolkit versions 4.5 to 4.5.5 (Pronk et al., 2013). During energy minimization phases, integration time steps between 0.05fs and 2fs were used. The 'em' integrator of GROMACS was used with a target number of steps of 10,000, thereby allowing the system to continue until no further minimization could be achieved; at no time was this target number of steps reached before minimisation was finished. A typical configuration file for the GROMACS `grompp` tool can be seen in Figure 2.2.

```
constraints              = none
integrator               = steep
dt                       = 0.002
nsteps                   = 1000
nstlist                  = 10
ns_type                  = grid
rlist                    = 0.9
coulombtype              = PME
rcoulomb                 = 0.9
rvdw                     = 0.9
fourierspacing           = 0.12
fourier_nx               = 0
fourier_ny               = 0
fourier_nz               = 0
pme_order                = 6
ewald_rtol               = 1e-5
optimize_fft             = yes
emtol                    = 1000.0
emstep                   = 0.01
```

Figure 2.2 Typical configuration file for GROMACS energy minimisation runs.

2.4 CamShift molecular dynamics simulations

During the recording of trajectories for analysis (the production run), integration times of 2fs were used throughout. Bond relationships were constrained using LINCS (Hess et al., 1997). For the implicit solvent model, the Onufriev-Bashford-Case Born radii calculation method was employed (Onufriev et al., 2004) and the electrostatic field was incorporated with cut-offs. In all simulations van der Waals interactions were accounted for with a cut-off 1.0nm. During energy minimization and equilibration within the chemical shift restraints, simulated annealing was employed consisting of 100ps at 300K, 100ps constant increase to 450K, 100ps at 450K and 300ps constant decrease to 300K. During production runs, the system was modelled as an NVT ensemble – a canonical ensemble – using temperature coupling with velocity rescaling (a "Bussi thermostat") (Bussi et al., 2007). Final production trajectories were conducted using the Amber99SB*-ILDN forcefield (Best and Hummer, 2009).

A typical mdp configuration file used for compiling the simulation system with the GROMACS `grompp` tool is shown in Figure 2.3.

```
integrator               = sd
tinit                    = 0
dt                       = 0.002
nsteps                   = -1
init_step                = 0
simulation_part          = 1
comm-mode                = angular
nstcomm                  = 1
comm-grps                =
bd-fric                  = 0
ld-seed                  = 1993
emtol                    = 10
emstep                   = 0.01
niter                    = 20
fcstep                   = 0
nstcgsteep               = 1000
nbfgscorr                = 10
rtpi                     = 0.05
nstxout                  = 1
nstvout                  = 1
nstfout                  = 1
nstlog                   = 1
nstcalcenergy            = 1
nstenergy                = 1
nstxtcout                = 1
xtc-precision            = 1000
xtc-grps                 =
energygrps               =
nstlist                  = 0
ns_type                  = simple
pbc                      = no
periodic_molecules       = no
rlist                    = 0
rlistlong                = 0
coulombtype              = cut-off
rcoulomb-switch          = 0
rcoulomb                 = 0
epsilon_r                = 1
epsilon_rf               = 1
vdw-type                 = cut-off
rvdw-switch              = 0
rvdw                     = 0
DispCorr                 = No
table-extension          = 1
energygrp_table          =
fourierspacing           = 0.12
fourier_nx               = 0
fourier_ny               = 0
fourier_nz               = 0
pme_order                = 4
ewald_rtol               = 1e-05
ewald_geometry           = 3d
epsilon_surface          = 0
optimize_fft             = yes
implicit_solvent         = GBSA
gb_algorithm             = obc
nstgbradii               = 1
rgbradii                 = 0
gb_epsilon_solvent       = 80
gb_saltconc              = 0
gb_obc_alpha             = 1
gb_obc_beta              = 0.8
gb_obc_gamma             = 4.85
gb_dielectric_offset     = 0.009
sa_algorithm             = Ace-
approximation
sa_surface_tension       = 2.092
tcoupl                   = No
```

```
nsttcouple               = -1
nh-chain-length          = 10
tc-grps                  = system
tau_t                    = 0.1
ref_t                    = 300
Pcoupl                   = No
Pcoupltype               = isotropic
nstpcouple               = -1
tau_p                    = 0.5
compressibility          = 4.5e-5
ref_p                    = 1.0
refcoord_scaling         = No
andersen_seed            = 815131
gen_vel                  = yes
gen_temp                 = 300.0
gen_seed                 = -1
constraints              = all-bonds
constraint-algorithm     = Lincs
continuation             = no
Shake-SOR                = no
shake-tol                = 0.0001
lincs-order              = 4
lincs-iter               = 2
lincs-warnangle          = 30
morse                    = no
energygrp_excl           =
nwall                    = 0
wall_type                = 9-3
wall_r_linpot            = -1
wall_atomtype            =
wall_density             =
wall_ewald_zfac          = 3
pull                     = no
disre                    = No
disre-weighting          = Conservative
disre-mixed              = no
disre-fc                 = 1000
disre-tau                = 0
nstdisreout              = 100
orire                    = no
orire-fc                 = 0
orire-tau                = 0
orire-fitgrp             =
nstorireout              = 100
dihre                    = no
dihre-fc                 = 1000
free-energy              = no
init-lambda              = 0
delta-lambda             = 0
foreign_lambda           =
sc-alpha                 = 0
sc-power                 = 0
sc-sigma                 = 0.3
nstdhdl                  = 10
separate-dhdl-file       = yes
dhdl-derivatives         = yes
dh_hist_size             = 0
dh_hist_spacing          = 0.1
couple-moltype           =
couple-lambda0           = vdw-q
couple-lambda1           = vdw-q
couple-intramol          = no
acc-grps                 =
accelerate               =
freezegrps               =
freezedim                =
cos-acceleration         = 0
deform                   =
```

Figure 2.3 Typical GROMACS configuration for CSMD trajectory capture.

Chemical shifts were incorporated as a restraint in the simulations using the method described in (Camilloni et al., 2012a). Briefly, the PLUMED plugin (Bonomi et al., 2009) to GROMACS is used, in conjunction with a modification to use camshift version 1.35 (Kohlhoff et al., 2009). Time averaging of 100 integration steps was used for the calculation of restraints. An example PLUMED configuration file is shown in Figure 2.4.

```
PRINT W_STRIDE 100 APPEND
UMBRELLA CV 1 KAPPA 0 SLOPE 5.0 AT 0 ANNEALING 300
CAMSHIFT LIST <prot> NRES 107 DATA data/ FF
a03_gromacs.mdb
ALIGN_ATOMS LIST <prot>
prot->
LOOP 1 1554 1
prot<-
ENDMETA
```

Figure 2.4 Example PLUMED+CamShift configuration file.

Systems were prepared for CS-MD by first performing energy minimisation of the system (with two replicas), and then equilibrating the system for circa 120ns within the forcefield absent of chemical shift restraints. Restraints were introduced by a 'csramp' process: restraints were initially applied with an alpha (weighting) value of 0.25 and subjected to a 1ns simulated annealing as described above. The alpha value was increased by 0.25 and the process repeated. This process continued iteratively – using the two last structures (1 from each replicon) as the starting structures for a higher-weighted CSMD run – until an alpha value of 6.5 was stably reached. If, at any stage, the system exploded, the alpha value was dropped by 0.25 and the last starting structure used again. In this way, a pair of starting structures that could be stably restrained by chemical shifts could be obtained.

Production trajectories for CSMD systems were obtained by taking the output structures of the csramp process and running circa 120ns simulations without simulated annealing and with an alpha value of 6.5. The simulation parameters were therefore identical to the unrestrained MD simulations, except for the addition of the restraints.

2.5 Selective simulation of nascent chains within RNCs

A protocol for performing RNC MD and CSMD was developed as described in chapter 6. The final protocol involved performing simulations similar to the non-RNC systems, though with frozen ribosome co-ordinates; that is, the co-ordinates of the ribosomal atoms and forces on those atoms were not calculated. To accomplish this, it was necessary to define the atoms of the system that corresponded to both the nascent chain and the ribosome. The PDB file used as input for the GROMACS `pdb2gmx` program was designed such that the nascent chain atoms were all adjacent and at the end of the PDB file. Inspection of the resultant 'gro' files allowed the identification of the atom numbers corresponding to the nascent chain. The `make_ndx` program was then used to define two additional groups: those atoms that belong to the ribosome, and those that belong to the nascent chain. The modified index file could then be used as input to the `grompp` program so that the parameters file could refer to these groups and ensure that they are treated appropriately. An example parameter file is shown in Figure 2.5.

```
integrator              = sd
tinit                   = 0
dt                      = 0.002
nsteps                  = -1
init_step               = 0
simulation_part         = 1
comm-mode               = angular
nstcomm                 = 1
bd-fric                 = 0
ld-seed                 = 1993
emtol                   = 10
emstep                  = 0.001
niter                   = 20
fcstep                  = 0
nstcgsteep              = 1000
nbfgscorr               = 10
rtpi                    = 0.05
nstxout                 = 500
nstvout                 = 500
nstfout                 = 500
nstlog                  = 500
nstcalcenergy           = 500
nstenergy               = 500
nstxtcout               = 0
xtc-precision           = 1000
energygrps              = 50S GelFac_NC
nstlist                 = 5
ns_type                 = simple
pbc                     = no
periodic_molecules      = no
rlist                   = 1.0
rlistlong               = 0
coulombtype             = cut-off
rcoulomb-switch         = 0
rcoulomb                = 1.0
epsilon_r               = 1
epsilon_rf              = 1
vdw-type                = cut-off
rvdw-switch             = 0
rvdw                    = 1.0
DispCorr                = No
table-extension         = 1
fourierspacing          = 0.12
fourier_nx              = 0
fourier_ny              = 0
fourier_nz              = 0
pme_order               = 4
ewald_rtol              = 1e-05
ewald_geometry          = 3d
epsilon_surface         = 0
optimize_fft            = yes
implicit_solvent        = GBSA
gb_algorithm            = obc
nstgbradii              = 1
rgbradii                = 1.0
gb_epsilon_solvent      = 80
gb_saltconc             = 0
gb_obc_alpha            = 1
gb_obc_beta             = 0.8
gb_obc_gamma            = 4.85
gb_dielectric_offset    = 0.009
sa_algorithm            = Ace-
approximation
sa_surface_tension      = 2.092
tcoupl                  = No
nsttcouple              = -1
nh-chain-length         = 10
tc-grps                 = system
tau_t                   = 0.1
ref_t                   = 300
Pcoupl                  = No
Pcoupltype              = isotropic
nstpcouple              = -1
tau_p                   = 0.5
compressibility         = 4.5e-5
ref_p                   = 1.0
refcoord_scaling        = No
andersen_seed           = 815131
QMMM                    = no
QMMMscheme              = normal
MMChargeScaleFactor     = 1
gen_vel                 = yes
gen_temp                = 300.0
gen_seed                = -1
morse                   = no
energygrp_excl          = 50S 50S
nwall                   = 0
wall_type               = 9-3
wall_r_linpot           = -1
wall_ewald_zfac         = 3
pull                    = no
nstdisreout             = 100
orire                   = no
orire-fc                = 0
orire-tau               = 0
nstorireout             = 100
dihre                   = no
dihre-fc                = 1000
free-energy             = no
init-lambda             = 0
delta-lambda            = 0
sc-alpha                = 0
sc-power                = 0
sc-sigma                = 0.3
nstdhdl                 = 10
separate-dhdl-file      = yes
dhdl-derivatives        = yes
dh_hist_size            = 0
dh_hist_spacing         = 0.1
couple-lambda0          = vdw-q
couple-lambda1          = vdw-q
couple-intramol         = no
freezegrps              = 50S
freezedim               = Y Y Y
cos-acceleration        = 0
```

Figure 2.5 Example MD parameter file for RNC simulations in which the ribosome is stationery.

2.6 S^2 Model-Free Order Parameter

A useful method for studying NMR and MD data – and for comparing the results of each – is the S^2 model-free order parameter (Lipari and Szabo, 1982, Henry

and Szabo, 1985). This parameter describes the level of 'order' associated with bond vectors, which can be used to infer the propensity of amino acid residues to be within a stable fold, loosely-constrained loop or entirely disordered. Of particular value is the fact that it requires no prior model for the system.

In NMR spectroscopy, the Lipari-Szabo model-free generalised order parameter is calculated as

$$S^2 = \left[\overline{c_0^2}\right]^{-1} \sum_{q=-2}^{2} \left| \overline{c_0 Y_2^q(\Omega)} \right|^2 \tag{2.1}$$

in which c_0 is the spatial variables, Ω is the orientation (ensemble-averaged by the overbar) in the molecular reference frame, and Y_2^q is a modified spherical harmonic function described in Table 2.1.

q	Y_2^q	Y_2^{-q}
0	$(3\cos^2\theta - 1)/2$	$(3\cos^2\theta - 1)/2$
1	$-\sqrt{3/2}\sin\theta\cos\theta e^{i\phi}$	$\sqrt{3/2}\sin\theta\cos\theta e^{-i\phi}$
2	$\sqrt{3/8}\sin^2\theta e^{i2\phi}$	$\sqrt{3/8}\sin^2\theta e^{-i2\phi}$

Table 2.1 Modified second-order spherical harmonics. Adapted from (Cavanagh et al., 2007)

The method to calculate S^2 values from MD trajectories has recently been described in literature (Best and Hummer, 2009). The S^2 parameter for nuclei *a* and *b*, S_{ab}^2 is

$$S_{ab}^2 = \frac{3}{2} tr\left\langle \Phi_{ab} \right\rangle^2 - \frac{1}{2}\left(tr\left\langle \Phi_{ab} \right\rangle\right)^2 \tag{2.2}$$

where Φ_{ab} is defined as

$$\Phi_{ab,\alpha\beta} = \hat{r}_{ab,\alpha}\hat{r}_{ab,\beta}, \alpha, \beta \in \{x, y, z\} \tag{2.3}$$

where $\hat{r}_{ab}$ is the *ab* unit vector.

The S^2 values quoted in this thesis were calculated by first collecting a suitable, nanosecond scale trajectory as output by the GROMACS simulation program, `mdirun`. This 'trr' format trajectory was converted to a portable binary format 'xtc' trajectory using the GROMACS `trjconv` program. Lists of bond vectors for which to calculate the S^2 value were extracted from the molecular topology using a custom Perl script. For all-atom simulations, the N-HN bond vector within each amino acid residue was used. For coarse-grained simulations, the bond vector between the C_α of residue i and residue i-1 was used. The bond vector list, structure topology and portable trajectory files were then used as input to an adapted build of the `s2` program, kindly provided by Dr Robert B. Best, Laboratory of Chemical Physics, National Institute of Health.

2.7 NMR experiments

NMR data presented here were collected on a 700MHz Bruker Avance III spectrometer. Samples were maintained at 25°C. A calibrated water frequency was used as the ^1H channel carrier frequency. Data were collected with Topspin 2.1.

Spectra were obtained from band-selective optimised flip angle short transient heteronuclear multiple quantum coherence (SOFAST-HMQC) experiments (Schanda et al., 2005).

2.8 Perl API library

The nature of the work under consideration in this thesis requires frequent manipulation of data and conversion between different file formats. For example, published NMR chemical shift data are typically provided in a format called NMR-STAR, while the same data are formatted differently for restraints in the GROMACS CS-MD restraint files, and again in another format for use with camshift. Manually converting each file between different formats would be time consuming and error prone (from, for example, typographical errors). Some tools are published for making some file format conversions, but work only for the original purpose. To handle these problems, as well as accomplishing many routine tasks, a software library may be authored that does not attempt to accomplish a single task but instead provides an interface for any new software

tool to accomplish the task. For example, a software library might be authored that allows reading of NMR-STAR files. This library would then be useful for any other software tool that wishes to read the files without having to replicate or duplicate the same work. If an error is found with the library, it may be fixed and have such improvements inherited in all software that uses the library.

Therefore, in order to allow new software tools necessary for accomplishing all other work conducted within this thesis, a software library was authored to accomplish any and all common tasks for which existing software libraries are unavailable in order to reduce the time needed to create new tools and to allow for few errors in any newly authored tools.

The Perl programming language is a mature scripting language that allows rapid development of new software tools and has a powerful and efficient text-processing facility. These features make it a suitable language to use for development of the software library described above.

As well as and API for reading NMR-STAR files, APIs were also written to allow the programmatic interaction with camshift, GROMACS tools (mdrun, pdb2gmx, grompp) HydroPro (Ortega et al., 2011), Sun Grid Engine, the Multiscale Modeling Tools for Structural Biology Toolset (Feig et al., 2009) and Sparta+ (Yokota et al., 2002).

Approximately 80 software programs that make use of the API for the handling and analysis of the data generated were authored. The library and these tools were made available via the group website at http://jcgroup.biochem.ucl.ac.uk/.

3 Cell-free RNC NMR sample preparation feasibility

3.1 Introduction

As detailed in Chapter 1, NMR study of biomolecules offers the ability to derive solution-state structural and dynamical properties. Moreover, it may offer tantalising glimpses into these properties of nascent chains as they emerge from a ribosome. It is therefore necessary to identify a protocol for the preparation of RNC samples that can be studied by NMR spectroscopy.

Preparation of RNC samples for previous, non-NMR studies has been performed principally by cell-free methods in small quantities. For study by NMR spectroscopy, much larger sample sizes (approximately 300µl at 5-10µM concentration) are required. Despite the dramatic advances in large-scale *in vivo* preparation of RNCs, samples produced by cell-free, *in vitro* methods offer tantalising advantages over their *in vivo* counterparts. These include the ability to precisely control the constituents of the sample, for example excluding co-factors such as Trigger Factor or incorporating selective amino-acid labelling, as well as a potentially simpler purification protocol, which would reduce the purification time and potential for sample degradation.

The aim of this avenue of work is to identify the most suitable method for the preparation of RNC samples for study by NMR. This includes investigating the feasibility of preparing such samples by *in vitro* methods and – if found to be feasible – to develop an optimised protocol for this application.

3.2 Methods

Most research on RNC sample preparation for NMR has focussed on *in vivo* production. This has been successful, with considerable data collected (Christensen et al., 2011, Hsu et al., 2007, Hsu et al., 2009b, Hsu et al., 2009c). The process entails significant growth of *E. coli* and is therefore more difficult to have precise control of cofactors such as Trigger Factor.

3.2.1 Plasmids

The αSyn RNC DNA vector was designed by Dr Lisa D. Cabrita (UCL). It is provided as a circular plasmid (αSyn-pLDC17) – a modified pLDC17 plasmid,

which uses the Novagen pET21b(+) plasmid as a backbone (Appendix 8.1). The RNC sequence is regulated by the *lac* operon, allowing expression to be induced by Isopropyl β-D-1-thiogalactopyranoside (IPTG) and consists of the αSyn sequence with sequence additions at the termini. At the 5'-terminus, six histidine residues and an NdeI-NheI restriction site has been added to provide a pre-translationally-removable His_6-tag. At the carboxy-terminus, six restriction site residues (KpnI, SpeI and EcoRI) followed by seventeen SecM residues, to provide a nuclease-labile SecM stall sequence. Other constructs available include α_1-antitrypsin (A1AT)-RNC and T17-RNC (a derivative of *Dictyostelium discoideum* filamin, ddFln, protein), both similarly based on the pLDC17 vector. Though not directly relevant to the project, these would produce nascent chains of different weights and properties to the αSyn-RNC polypeptide, which could be useful for diagnostic purposes.

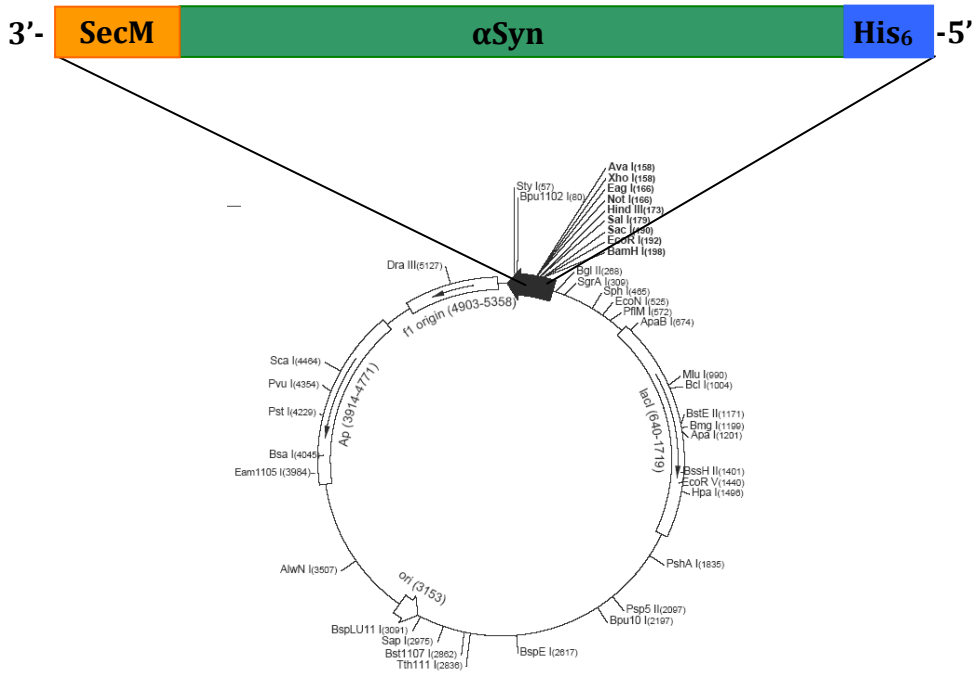

Figure 3.1. Schematic of the pLDC17 plasmid incorporating the αSyn-RNC sequence. Adapted from (Oldfield, 2002).

Plasmid stock was produced by growth in DH5α *Escherichia coli* cells and use of QIAGEN QIAprep Spin Miniprep and Maxiprep Kits, following the protocols provided. When initially I used my αSyn-RNC plasmids in cell-free reactions I

found no observable activity. Upon running the samples in an agarose gel, there appeared to be considerable contamination, thought to be RNA (Figure 3.2A). Repeated Minipreps showed removal of the contamination but still predominantly nicked DNA. Reactions still yielded no observable product. Eventually, it was found that modifying the protocol by treating the cell culture with lysis buffer for 5 minutes, rather than 30 seconds, more supercoiled DNA was obtained (Figure 3.2B) and the cell-free reaction yielded additional observable product.

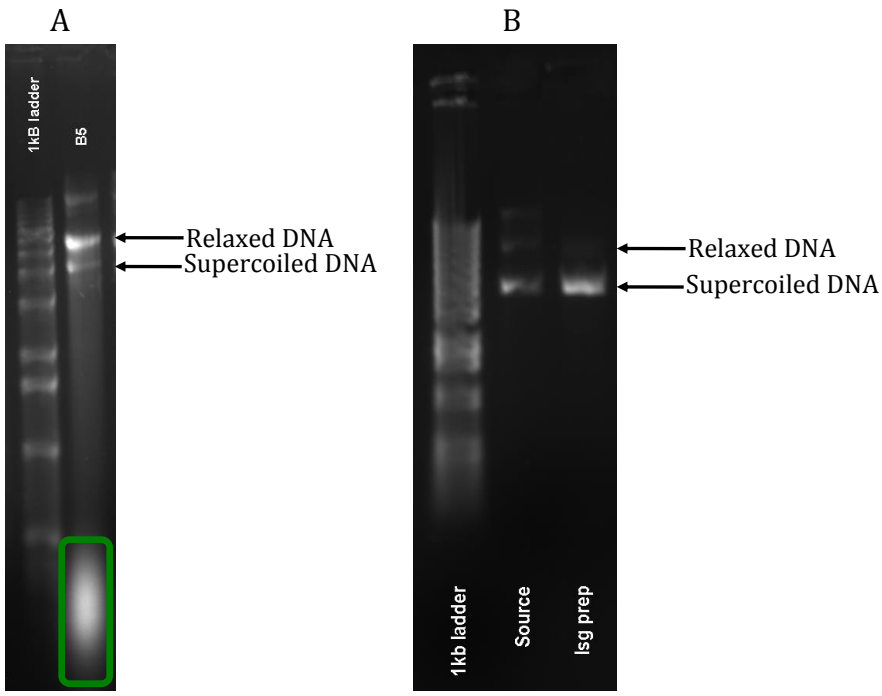

Figure 3.2. EtBr/Agarose gels of selected attempts at producing αSyn plasmid with samples loaded from the top. A) The plasmid from the first Miniprep attempt. The large indiscrete band (demarked green) corresponds to a large amount of heterogenous, low-mass contaminant. B) The original plasmid and successful copy, produced with a longer lysis step. The contaminant present in previous preparations has been removed and the ratio of supercoiled DNA to relaxed DNA has increased from 1:5 to 10:1.

3.3 Results

3.3.1 *In* vitro sample synthesis

Cell-free reactions were conducted using Roche RTS100 E. coli HY kits with the addition of RNase-inhibitor to prevent degradation of mRNA by contaminants. The initial plans were to confirm the production of RNCs using small-scale (10µl) cell free reactions using RTS100 kits and then to identify the conditions that

yield optimal ^{15}N- and/or ^{13}C- labelled RNC samples. Once this was complete, large-scale reactions, using Roche RTS500 E. coli kits, would be used to produce sufficient sample (300µl at > 1µM) for study such as by NMR spectroscopy.

Initial test runs were performed on cell-free synthesis of isolated, released YFP as a suitable plasmid and purified YFP was readily available, providing sufficient resources for training and testing. Experiments were also conducted to assess the effect of reaction time (5 to 120 minutes) and IPTG (0.2mM to 1mM in reaction) in an attempt to increase yield. Results showed that high concentrations (>8mM) of IPTG proved toxic to the reaction as no synthesised protein could be detected, but lower concentrations (≈0.4mM) could increase YFP yield. While the yield of YFP does not necessarily imply that an improved yield of RNCs would be achieved (since each ribosome is only used once), these showed that addition of 0.4mM IPTG may give larger concentrations of mRNA transcripts, which could increase occupancy of the ribosomes in RNC reactions.

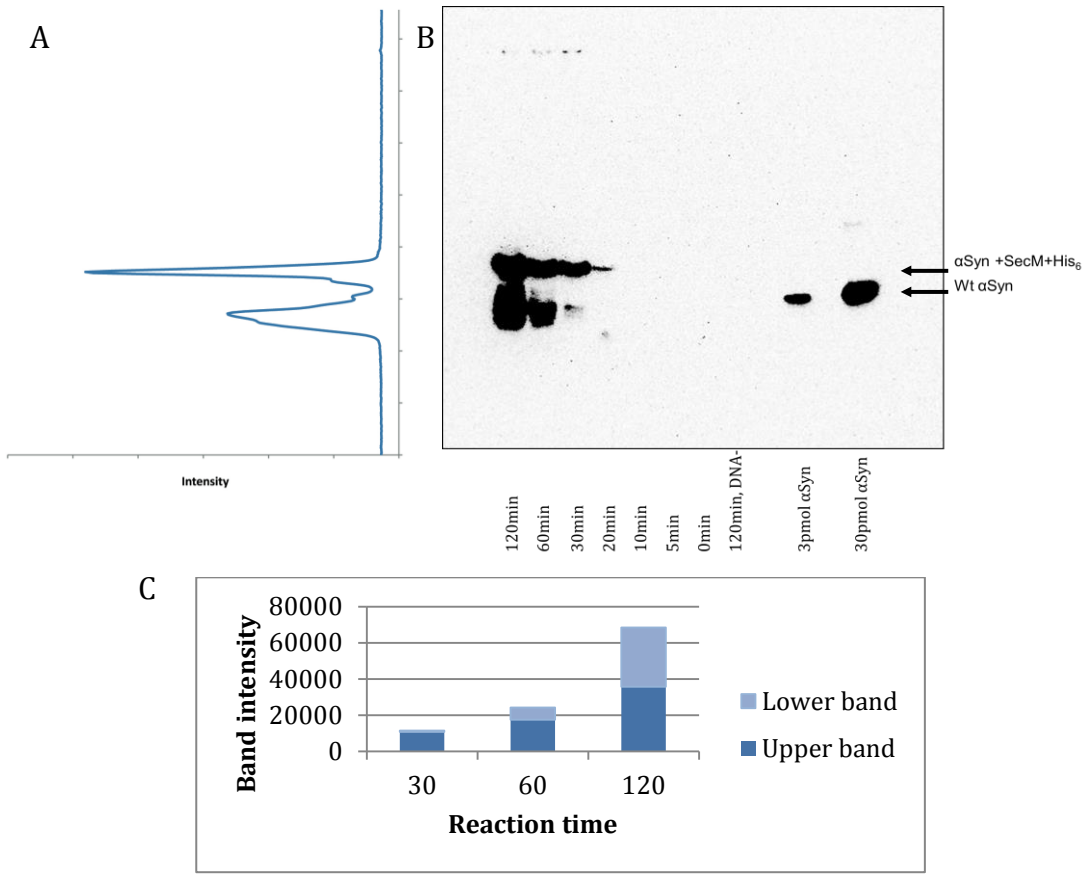

Figure 3.3. Effect of reaction time on αSyn-RNC synthesis. A and B) Anti-αSyn western blot of pellets after sucrose cushion centrifugation of different reaction times of αSyn-RNC synthesis without protease-inhibitors and the intensity of signal across the first lane of the western showing multiple peaks contributing to the lower-mass material. Two concentrations of pure αSyn are included for reference indicating greater than 30pmol yield after 120 minutes of reaction. The antidbody used is BD Transduction Laborities Purified Mouse anti-α-synuclein 610786. C) Comparisson of the peak intensity of the upper band (αSyn-RNC) and the lower band over the period for which both could be reliably measured.

αSyn-RNC reactions unexpectedly produced two observable products that on a western blot (using anti-αSyn antibody) with molecular weights of marginally above and below that of isolated wt αSyn (Figure 3.3B). The expected αSyn-RNC product is 17kDa, so the additional lower-mass product could be a proteolysed form of the product Consistent with this, the lower-mass signal appears after the αSyn-RNC band (Figure 3.3C) in a time-course expression. Moreover, the lower mass band was found to comprise multiple overlapping peaks, suggesting that multiple proteolysis products were produced (Figure 3.3A). Protease inhibitors

were then subsequently added to the reaction mixtures, yielding a single product visible by anti-syn western blot (Figure 3.4).

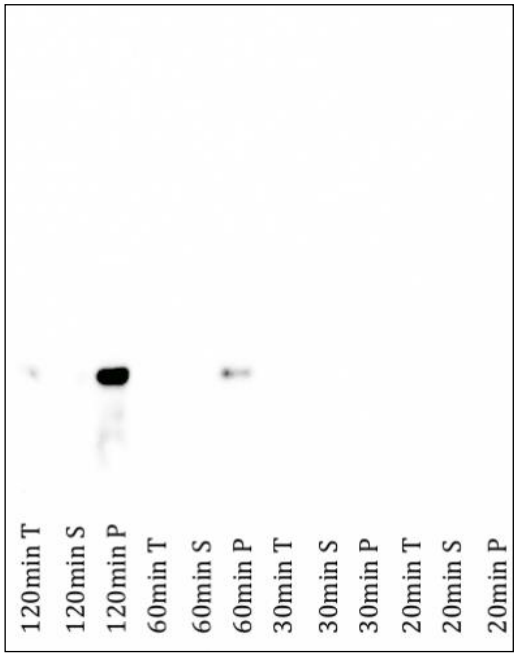

Figure 3.4. Representative western-blot image of αSyn-RNC synthesis with protease inhibitors present in the reaction. Aliquots were taken from a single reaction at different times and run through a 1ml sucrose cushion (see section 3.3.3). Lanes marked 'T' were loaded with the top 200µl of the cushion; lanes marked 'S' were loaded with the remaining supernatant; and lanes marked 'P' were loaded with the resuspended pellet. All other reaction conditions were the same as in the reaction that produced the western in Figure 3.3B.

3.3.2 Labeled Amino Acid Synthesis

The amino acid source for the unlabelled reactions was lyophilized amino acid mixture resuspended in a "reconstitution buffer" supplied with the Roche RTS100 E. coli HY kits. Two ^{15}N-labelled amino acids mixtures were available: Isotec 608947 and Cambridge Isotopes Ltd NLM6695. The Isotec mixture was less expensive, so initial reactions were conducted with this. Literature for the Roche cell-free expression kits suggests an in-reaction concentration of the amino acids ranging from 1.51mM (for Leu) to 2.0mM (for Ala, Arg, Asp, Gly, His, Iso, Lys, Ser, Pro, Thr, Val, Asn, Met, Cys, Tyr, Trp and Phe) for small-scale (≤50µl) reactions. The solubility limit of the Isotec mix in reconstitution buffer was found to be 82.67mg per ml, which corresponded to concentrations of amino acids in the reaction ranging from 0.61mM for Cys to 21.52mM for Ala (Table 3.1). 10µl

αSyn-RNC reactions were performed with the 2.4µl unlabelled amino acids substituted with reconstitution buffer saturated with Isotec 608974 amino acids and the 0.2µl methionine substituted for 0.2µl 100mg/ml Trp. Initial results showed synthesis of labelled αSyn-RNC of similar yield and behaviour to that from reactions with unlabelled amino acids (Figure 3.5).

	Maximum concentration in reaction (mM)	
Amino acid	**Isotec 608947**	**CIL NLM6695**
Ala	21.5	8.4
Arg	4.5	8.4
Asn	7.8	7.0
Asp	7.8	11.2
Cys	0.6	4.2
Gln	7.8	7.0
Glu	7.8	12.6
Gly	16.6	7.0
His	1.4	1.4
Ile	8.9	4.2
Leu	14.8	12.6
Lys	6.0	16.7
Met	4.2	1.4
Phe	6.8	5.6
Pro	7.2	7.0
Ser	6.8	5.6
Thr	9.5	5.6
Trp	0.0	4.2
Tyr	1.2	4.2
Val	12.5	5.6

Table 3.1. Concentration of each amino acid within each reaction for each ^{15}N-labeled amino acid mix when using a saturated stock solution. Note that Tyrosine is not present in the Isotec 608947 mix and so unlabeled tyrosine was added at 2mM in accordance with the concentration advised by the Roche RTS literature (Katranidis et al., 2011).

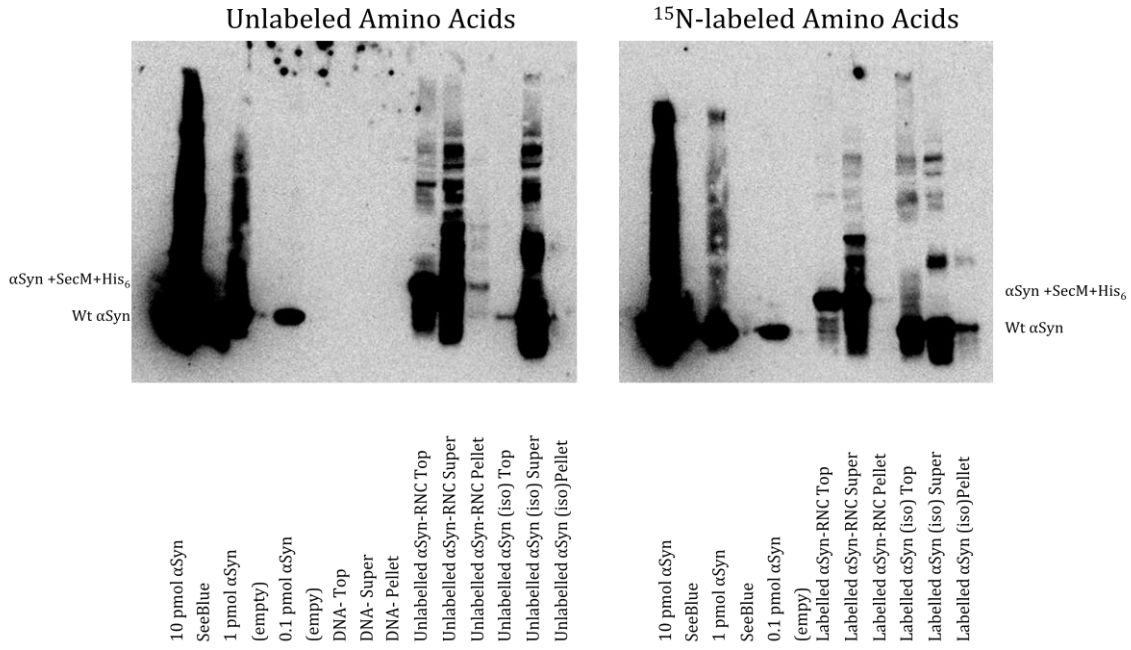

Figure 3.5. Anti-αSyn western showing expression of αSyn with ^{15}N-labeled amino acids and unlabelled amino acids.

To investigate whether the higher minimum concentration of amino acids offered by the CIL NLM6695 amino acids improved yield, concurrent reactions were conducted where aliquots of reaction mixtures without amino acids were taken and equal volumes of saturated amino acid mixtures were added. Analysis of the reaction products showed similar yield and banding patterns in western-blots (Figure 3.6A). Furthermore, use of different amino acid mix did not significantly improve the relative yield of RNC within the cushion ($P>0.1$, Figure 3.6B). As the Isotec mixture appeared to confer no disadvantage over the more expensive alternative, it was used for all subsequent reactions with labelled amino acids.

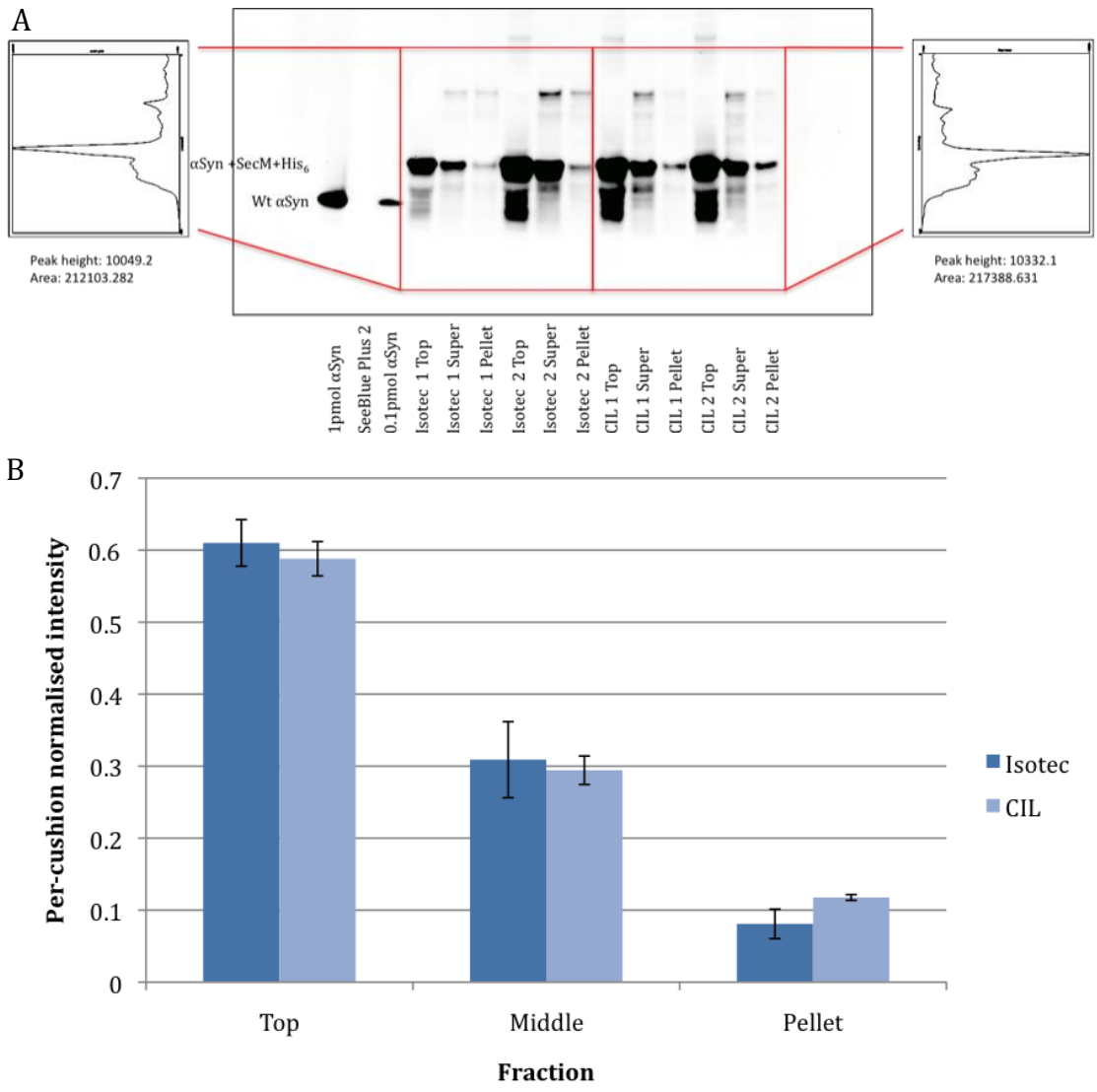

Figure 3.6. The effects of different labelled amino acid sources. A) Anti-αSyn western showing reaction product from reactions using Isotec and Cambridge Isotopes Ltd ^{15}N-labeled amino acids. Graphs at the side show the average intensity along the Isotec 608974 lanes (left) and CIL lanes NLM6695 (right). B) Per-cushion normalised intensities of the αSyn-RNC peak for each cushion fraction. Bar heights denote average band intensity; error bars denote the maximum and minimum value.

3.3.3 Purification

After synthesis of isotopically-labeled αSyn-RNC had been confirmed, efforts were made to improve the yield of RNCs purified from the reaction mixture. It would need to be demonstrated that a single large-scale (5ml) reaction could produce a useable NMR sample of 300µl at ≥5µM RNC, so small-scale (10µl) reactions would need a yield of at least 3×10^{-12} moles.

Two protocols were used for the analytical purification of RNC reactions: acetone precipitation to readily isolate all proteins from a reaction mixture and sucrose cushioning with methanol chloroform/extraction to separate ribosome bound proteins from released proteins.

Acetone precipitation involved adding 100µl acetone (pre-chilled to -20°C) per 10µl reaction, leaving on ice for 20 minutes to precipitate proteins, centrifuging in an Eppendorf 5424 Centrifuge for 10 minutes at 14,680rpm, decanting the supernatant and drying the pellet. The precipitate can be resuspended in Tico buffer (Appendix 8.4).

By placing the sample on the top of a buffer containing a carefully selected concentration of sucrose and centrifuging this under certain conditions, more dense material (including ribosomes and RNCs) move to the bottom of the solution and produce a pellet, while the less dense material (including released nascent chains) remain near the top. This technique – referred to as sucrose cushioning – provides a fast, non-destructive means to separate released nascent chain from that still attached to the ribosome. Previous work by Dr Lisa Cabrita had shown that sucrose concentrations between 0.5M and 1.5M with centrifugation at 120,000rpm in a Beckman Optima Max centrifuge with TLA-120.1 rotor for 30 minutes would separate other RNC constructs from their isolated (released) proteins.

Methanol chloroform extraction involved adding 150µl chloroform to 150µl sample, vortexing, adding 450µl water, incubating on ice for 2 minutes, centrifuging in an Eppendorf 5424 centrifuge at 14,680rpm for 5 minutes, removing the aqueous top layer, adding 600µl methanol, vortexing, centrifuging for a further 5 minutes and discarding the supernatant.

While optimisations for reaction conditions were taking place, many of the purification conditions were investigated so as to improve RNC yield. Of particular concern was that substantial (>99% in the first labelled-amino acid reactions) αSyn-RNC material was being produced but was not pelleted in the sucrose cushion (Figure 3.7A). Moreover, most non-pelleted material diffused into the bulk supernatant without being forced to the top (Figure 3.7B), as would be expected of released proteins by more dense sucrose solution. These

suggested that either RNCs were not sufficiently being pelleted, or that RNCs were degrading during centrifugation and nascent chains moving to the top.

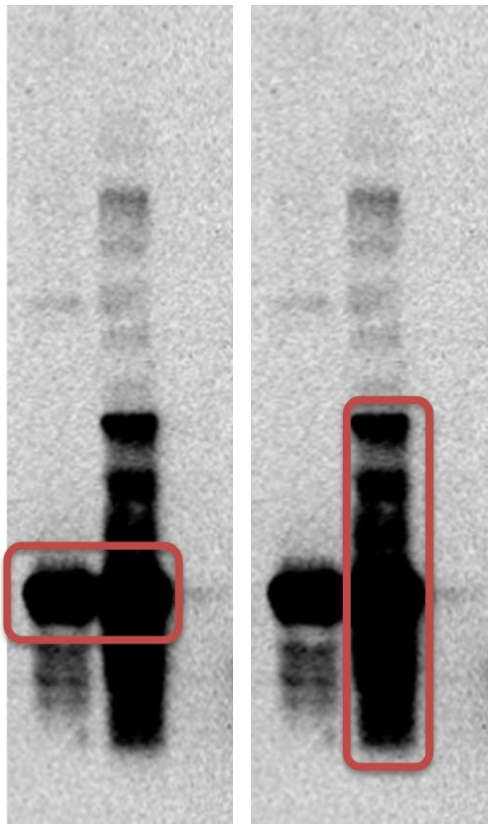

Figure 3.7. Representative anti-αSyn western-blot image of RNC synthesis products following sucrose cushioning illustrating concerns regarding yield. Both images are of the same three lanes taken from the same image as in Figure 3.5, loaded with (left to right) the top 300µl of supernatant, the remaining 760µl supernatant and the pellet. A) the red-enclosed region shows the αSyn-RNC protein material that has not been pelleted during sucrose cushioning. B) The red-enclosed region shows the material that has neither been pelleted nor forced to the top of the cushion. That multiple overlapping bands are produced suggests a complex interaction or reaction has occurred that produces material with the anti-αSyn epitope at multiple molecular weights.

The centrifugation speed was varied while keeping constant the sedimentation coefficient, *s*, defined as

$$s = \frac{k}{t} \tag{3.1}$$

where *t* is the centrifugation time and *k* is the clearing factor, defined as

$$k = \frac{253{,}000\left[\ln\left(\frac{r_{max}}{r_{min}}\right)\right]}{\left(\frac{rpm}{1000}\right)^2} \tag{3.2}$$

where r_{max} and r_{min} are the maximum and minimum radii of centrifugation respectively and *rpm* is the revolutions per minute of the rotor (Schuwirth et al., 2005). As the rotor type (and therefore r_{max} and r_{min}) is kept constant, a new centrifugation time, t_2, for a new centrifugation speed, rpm_2, with the same sedimentation coefficient can be calculated as

$$t_2 = t_1\left(\frac{rpm_1^2}{rpm_2^2}\right) \tag{3.3}$$

Using the initial conditions of 30 minutes and 120,000rpm for t_1 and rpm_1 respectively, three centrifugation speeds of 120,000rpm, 90,000rpm and 60,000rpm were assessed initially, corresponding to durations of 30, 53 and 120 minutes respectively. It was decided that slower speeds would be unnecessary as these speeds already cover a 75% drop in centrifugal force experienced by the sample, exposing any effect potentially present. Moreover, it was considered that for slower centrifugation speeds, the extra preparative time would contribute more to the sample degradation than the centrifugation itself. As can be seen in Figure 3.8, lower centrifugation speeds did not demonstrate a statistically-significant increase in pellet yield ($P > 0.3$), so the highest speed possible was used thereafter to reduce the preparation time.

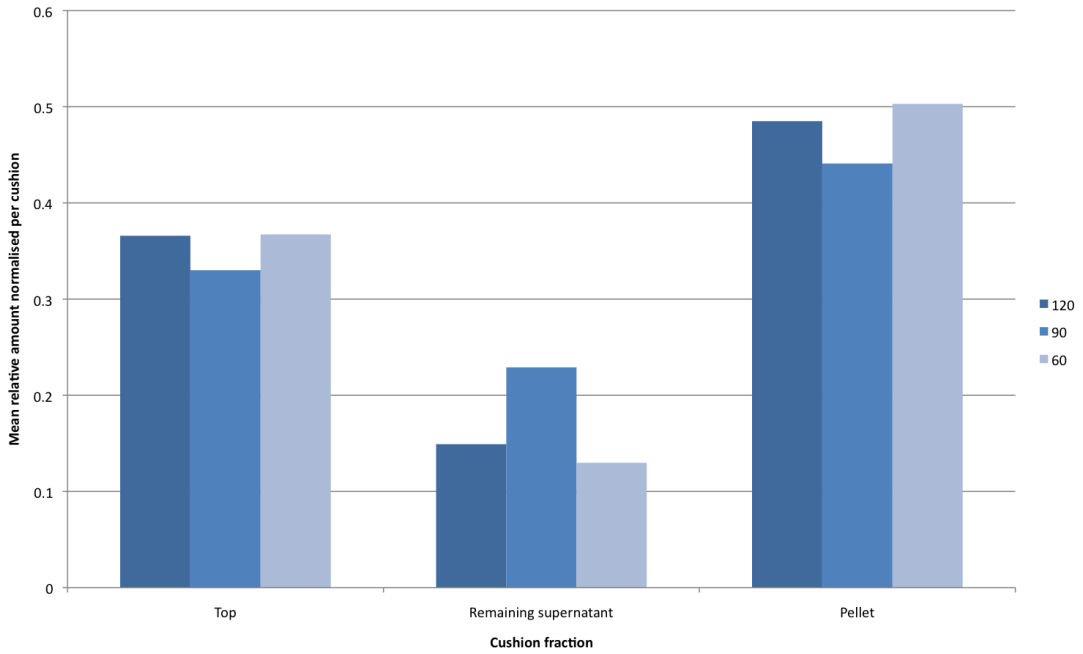

Figure 3.8. Chart of mean relative αSyn-RNC protein detected in each of three fractions – top 300µl of supernatant, remaining 760µl of supernatant and pellet – under three centrifugation regimes – 120,000rpm for 30 minutes, 90,000rpm for 53 minutes and 60,000rpm for 120 minutes.

Previous studies had suggested that a high (0.5M to 1M) concentration of potassium acetate improves the stability of RNCs during purification (Schaffitzel and Ban, 2007). In agreement with this, it was found that increasing the concentration of potassium acetate in the sucrose cushion reduced RNC decomposition, more than doubling the yield of pelleted αSyn-RNC (Figure 3.9).

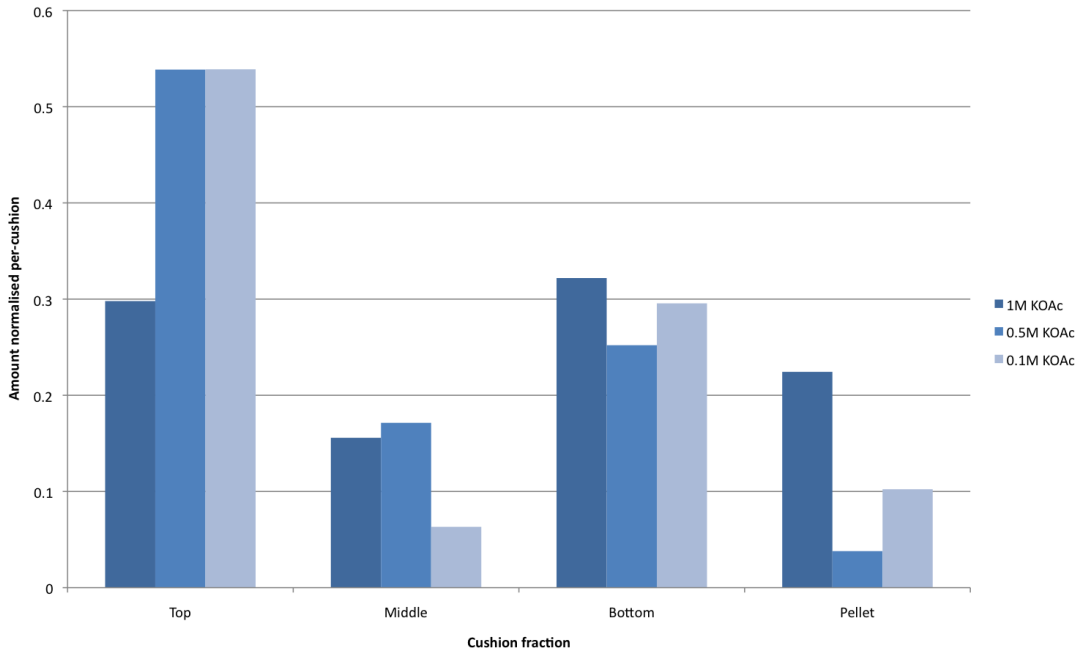

Figure 3.9. Chart of per-cushion-normalised mean relative αSyn amount in each of four sucrose cushion fractions – Top 300µl, middle 380µl, bottom 380µl and pellet – under three different potassium acetate concentrations.

3.3.4 Detection of Nascent Chains

Coomassie- and silver- stained SDS-PAGE gels of RNCs yield many bands produced by ribosomal proteins, which makes identifying the band for the desired protein difficult and error-prone. Western blots with antibodies targeting the His_6 tag should be much more selective, binding only to the desired protein and few background bands (due to off-target affinity), and potentially much more sensitive.

Initial western blots were performed using QIAGEN Penta His HRP Conjugate, an antibody that recognises an epitope of five consecutive histidine residues. This produced many background bands associated with ribosomal proteins. When synthesising isolated YFP, the yield of protein was sufficient that the protein could be identified by comparing results to a reaction lacking DNA. However, it was found that one ribosomal protein band migrated at the same molecular weight as the αSyn-RNC protein, masking the presence (or otherwise) of the protein. Therefore, an alternative antibody was used to specifically target the carboxy-terminal residues of αSyn (Figure 3.10).

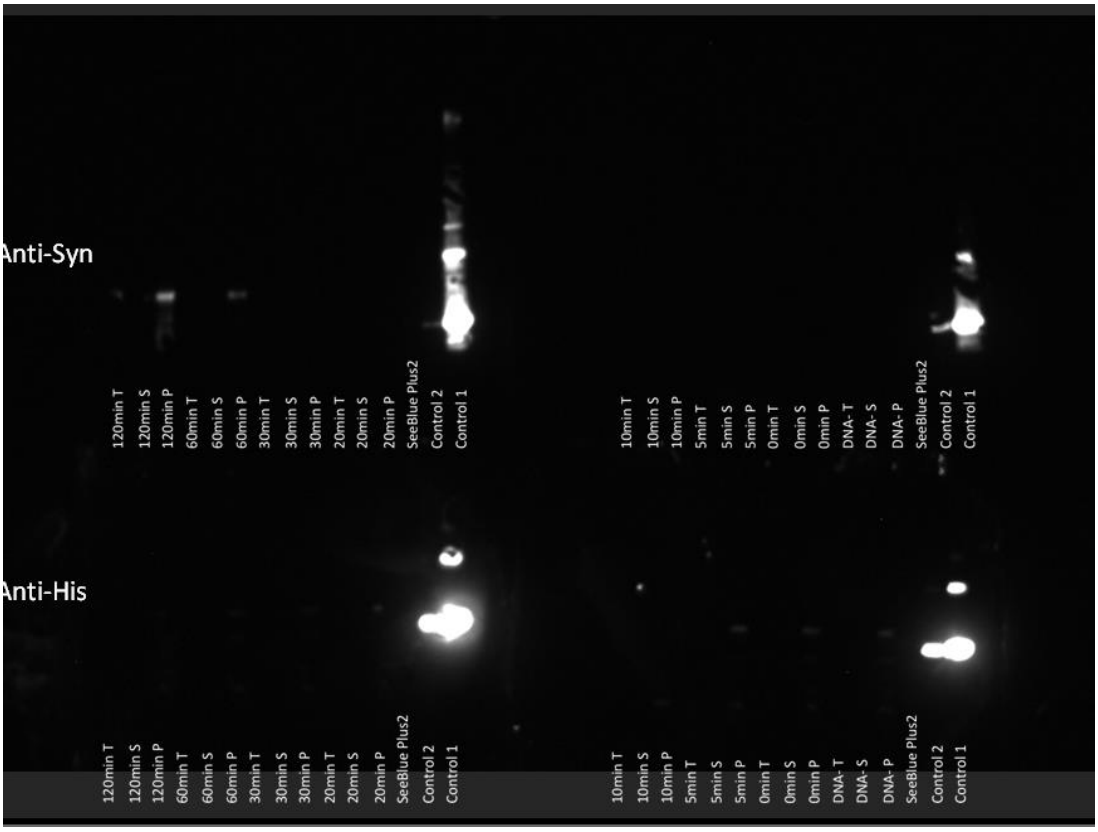

Figure 3.10. Western blots of αSyn-RNC production over varying time intervals from 0 minutes to 120 minutes with protease-inhibitors present. The two top membranes were treated with anti-αSyn, the two bottom membranes with anti-penta-his. After the reactions were complete, the samples were placed on a sucrose cushion to separate the ribosome-bound protein from released protein. Each cushion was separated into three fractions: the top 150µl ('T'), the remaining supernatant ('S') and the pellet ('P'). Each membrane was also loaded with the same two controls containing two concentrations of purified synuclein and and YFP.

The sensitivity of the western blots using the existing protocol was capable of detecting no less than 30pmol of YFP, less than observable by staining the nitrocellulose membrane with Ponceau-S solution (a non-specific protein stain). It was hypothesised that the blocking solution (skimmed milk) was too stringent and was preventing binding of the antibody. The effect of using a proprietary blocking solution (QIAGEN Blocking Reagent) was investigated by comparison to skimmed milk at various concentrations of both blocking solution and antibody to see if improved sensitivity could be achieved while maintaining selectivity. Furthermore, the existing protocol involved incubating the membrane with antibody in the washing buffer (TBS-T: 20 mM Tris-HCl , pH 7.5, 150 mM, NaCl,

0.05% Tween20) while the Blocking Reagent protocol advised incubating in blocking solution. The effect of this was therefore investigated in tandem. As can be seen in Figure 3.11, blocking with Blocking Reagent followed by incubating in Blocking Reagent gives considerably greater sensitivity, to 0.3pmol YFP.

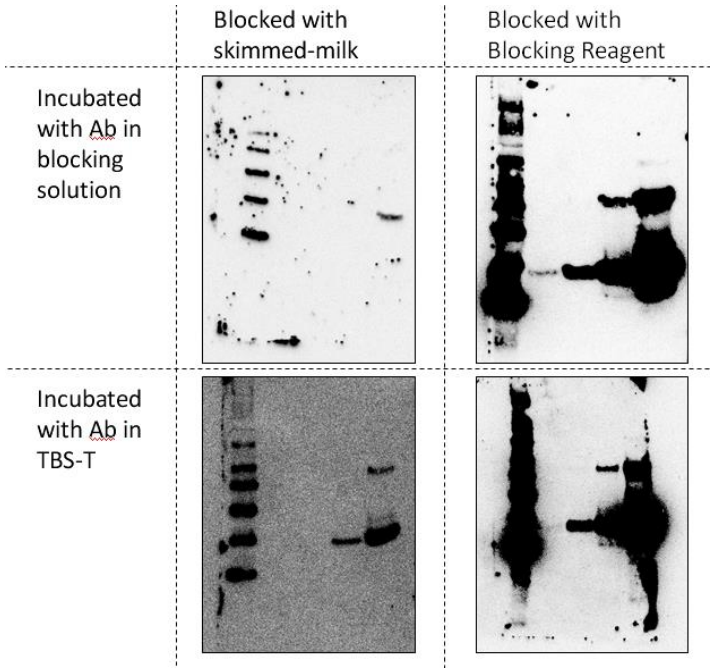

Figure 3.11. Images of chemiluminescent western blots blocked and incubated under different conditions. Each membrane is loaded identically: lane 1: 1µl MagicMark XP marker; lane 2: 0.3pmol YFP; lane 3: 3pmol YFP; lane 4: 30pmol YFP; lane 5: 300pmol YFP.

Anti-syn western blotting was optimised similarly. No preconjugated anti-αSyn – HRP antibody was available, so a two-stage incubation was followed. When maximum sensitivity was required, the primary anti-αSyn antbody placed in blocking solution at 1:1000 dilution and incubated with the membrane for 12-16 hours at 4°C on a rotary shaker. The membrane was washed 3 times for 10 minutes in TBS-T and then incubated with the anti-IgG-HRP secondary antibody at 1:10,000 dilution in blocking solution, before washing and developing as with the anti-penta his antibody.

Following the advice from QIAGEN, it was found that a solution of 0.5% casein in TBS-T provided comparable blocking to the proprietary Blocking Reagent, while being less expensive and easier to handle.

Anti-αSyn western blots following this protocol were able to achieve sensitivity to less than 1pmol of αSyn when using the Thermo Scientific SuperSignal West Pico Chemiluminescent Substrate.

3.3.5 Quantification

To allow comparison between reactions, and to assess the yield obtained, it was intended that the yield from each small-scale reaction would be quantified. Since western blots were being produced for detection already, it was decided that these would be used for quantification. Since using different known concentrations of purified protein to produce a calibration curve for the signal intensity gives most reliability (Charette et al., 2009), it was hoped that each western blot would include several lanes of purified YFP to produce such a calibration curve. Unfortunately, due to the limited number of lanes available, this was often infeasible. Instead, since each western blot included a lane containing a ladder of known-weight markers, each of different intensity, calibration curves of YFP intensity and marker intensity were produced from the same western blot using intensities measured using ImageJ (Atieh et al., 2010) (Figure 3.12). This would allow bands on future western blots to be quantified based on a calibration curve for the markers. However, once it was found that anti-synuclein antibody must be employed (which shows no affinity for any of the marker bands), this method was no longer of use. Instead, quantification was conducted by including at least three different concentrations of purified α-synuclein on each western blot.

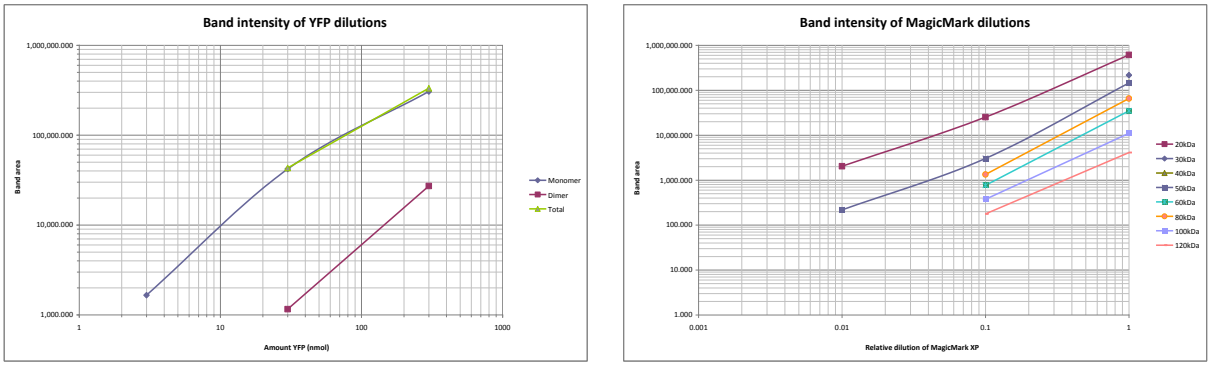

Figure 3.12. Calibration curves for YFP dilutions (left) and marker dilutions (right). Future western blots could be quantified by producing a new calibration curve for the marker lane, comparing the desired band to the new calibration curve to find the equivalent intensity on the above curves, yielding an equivalent amount of YFP.

3.3.6 *In vivo* Sample Preparation

In addition to *in vitro* methods, the preparation of an αSyn-RNC was also explored using *in vivo* methods in *E.coli*. The plasmid previously used for the *in vitro* preparation was transformed into *Escherichia coli* of the strain BL21-DE3 – a cell line adjusted to synthesise large quantities of protein via IPTG induction (Studier and Moffatt, 1986), although replication of DNA is more error-prone than in the DH5α cell line used for plasmid production (Grant et al., 1990).

The cells were first grown up in MDG medium (Appendix 8.2) to produce large numbers of log-phase cells with unlabelled ribosomes. The cells were then pelleted and resuspended in EM9 medium (Appendix 8.3) containing $^{15}NH_4Cl$ and transcription of the RNC sequence was induced by IPTG. Rifampicin is added to the medium prior to induction to prevent the synthesis of undesired protein products, which would incorporate ^{15}N and so be visible in ^{15}N NMR spectroscopy.

Aliquots of the expression media were taken at 5 time intervals over a 2 hour period and the cells immediately pelleted at 4°C to reduce further expression. The cells were lysed and their contents run on a sucrose cushion. A western blot of the supernatant and pellet from each time interval was saturated, but showed that significant RNC was produced after 15 minutes, and significant released αSyn was present after 30 minutes. This suggested that expression of the αSyn-

RNC should last just 15 minutes to reduce the released material present in the sample.

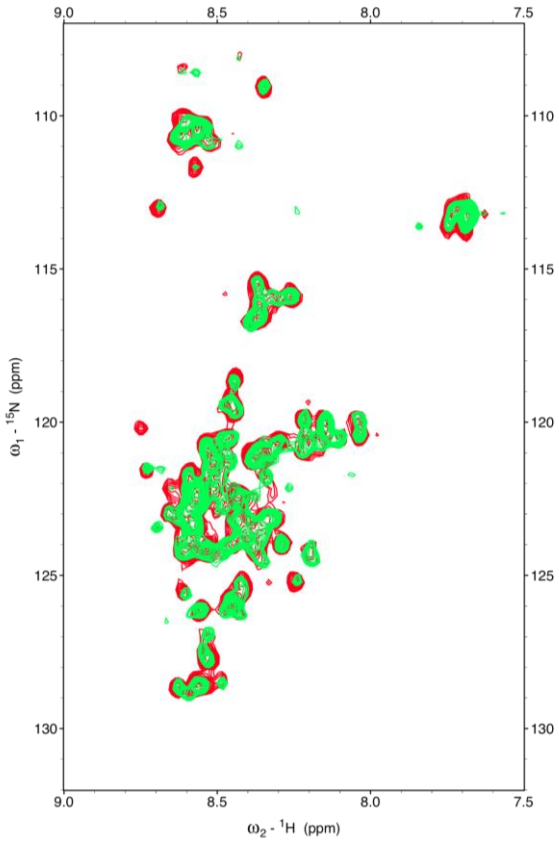

Figure 3.13. ^{1}H-^{15}N HSQC NMR spectrum of αSyn-RNC (red) and αSyn-RNC with Trigger Factor (green) provided by Dr Lisa Cabrita (UCL) and Dr Chris Waudby (UCL).

Another expression was conducted to produce an NMR sample. A larger culture of cells was used, so as to produce a suitable quantity of RNC for the NMR sample. Although the construct includes an N-terminal His_6 tag that should be suitable for binding and purification on a nickel column, the anti-penta his western blots repeatedly failed to clearly bind to the desired protein. Furthermore, previous attempts to purify this construct on such a column have failed to show significant binding to the column. In order to reduce the time during which that the sample could deteriorate, it was decided to rely on the sucrose cushion and sucrose gradient as the primary means of purification of the sample. This would remove released protein and non-70S ribosomes but vacant ribosomes would still be present in the sample, reducing the occupancy. As ribosomes become unstable at concentrations much higher than 10μM, having a

lower occupancy reduces the maximum concentration of RNCs in the NMR sample, which reduces the signal intensity.

Following sucrose gradient fractionation, the fractions containing ribosomes (as identified by absorbance of UV light at 254nm) were pooled, concentrated and exchanged into a Tico buffer. 10% D_2O (v/v) with 1% DSS were added to create a solution of 330µl at 3.6µM ribosomal material. A ^{1}H-^{15}N SOFAST HSQC spectrum was recorded for the sample.

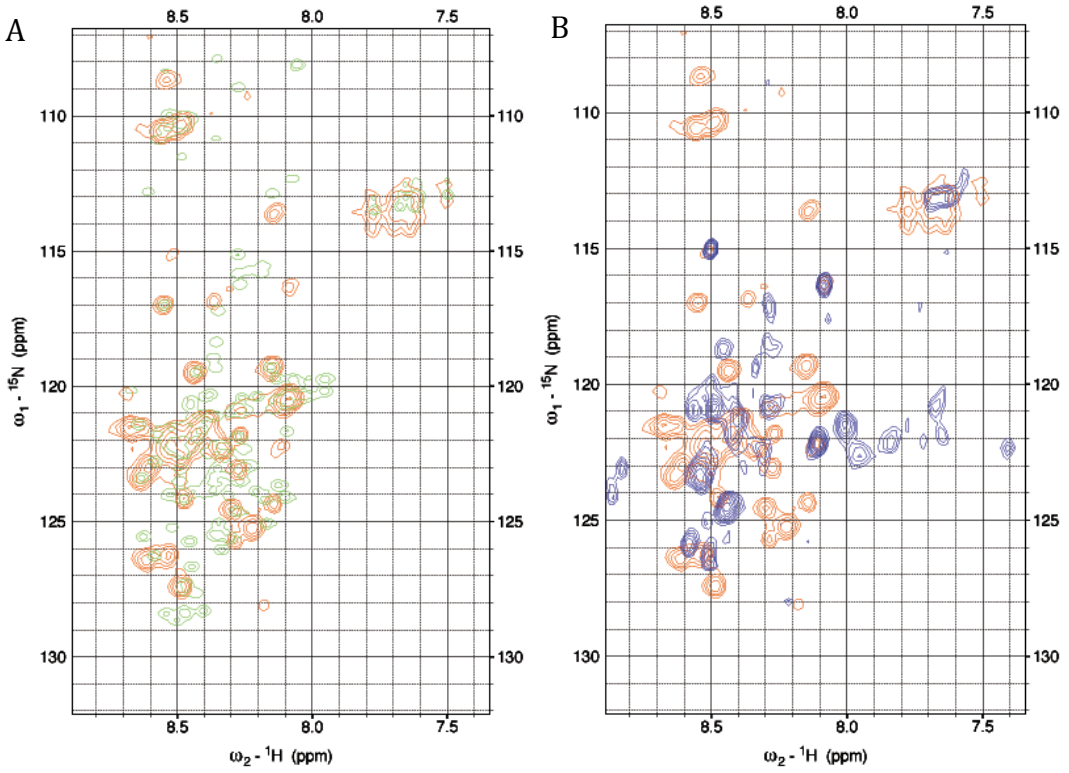

Figure 3.14. The ^{1}H-^{15}N spectrum of the αSyn-RNC sample (red) overlaid with that of isolated αSyn (A, green) and that of labelled L7/L12 (B, blue). The spectrum of isolated αSyn was provided by Dr Carlos Bertoncini (Cambridge University) and that of ribosome by Hélène Launay (UCL).

Comparison of the RNC spectrum with that of isolated αSyn showed significant differences; some peaks clearly overlaid but much of the signal did not, suggesting that the signal did not derive from a complete αSyn sequence (Figure 3.14A). Ribosomal protein L7/L12 is known to readily dissociate from the ribosome. It was therefore considered that L7/L12 might have synthesised during RNC expression, incorporating ^{15}N and giving rise to signal in the

spectrum. Indeed, some peaks appeared to overlay well with a spectrum of labelled ribosome L7/L12 alone (Figure 3.14B).

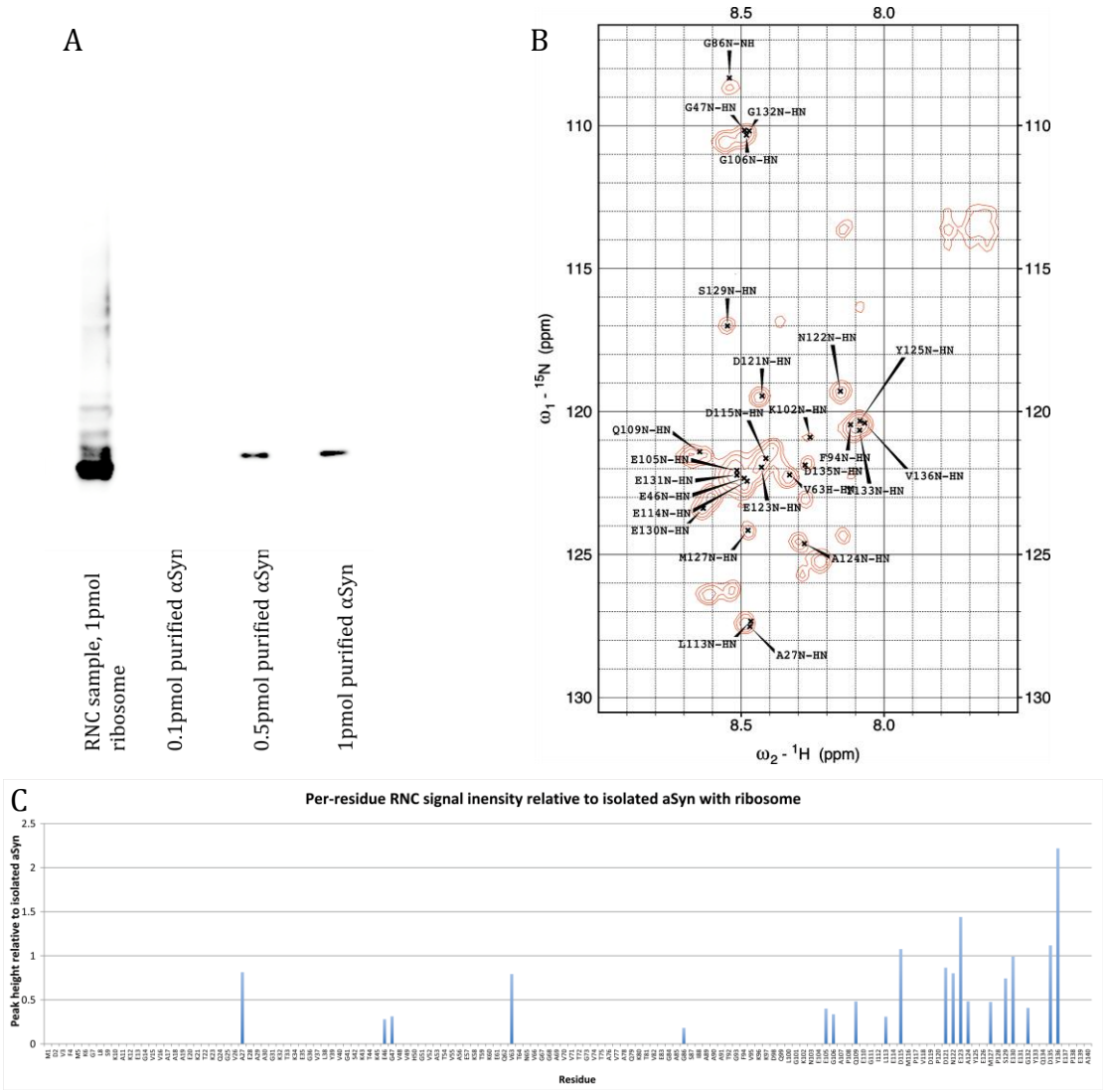

Figure 3.15. Properties of the first αSyn-RNC NMR sample. A) Anti-αSyn western blot of the RNC sample (1pmol of ribosome loaded) showing multiple bands, the most intense of which runs faster than isolated αSyn and therefore cannot be the full-length αSyn-RNC construct. B) The ^{1}H-^{15}N spectrum of the αSyn-RNC sample with the peaks that overlay with isolated αSyn labelled. Some peaks – including most of those below residue 100, namely A27, E46, G47, V63 and F94 – overlap and so cannot be clearly distinguished. The only peak below residue 100 that does not overlap is that of G86, which does not overlay well with that of the reference spectrum: Δδav = 0.104ppm, using a weighted Pythagorean relation (Frenkel and Smit, 2002) C) For those peaks that overlay with isolated αSyn, per-residue peak intensity relative to isolated αSyn with ribosome, normalised for concentration and number of scans.

Western blot analysis of the sample suggested that multiple species were present with the αSyn C-terminal tail bound by the anti-αSyn antibody (Figure 3.15A). Furthermore, all the peaks that could be confidently assigned without overlapping other peaks resided in the C-terminal tail (Figure 3.15B and C). It can therefore be concluded that significant proteolysis has occurred within the sample.

3.4 Conclusion

Outlined above are the successful preparation of αSyn-RNCs by both *in vitro* and *in vivo* methods, as well as methodological enhancements that were necessary for the detection and quantification of samples.

As stated, a sample required for NMR study would require approximately 300µl at 5-10µM concentration. Furthermore, the instability of ribosomes in concentrations above this concentration requires that almost all ribosomes present in the sample must be occupied by a nascent chain lest the signal be too week (if under-occupied) or the ribosomes prematurely degrade (if over-concentrated).

In vitro preparation of ^{15}N-labeled αSyn-RNCs was performed, with positive results (Figure 3.5). A number of factors to improve stability and occupancy of samples were explored, including plasmid quality (Figure 3.2) reaction time (Figure 3.3), presence of protease inhibitors (Figure 3.4), labelled amino acid source (Figure 3.6) and sucrose cushioning parameters (Figure 3.9).

With these advances in preparation protocol, it was calculated that the necessary materials to provide NMR-scale samples could indeed be supplied using the Roche Diagnostics Rapid Translation System RTS5000 cell-free protein biosynthesis kits. By modulation of the amino acid source, it is possible to precisely control the labelling of amino acids within the resultant RNC. It would therefore follow that it is possible to use selective ^{13}C labelling to further screen the NMR data to shown only specific residues of interest. The cost of this larger cell-free kit, however, was prohibitive for the studies undertaken here.

By adapting existing *in vivo* RNC sample preparation strategies, this work has demonstrated the ability to prepare NMR-scale samples of ^{15}N-labelled αSyn-

RNC constructs. Furthermore, preliminary investigations by NMR spectroscopy suggested that the αSyn-RNC is present, but that it had undergone sample degradation, leading to proteolysis by-products, and contamination by ^{15}N-labelled ribosomal proteins (Figure 3.14). While these imperfections render the initial sample unsuitable for further study, they pose problems for which potential solutions are evident, such as the addition of protease inhibitors and siRNA to inhibit ribosomal protein translation during RNC expression.

4 Development of techniques for Molecular Simulations of Ribosome-Nascent Chain Complexes

4.1 Introduction

As described in section 1.5.2, the use of chemical shifts to produce refined structures of proteins requires the molecular dynamic or Monte Carlo simulation of all the atoms in the relevant molecular system. Existing techniques cannot be applied directly to RNC systems for several reasons, including the fact that the all-atom simulation of a ribosome would be unfeasibly slow due the enormous number of atoms present – there are more than 200,000 atoms in an RNC system, compared to the approximately 2,000 in a typical CS-MD system. It is therefore necessary to define a protocol whereby the all-atom dynamics of a nascent chain may be feasibly simulated in the presence of a ribosome.

Moreover, the ribosome also presents a novel challenge to the preparation of starting structures for simulations, as the structure must thread the exit tunnel of the ribosome while still exhibiting various characteristics beyond the exit port, such as a linearised structure, a randomised structure or a pre-folded structure.

This chapter describes solutions to these and other problems associated with performing CS-MD of RNCs such that the resulting calculations are rendered feasible.

The aim of this work was to identify simulation parameters that would render feasible the CS-MD simulation of RNCs. This would require the ability to perform all-atom unrestrained MD of RNCs. Moreover, the ability to generate suitable starting structures for different modes of study would be required. This includes model-free structure generation – ie generating structures that do not require *a priori* knowledge of the structure – as well as defining RNC structures based on known folding pathways of the isolated protein.

4.2 Results

4.2.1 Preparation of initial RNC structural models

Before simulations of an RNC can be performed, it is prudent to identify known structural properties of the nascent chain. For example, from high-resolution structure data of the prokaryotic ribosome, it is know the exit tunnel is not clearly defined, and more than one path can be traced by the nascent chain (Seidelt et al., 2009). Furthermore, as detailed in the Literature Review there are two Universal Adapter Sites (UAS1 and UAS2) that allow multiple translation-coupled interactions with cofactors about which structural information may be required. In particular, the Trigger Factor is known to bind to protein L23 and limited high-resolution data exists for this interaction (Ruth, 1983). To form the starting structure for RNC and RNC-TF interactions, a model was built by combining multiple experimentally-determined structures to produce a high-resolution ribosome-nascent chain-trigger factor complex (RNTC) of as high accuracy as possible.

A likely nascent chain structure – produced by cryoelectron microscopy-restrained MD – is available in the Protein DataBank (ID 2WWQ) for a poly-A TnaC leader peptide (Seidelt et al., 2009). This structure includes only a small proportion of the surrounding ribosome, so a high resolution prokaryotic ribosome structure that includes as much of the ribosome as possible was found – PDB ID 2J00 and 2J01 (Selmer et al., 2006), which also includes co-ordinates for parts of the P-site and E-site tRNAs – and superimposed over the ribosomal region in 2WWQ using the PyMol 'super' command (Press et al., 2007). All but the nascent chain data derived from structure 2WWQ were then hidden, providing a structure that should a likely path for a nascent chain without forming a steric clash with the ribosome (Figure 4.1 B and C).

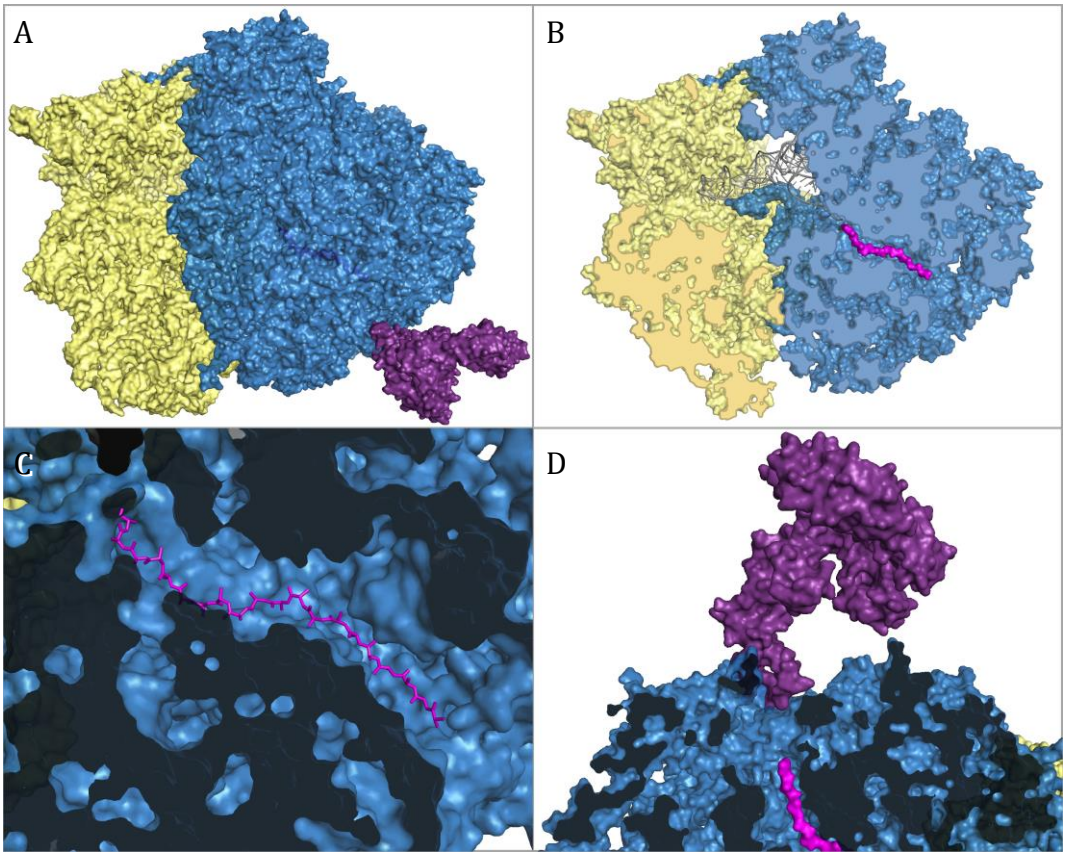

Figure 4.1. Models of a ribosome-nascent chain-trigger factor complex (RNTC). A) Solvent accessible surface of the RNTC model (see text for construction). The small subunit is yellow and the large subunit blue. The Trigger factor is shown in purple and is bound to protein L23 at the exit port of the ribosome. B) Cross-section through the RNC, with the nascent chain shown as a magenta solvent-accessible surface. The P-site tRNA is shown as a grey cartoon representation. C) Detail of exit tunnel cross-section, with the model shown in the same orientation as B. The nascent chain is shown in stick form. D) Cross section of the RNTC showing the exit port of the large subunit with the nascent chain emerging from the tunnel below into a cavity formed by the ribosome and the trigger factor. .

To add a docked model of the Trigger Factor, two x-ray crystal structures were used: that of a high-resolution Trigger Factor homodimer, 1W26, and of the Trigger Factor ribosome-binding domain bound to ribosomal protein L23, 1W2B (Ruth, 1983). The latter structure was first superimposed over the RNC structure using just the residues of L23 present in 1W2B as a template from the RNC model. A single Trigger Factor model – extracted from the homodimer structure – was then superimposed over the ribosome binding domain portion present in 1W2B. The bridging structure was then hidden, leaving a single ribosome, a

single nascent chain and a single Trigger Factor, as seen in Figure 4.1A. In agreement with previous observations, the Trigger Factor was found to form a cradle over the exit port of the ribosome (Figure 4.1D).

Two GelFac-RNC models were produced: $GelFac_{646\text{-}750}$-RNC and $GelFac_{646\text{-}761}$-RNC. Each construct comprises and N-terminal histidine tag (His_6-AS), the denoted residues of Gelation Factor and C-terminal stalling sequence and associated construct-derived residues (TSEF-SecM). For both of these structures, the folded domain 5 structure (residues 646-750) was used. To this, the histidine tag was manually modelled in and relaxed by steepest-descents energy minimisation followed by 20ns implicit solvent stochastic dynamics in the absence of the ribosome. For the $GelFac_{646\text{-}750}$-RNC, the final 8-residue G strand of the folded domain was removed and manually modelled to connect to the exit-tunnel backbone co-ordinates described above. The side chains of residues 743-750, followed by TSEF-SecM were placed on the exit-tunnel structure. For $GelFac_{646\text{-}761}$-RNC, the side chains of residues 754-761, TSEF-SecM were used for the exit-tunnel structure. Residues 751-753 were manually modelled to connect the domain 5 structure to the exit-tunnel structure.

4.2.2 All-atom molecular dynamics simulations of RNC systems

We used the following methods in order to perform molecular dynamics simulations of RNC systems, both by themselves and in the presence of chemical shift restraints. Initially, simulations were performed using the 50S ribosomal subunit molecules from PDB 2J01 (Selmer et al., 2006). As explained in the previous report, the dynamics of the ribosome were assumed to be not critically important for studies of extra-ribosomal nascent chain dynamical behaviour. Therefore, the atoms of the ribosome structure were held in fixed postions by use of the “freezegrps” simulation parameter while the dynamics of the nascent chain were simulated. Since in the approach interatomic energies within the ribosome are not required, they were excluded from the calculations by use of the “energygrp_excl” parameter. The generalised Born/solvent-accessible surface area (GB/SA) implicit solvent model was used in place of explicit solvent molecules to reduce the number of particles in the system. However, despite these measures, there was still so much calculation unnecessary required for

fixed atoms, that the sampling rate of the simulations was prohibitively slow. Further means to increase the sampling rate were therefore required.

4.2.2.1 Adjusting the Simulation Parameters

It is possible, at least in principle, to increase the sampling rate of some simulations by changing the temperature coupling time constant (the exponential decay time constant for temperature deviation), τ_T, and the Brownian dynamics friction coefficient (the motion dampening coefficient applied as part of the stochastic dynamics), ξ_{BD}. Both parameters affect dynamic characteristics without affecting its Hamiltonian. Varying these two parameters can therefore affect the timescales of motions without affecting the energy associated with the system. Their effects were investigated by monitoring the root mean square fluctuation (RMSF, a measure of how much movement occurs) of a three-alanine peptide in the same force field and implicit solvent as was to be used for the RNC systems. As seen in Figure 4.2, despite using values over 5 orders of magnitude, no significant increase in RMSF was observed. Over the range of values for τ_T and ξ_{BD} employed, no significant improvement in sampling was observed.

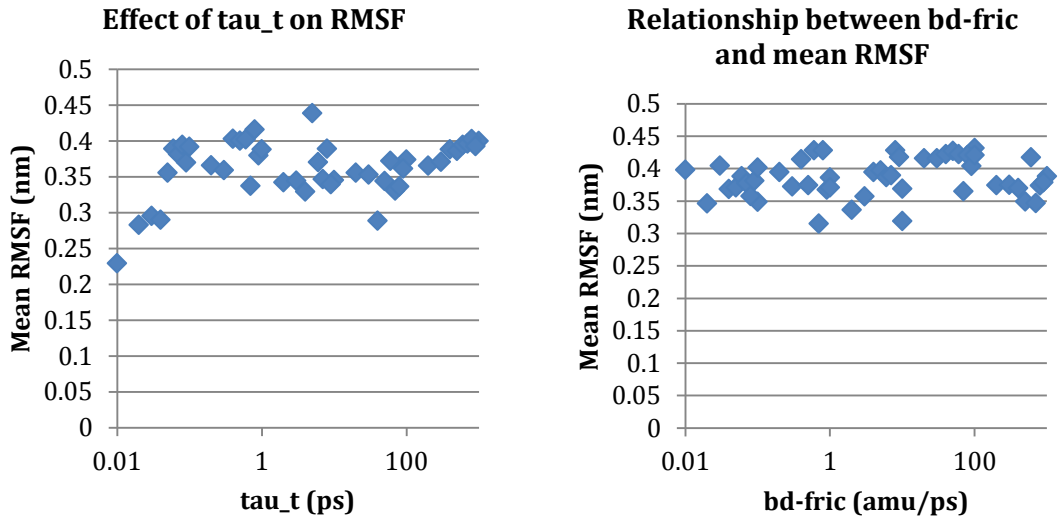

Figure 4.2. Effects on the RMS fluctuations (RMSF) of a poly-alanine structure as a function of the temperature coupling time constant, τ_T, and Brownian dynamics frictional coefficient, ξ_{BD}. High RMSF values indicate greater motion and therefore greater sampling within the simulations.

4.2.3 Approaches to removal of unnecessary atoms

A straightforward method to increase the sampling rate of the complex system is to remove some of its atoms. The accurate identification of the atoms that can be removed without significantly affecting the properties of the system, however, is often difficult. Initially, whole proteins of the 50S subunit not interacting with most nascent chains were removed. As shown in Figure 4.3A, of the approximately 168,000 atoms present in the starting structure, about 150,000 remained after this approach. There were therefore still far too many atoms present in the system, and the structure may not be suitable for other nascent chains with greater extension.

A second approach that we considered was to manually remove atoms that seemed unlikely to interact with a given nascent chain (Figure 4.3B). This left far fewer atoms (approximately 30,000), but was highly error-prone, as it is impossible to ensure that nascent-chain-accessible cavities are not inadvertently removed, as was later discovered to be the case.

Finally, using existing tools it is possible to identify residues with solvent-(water) accessible atoms for retention (Figure 4.3C). This reliably includes all

atoms that may interact with the nascent chain but also includes many that are in pockets accessible to water molecules but not to the nascent chain, meaning that a prohibitively large number of atoms is retained.

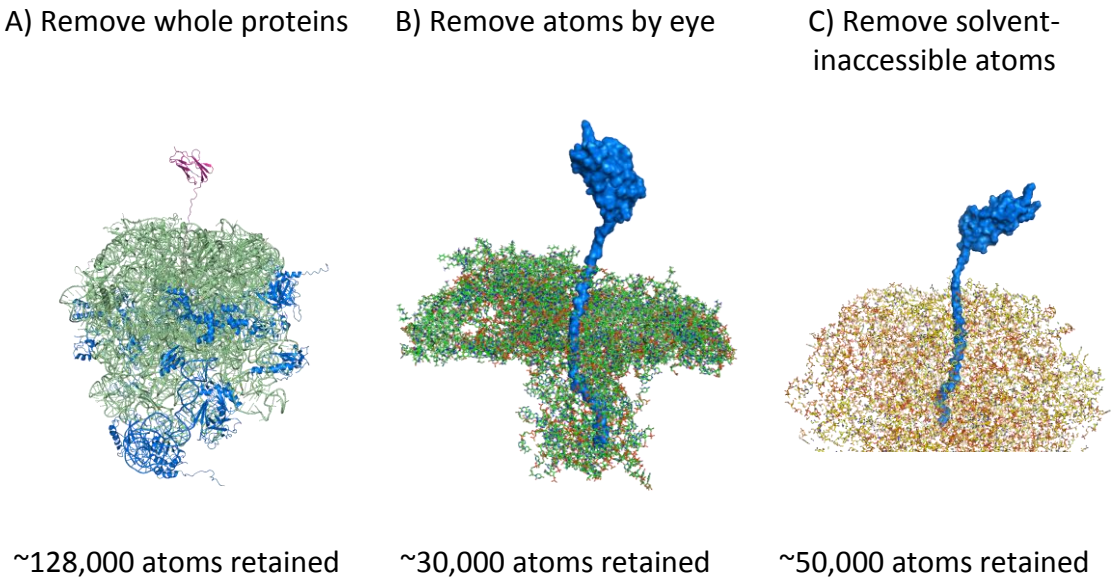

Figure 4.3. Illustrations of three possible approaches for removing unnecessary atoms from the 50S subunit structure. Each illustration is shown with a representative nascent chain and the number of 50S subunit atoms left in the structure displayed. A) The large subunit is shown as green cartoon with the whole proteins that do not interact with the nascent chain shown in blue. B) The ribosome (green and red ball-and-stick) is visualised in PyMol and atoms that look unlikely to interact with the nascent chain (blue solvent-accessible surface). C) Using PyMol's solvent-accessible surface calculation, residues with atoms that interact with solvent (yellow and red ball-and-stick) are selected and other atoms removed.

To address these problems, a novel nascent-chain-accessible distance calculation algorithm was developed inspired by the Level-Set Molecular Surface (LSMS) algorithm (Can et al., 2006). The new algorithm, illustrated in Figure 4.4 and termed Probe Accessible Distance Calculator (PADC), is organised as follows.

1. Based on a desired resolution, an optimized three-dimensional discrete grid is fit over the co-ordinate space of the structure.

2. Any grid-cell ("voxel") in which the co-ordinates of an atom reside is marked as "occupied" by the corresponding atom.

3. Search around each occupied voxel for neighbouring voxels whose centre co-ordinate is within the van der Waal's radius of the starting voxel's atom.

4. Starting from the external boundary of the matrix, mark voxels as being occupied by the probe. Search in all directions from each starting voxel to see if each voxel within the probe's radius is unoccupied by an atom. If so, mark each searched voxel as occupied by the probe and begin the search again from there. Repeat this process until no more voxels can be occupied by the probe. A probe with the diameter of an amino acid residue, 3.35Å, is used as this mimics the potentially-explored space of the nascent chain. This gives a three-dimensional grid space that can be explored by a nascent chain.

5. Starting from the PTC, the probe-accessible voxels are explored and marked with the shortest distance through probe-accessible voxels from the PTC. When a voxel occupied by an atom is encountered, the atom is assigned the shortest probe-accessible distance encountered for any atom of that residue.

6. All residues with one or more probe-accessible atoms are now assigned their minimum distance from the PTC. All residues with no distance assigned cannot make contact with the nascent chain and can be discarded. All atoms with a greater distance than the length or extension of the nascent chain can be discarded.

The extension of a nascent chain is the maximum distance that can – or is likely to – be reached by the nascent chain during simulation. For simulation of a nascent chain with a stable folded domain, this would be the distance from the C-terminal residue to the furthest point in the structure with the domain folded and unfolded region fully extended.

The distances measured by PADC are not entirely accurate: distances are calculated as the sum of distances between neighbouring (adjacent and diagonal) probe-accessible voxels; and it is possible that nascent chain side chain branches

may reach into smaller cavities than the probe size. However, use of a high enough resolution grid reduces measurement error, and retention of entire residues reduces chance of significant atoms being removed to near zero.

PADC allows determination of the length of nascent chain required to reach different atoms of the ribosome to adjust the atoms kept according to the reach of the nascent chain under investigation, such that the number of atoms retained is as low as possible, ranging from approximately 4,000 atoms for a nascent chain that covers just the exit tunnel, to approximately 36,000 atom for a nascent chain that could reach the 30S subunit.

The implementation of PADC used, in the Perl scripting language, is slow for distance measurement, taking approximately 12 hours on a single processor. However, after calculation of distances for a particular ribosome structure, grid resolution and probe diameter, structures for any desired maximum nascent chain extension can be produced in less than 30 seconds on the same processor.

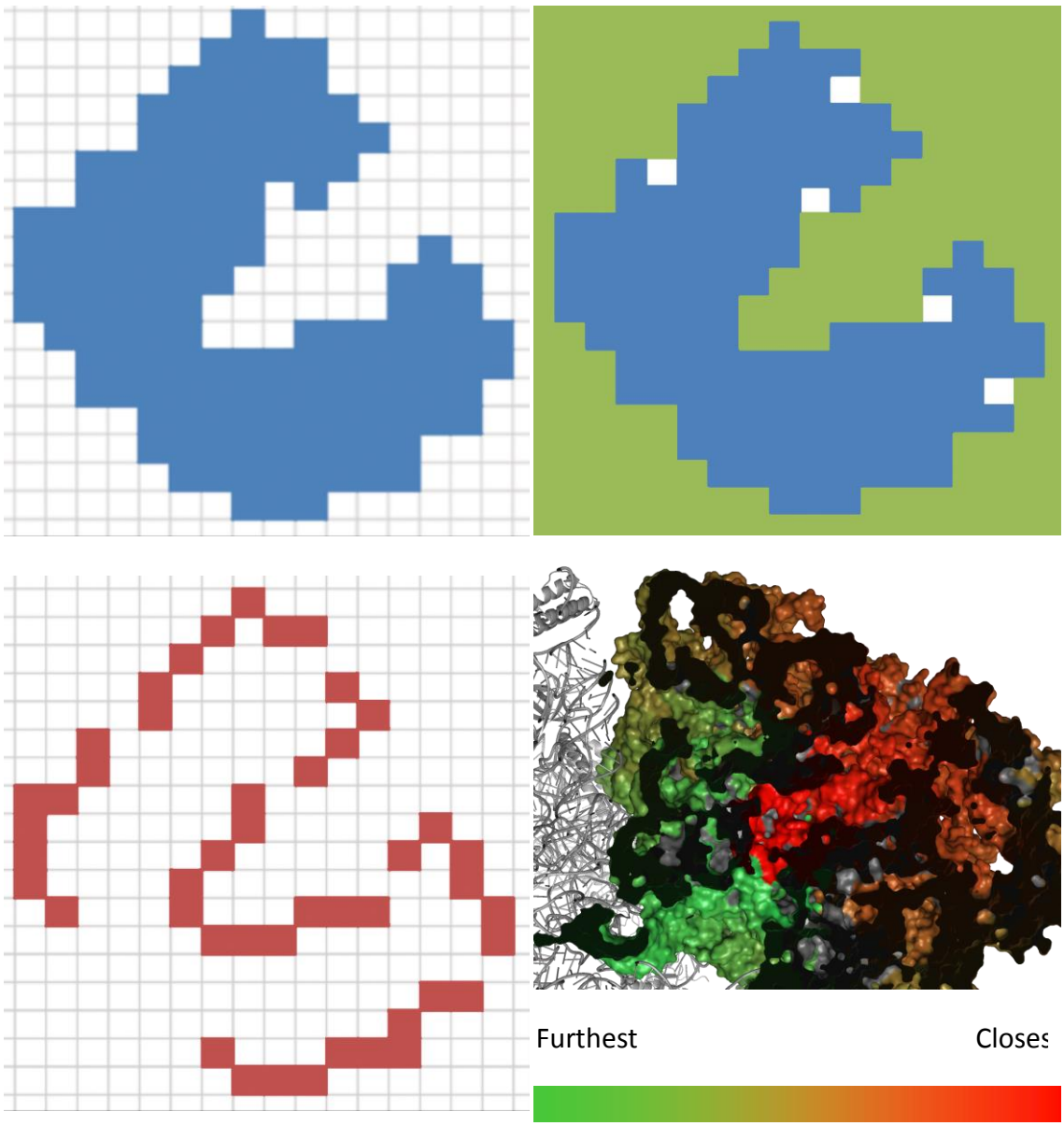

Figure 4.4. Illustration of the PADC atom-stripping algorithm. Atom coordinates are first aligned into a three dimensional grid (here represented in two dimensions) and voxels containing atoms marked as atom-occupied, shown as blue squares. Voxels that could be occupied by a probe sphere are then marked as probe-accessible (green squares). Starting from the PTC, the distance is measured to all other probe-accessible regions of the grid and all atom-occupied voxels (red squares) encountered are marked with the minimum probe-accessible distance from the PTC. Finally, all residues containing atoms within a chosen distance are noted, and all other atoms removed from the structure. The distance from the PTC is shown on the solvent-accessible surface of a cross-section through the 50S subunit of the ribosome. The exit tunnel is seen as a red channel through the subunit (being closest to the PTC) while the P-site is seen as green, as this is the longest distance a nascent chain would have to traverse to be reached, if not occluded by the 30S subunit.

Number of cores	RAM per core (GB)	Structure trimming	Use of energy exclusion groups	Simulation time (fs)	Duration	Simulation rate (ps/day)
256	2	Proteins	Yes	10000	5:08:07	46.74
256	2	No	Yes	10000	5:33:35	43.17
256	2	Proteins	Yes	100	0:01:56	74.48
256	2	No	Yes	100	0:03:31	40.95
32	16	No	Yes	100	0:24:09	5.96
32	16	No	No	100	0:27:46	5.19
2	2	PADC	Yes	434830	12:00:00	869.66
4	2	PADC	Yes	184930	12:00:00	369.86
8	2	PADC	Yes	167930	12:00:00	335.86

Table 4.1. Effect of different simulation parameters on the sampling rates, measured in picoseconds of simulation time per day of computation. In each simulation, the system includes a SecM nascent chain and the 50S subunit whole ("no" trimming), with some proteins removed ("proteins"), or with the probe accessible distance calculator used to remove some atoms ("PADC"). In all simulations, the ribosome atoms were held in fixed positions.

4.2.4 Coarse-grained simulations of RNC systems

In parallel to the all-atom chemical-shift restrained RNC molecular dynamics simulations described in Chapter 6, we also performed coarse-grain unrestrained simulations (set up according to the methods developed by Dr Edward O'Brien, University of Cambridge (Case, 1995)). These simulations were carried out using a two-centre model (in which each residue is represented as two spheres; one for the C_α and the other for the sidechain), with no electrostatic interactions, no attractive vdW interactions between the nascent chain and ribosome and all ribosome-nascent chain attraction mediated via a single Lennard-Jones potential of varying potential well depth, ε, for different simulations. The resulting ensembles of between 10,000 and 128,000 structures

were made available for further analysis. RNCs of both αSyn and Gelation Factor$_{646\text{-}839}$ and Gelation Factor$_{646\text{-}770}$ were produced.

Initially, appropriate behaviour of the nascent chain within the exit tunnel was verified by comparing the extension of this portion of the nascent chain to the cryo-EM structure. As shown in Figure 4.5, even the most extend ensemble had a mean extension approximately 96% of the length of the cryo-em structure. The difference is, however, well within the 5.8Å resolution of the published structure, which was also solved at much lower temperature than the energy at which the simulation was conducted.

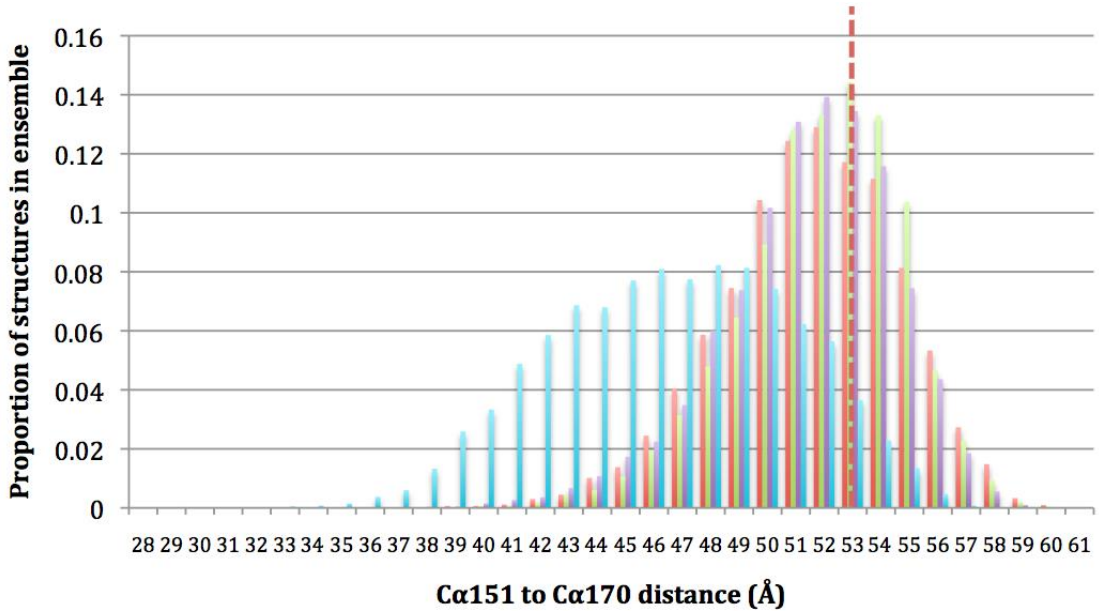

Figure 4.5. Distribution of nascent-chain extensions in different coarse-grain ensembles. Extension, measured as the distance between the C-terminal C_α atoms, C_α151 and C_α170 is plotted against the proportion of structures with that extension length in coarse grain ensembles with Lennard-Jones potentials of ε=0.1 (blue, mean 46.0Å), 0.04 (orange, mean 50.9Å), 0.03 (green, mean 51.4Å) and 0.001 (purple, mean 51.1Å). The extension observed in the published cryo-electronmicroscopy structure, 53.3Å, is marked as a dotted red line.

4.2.4.1 Analysis of the dynamical behaviour of RNCs shows unexpected NMR properties

To investigate the visibility of nascent chain residues by NMR spectroscopy, the root mean square fluctuation (RMSF) – the standard deviation of atomic

positions distances from the ensemble average – was calculated for the most-extended (ε=0.001) αSyn ensemble. Experimentally, the last residue observable by NMR for the αSyn-RNC is GLU137, which is 25 residues (3 subsequent αSyn residues, 6 vector-derived residues and 16 SecM stalling residues before the stalled P-site proline) away from the PTC. It was expected that, being NMR-visible, this residue would tumble more freely than the subsequent residues in the sequence. As seen in Figure 4.6, however, a small increase in motion is seen for this residue, but greatest increase in mobility is seen for residues prior to GLU130.

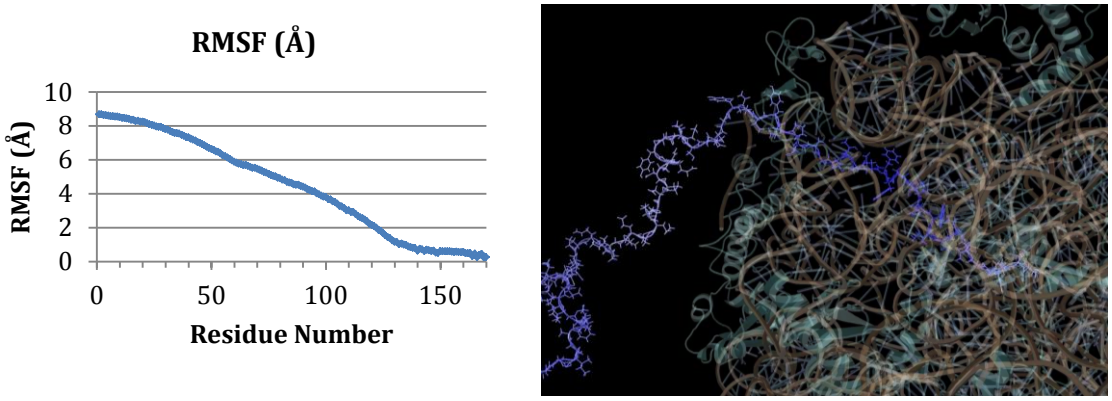

Figure 4.6. RMSF as a function of the residue number of the nascent chain. The last 40 residues of the nascent chain show minimal increase in motion with increasing distance from the PTC. While residue 140 is outside the ribosome, it is not until reside 130 that motion begins to increase significantly. This is illustrated by a representative structure in which the nascent chain colour is assigned according to its RMSF value at each residue. It can be seen that many residues beyond the exit port are assigned the same colour as those within the exit port.

This result can be explained in one of five ways: 1) The nascent chain is not as extended within the CG ensemble as is the case experimentally; 2) The CG ensemble is not sampling a sufficient conformational space to exhibit the appropriate RMSF; 3) The threshold RMSF value to determine NMR-visibility is low and discrete; 4) Some other process not replicated in the CG ensemble is preventing downstream residues of GLU137 from being observed, such as interaction with the ribosome; 5) RMSF is not a suitable proxy for NMR visibility.

Of these explanations, 1 and 2 are unlikely as the CG ensemble was produced from a high temperature, low interaction, long duration simulation, allowing

maximal extension of the nascent chain and sampling of all feasible conformations. Given that the changes in RMSF between NMR-visible and non-NMR-visible residues is very small, explanation 3 is similarly unlikely. Explanation 5 is worthy of exploration as RMSF measures only translational movement of residues and not tumbling of residues, which is the determinant of NMR visibility. An alternative approach would therefore be to measure the order parameter (see section 2.6) of residues, which more closely measures an NMR-relevant parameter.

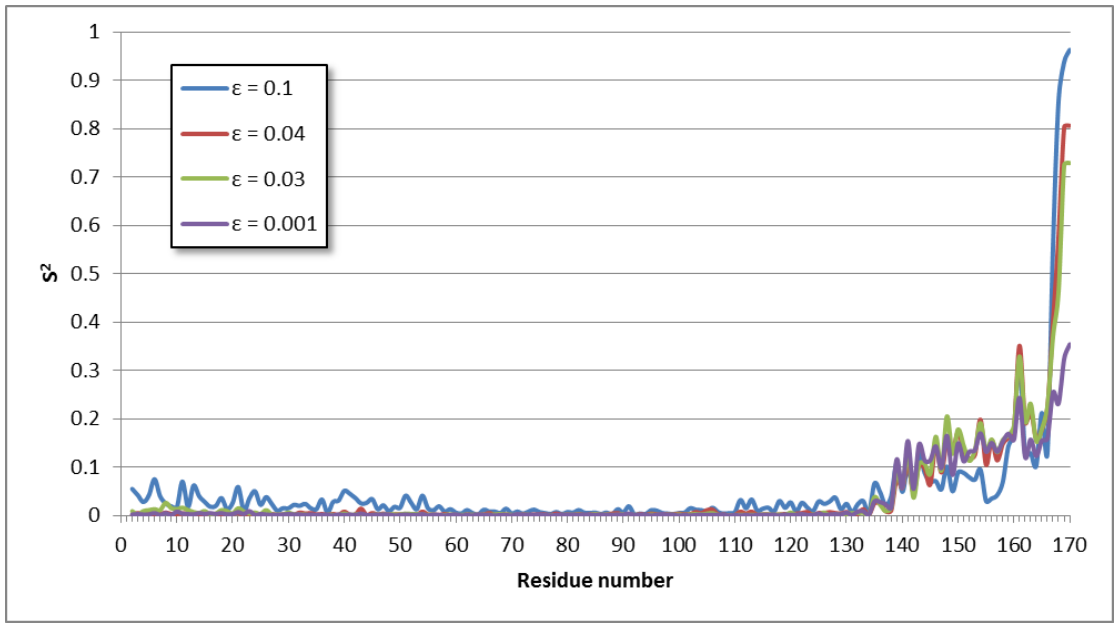

Figure 4.7 Per-residue C_α-C_α S^2 order parameter values across four αSyn CG MD systems with different ε values.

As seen in Figure 4.7, the C_α-C_α S^2 values were calculated for each of the ensembles as described in section 2.6. Note that C_α-C_α values are used as the CG simulation does not include other backbone atoms, and replacing missing atoms would incorporate biases intrinsic to the molecular completion protocol. The values are almost entirely low – suggesting almost complete disorder – which is as would be expected for a high-temperature simulation, as the system would adopt a large range of conformations than would otherwise be feasible. As such, interpretations reliant on the absolute S^2 value are inapplicable. Of note, however, is the clearly higher S^2 value for residues approximately 136 to 170, which corresponds to those residues that would be expected to be within the exit

tunnel of the ribosome. The high S^2 value for the residues closest to the PTC is likely due to the highly confined sterically-accessible space for the amino acid residues to sample. Further from the PTC, across residues 136 to 165, despite the high temperatures leading to an almost entirely-disordered system, the confines of the exit tunnel clearly impose an impediment to local motion.

4.3 Conclusion

In section 4.2.1, I outlined the complexities associated with preparing all-atom models of GelFac RNCs as suitable starting structures for MD simulation, based on known structural properties of the isolated protein. Moreover, this section demonstrates the ability to incorporate the Trigger Factor folding cofactor within the model based on known docking and folding. The core principle involved with generation of the RNC – using the cryo-EM observed backbone co-ordinates of a nascent chain (Seidelt et al., 2009) with sidechain replacement, placed within a high-resolution ribosome structure (Selmer et al., 2006), followed by modelling of the trans-exit tunnel region and MD energy minimisation – can be re-used for any RNC system of interest. This leads to a high-resolution, sterically sound, empirically-derived model of RNCs for any nascent chain of interest.

In section 4.2.2, it is demonstrated that adjustment of commonly-used simulation parameters (τ_T and ξ_{BD}) to increase the sampling rate of the RNC structure do not impart any perceptible improvement on the range of conformations of the nascent chain. The number of atoms present in the system compared to the number of atoms in the region of interest (the nascent chain) is so great as to render this approach ineffective.

Instead, based on prior publications indicating that the dynamics of the ribosome are immaterial to the conformational biases of the nascent chain (O'Brien et al., 2010), a strategy involving the freezing of ribosomal atoms means that atoms that do not interact with the nascent chain can be removed. Three obvious routes to the removal of such atoms – removal of whole ribosomal proteins, removal of atoms by eye and removal of solvent-inaccessible atoms – are shown to either remove too few atoms, too many atoms or both in section

To solve the problem of identifying ribosomal atoms that cannot interact with the nascent chain, a new algorithm has been developed called the Probe Accessible Distance Calculator (PADC). This discretises the three-dimensional co-ordinate space of an RNC system to any selected resolution, and uses a three-stage search process to 1) identify ribosome-occupied voxels, 2) identify those voxels that are accessible to a probe of any given radius and 3) calculate the shortest distance through accessible voxels to every probe-accessible ribosome-occupied voxel. These distances are then mapped back to the ribosomal atoms by which they are occupied.

The PADC has been successfully used to cull unwanted ribosomal atoms from an RNC system with the effect of increasing simulation rate by an order of magnitude.

Using high temperature, coarse grained αSyn-RNC Lennard-Jones simulation trajectories supplied by Edward O'Brien (Cambridge University), it has been shown how the full extent of accessed structures observed under these conditions can be used to predict the range of motion exhibited by such an RNC. In particular, regardless of choice of the potential well depth, as seen in Figure 4.7, the model-free order parameter S^2 shows a clear boundary between those amino acid residues that exhibit unhindered motion beyond the ribosomal exit tunnel, and those whose motion is impinged by the ribosome. This suggests that even relatively simple CG simulations of the RNC may predict nuclei that are observable by NMR, thereby assisting with the design of RNC constructs.

The prior sections of this chapter demonstrate that the principle aim of this work – rendering feasible the CS-MD of RNCs – has been achieved.

5 NMR-Based Comparison of the Free Energy Landscapes of Four Truncated Forms of Gelation Factor Domain 5

5.1 Introduction

As described in section 1.6.1, Gelation factor provides a suitable model for the study of co-translational folding of multiple immunoglobulin-like domains. Towards that end, one approach to studying its co-translational behaviour is by synthesising isolated proteins derived from one of the domains, with progressive C-terminal truncations. This allows the study of stability and conformational preferences of the domain as additional amino acids are available to be incorporated into the folded protein.

Since removal of C-terminal residues is expected to result in reduced stability of the domain, the solution-state properties of the constructs are of particular interest, including the energetic favourability for the native fold and whether intermediates are favoured when the full domain is not present.

Techniques that may provide information relevant to such study include biochemical characterization, NMR linewidth analysis, chemical shifts, residual dipolar couplings and hydrodynamic radius measurements. Data have been recorded using such techniques, but my use of appropriate computational techniques (as described in section 1.5) it may be possible to derive atomic-resolution models of the proteins and high-resolution information about the solution-state ensembles of each construct.

In this chapter application of molecular dynamics simulations – including the use of chemical shift data as restraints – to truncations of domain 5 of Gelation factor are described.

5.1.1 Gelation Factor constructs under investigation

Many truncation constructs of Gelation factor domain 5 have been prepared and studied (Maria-Evangelia Karyadi, University College London), of which

appropriate NMR data were available for three truncations. These three constructs (Table 5.1) comprised the entire domain 5 ($GelFac_{644-750}$), the domain with most of the final beta strand missing ($GelFac_{646-744}$) and the domain with half of the final beta strand missing ($GelFac_{646-746}$).

Name	Composition
$GelFac_{644-750}$	Gelation Factor residues 644 to 750
$GelFac_{646-746}$	HHHHHHAS- Gelation Factor residues 646 to 746
$GelFac_{646-744}$	HHHHHHAS- Gelation Factor residues 646 to 744

Table 5.1 Constructs under investigation

Dataset	Description	Available shifts (%)					
		C_α	C_β	C′	HN	N	H_α
$GelFac_{644-750}$	Native	98.1	87.9	93.5	90.7	90.7	88.8
$GelFac_{646-746}$	Native-like	88.1	63.4	91.1	84.2	84.2	0.0
$GelFac_{646-744}$ F2	Native-like	66.7	54.5	63.6	54.5	55.6	0.0
$GelFac_{646-744}$ F1	Less native-like	83.8	67.7	80.8	71.7	71.7	0.0
$GelFac_{646-744}$ U	Unfolded	93.9	82.8	93.9	86.9	86.9	0.0

Table 5.2 Datasets employed. For each dataset, the percentage of the residues of Gelation Factor for which the specified chemical shifts are available is listed.

The C_α, C_β, C′, H_α, HN and N chemical shifts for Gelation Factor residues 644-750, both native and denatured in urea have already been published (Hsu et al., 2009a). Each truncation has an N-terminal histidine tag (His_6-AS-). The shorter of these truncations $GelFac_{646-744}$ has three datasets: one set of peaks close to the native full domain chemical shift ('F2'), one set of folded peaks that differ more significantly from the native state ('F1') and one set of peaks corresponding to an unfolded species ('U'). The F1 and F2 datasets were recorded simultaneously from the same sample, suggesting two states in slow exchange.

5.1.2 Aim of the Study

This work aims to demonstrate that chemical shift restraints from an analogue for the RNC system (in which additional residues of a domain are progressively available for folding) can be used to restrain an MD system. To that end, it aims to demonstrate that the chemical shifts are altering the equilibrium conformational ensemble. Furthermore, it aims to demonstrate that the subtle differences in chemical shifts arising from the presence of different numbers of amino acid residues at the C-terminal tail of Gelation Factor domain 5 have an appreciable effect on the equilibrium ensemble of CS-MD simulations. Ultimately, it is intended that this information will assist with elucidating the conformational biases of Gelation Factor domain 5 as it emerges from the ribosome.

5.2 Methods

Chemical shift data were provided as NMR-STAR-formatted datasets, which were validated by a comparison with the published $GelFac_{644\text{-}750}$ to identify typographical errors and frame shifting. To assist with this, a Perl module (see section 2.8) for parsing and error checking NMR-STAR formatted files was written. By comparing the data to that of the (published) native state chemical shifts and producing a heat map of difference from the published data, it became reliable and efficient to spot and fix frame-shift errors that occur.

The starting structure for $GelFac_{644\text{-}750}$ was obtained by extracting the corresponding residues from a published crystal structure (PDB 1QFH) (McCoy et al., 1999). This structure was minimized using a steepest-descents algorithm for approximately 120 steps followed by 20ns unrestrained simulation using the Amber99SB*-ILDN force field, Generalized Born Solvent Accessible Surface Area implicit solvent and a stochastic dynamics integrator.

The starting structure for His_6-AS-$GelFac_{646\text{-}746}$ was obtained by removing two N-terminal and 4 C-terminal residues from the GelFac644-750 starting structure. The additional N-terminal residues were manually modelled by extending the N-terminal maintaining the same peptide bond angles as residue 646. This structure was minimized using the same steepest descents and unrestrained SD as for GelFac644-750.

The starting structure for His6-AS-GelFac646-744 was obtained the same as for His6-AS-GelFac646-746 except that two additional C-terminal residues were removed at the first step.

We applied to each structure the CamShift-restrained Molecular Dynamics (CS-MD) method with the same Amber99SB*-ILDN force field and implicit solvent in Gromacs. Two replicas of each construct were produced from the same starting structure but different initial velocities. The system underwent simulated annealing cycling between 300K and 450K for 100ps each with 100ps for each temperature change.

To monitor convergence, the RMSΔδ was calculated (see below) as an indicator of whether any improvement of chemical shift differences was occurring. Each production CS-MD simulation was repeated without restraints for comparison. All other simulation parameters were kept the same.

5.3 Results

5.3.1 Chemical shift restraints are biasing the trajectories of the folded domains towards ensemble structures that have chemical shifts closer to the experimental values

As a means to monitor equilibration within the chemical shift restraints, the RMSΔδ over time was calculated for each trajectory. This procedure entails using CamShift (or any other chemical shift predictor) to predict the chemical shifts of each residue for each replica at intervals throughout the trajectories. The per-shift, per-residue, per-replica difference in chemical shift (Δδ) between each structure and the experimental data is then calculated and subsequently the mean per-shift, per-residue Δδ across all replicas. Finally, the root mean square (RMS) value across each shift at each time point is calculated and plotted against the time frame for each value. This method replicates the ensemble-averaging of shifts exhibited both by the recording of NMR data and in the restraint application.

The initial ensembles of the restrained simulations showed large RMSΔδ values, which dropped rapidly within the first few picoseconds of simulation; for

comparison, these large RMSΔδ values are not reduced in the unrestrained simulations. Each of the simulations shows significantly lower RMSΔδ values for all shifts in the restrained trajectories than in the unrestrained trajectories (Figure 5.1). Although the RMSΔδ were initially considered to have reached their equilibrium values within the first 10ns of restrained simulation, it became apparent that the mean RMSΔδ were still decreasing in the subsequent 100ns of CS-MD. It can therefore be concluded that the restraints are exerting biasing the trajectory of the ensembles towards structures that match the experimental chemical shifts used as restraints. Further investigations are under way to determine whether with greater time using the same parameters, the RMSΔδ will drop further.

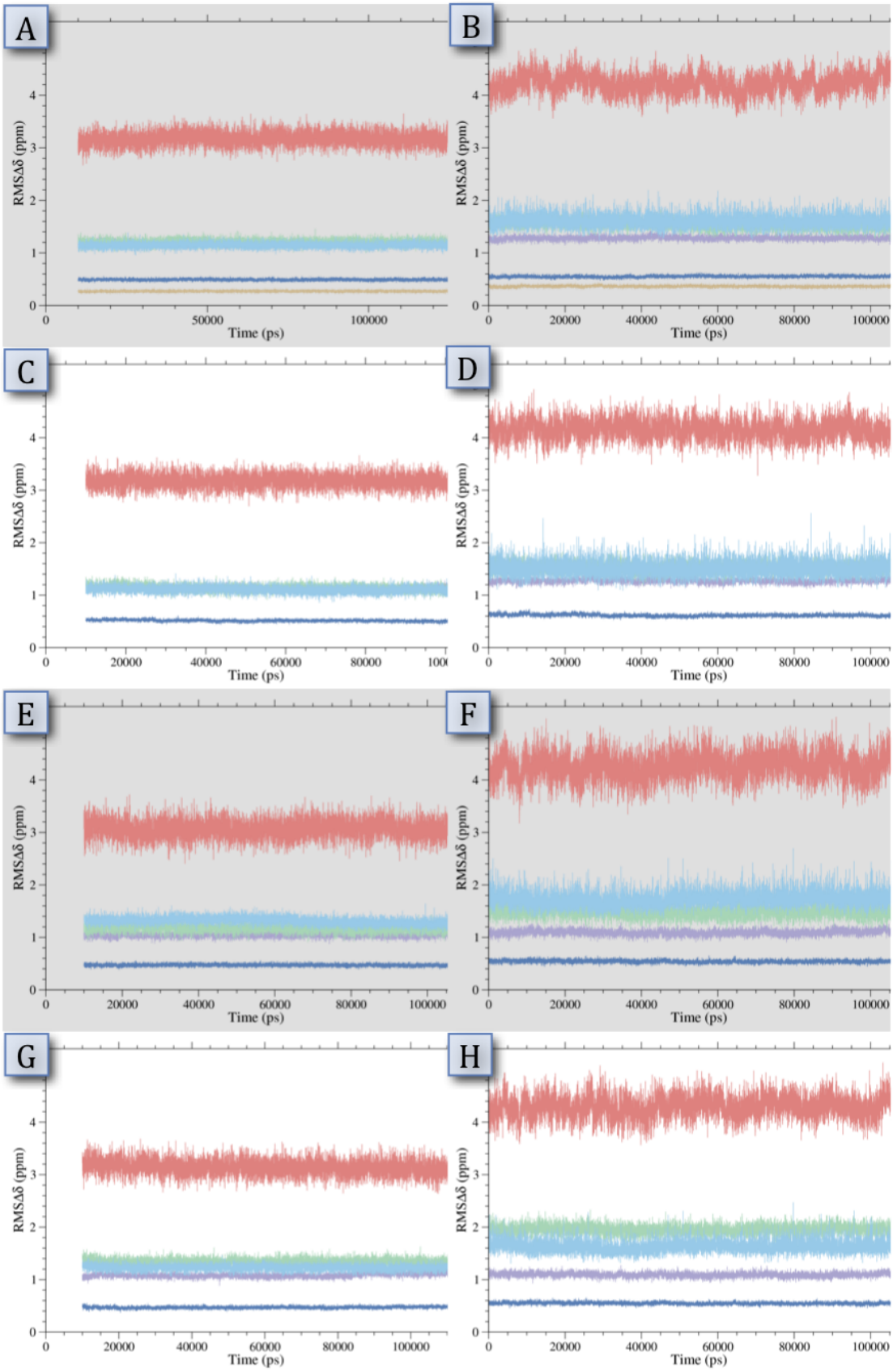

Figure 5.1. Time series of the RMSΔδ values for isolated Gelation Factor domain 5 constructs during the ensemble-chemical-shift-restrained- (A, C, E, G) and

unrestrained- (B, D, F, H) ensemble molecular dynamics simulations. A,B: GelFac$_{644-750}$; C,D: GelFac$_{646-746}$; E,F: GelFac$_{646-744}$ F2; G,H: GelFac$_{646-744}$ F1. Each line denotes the RMSΔδ for a different shift: red, N; green, C$_\alpha$; cyan, C$_\beta$; violet, C′; navy, HN. Note that one trajectory for the unrestrained GelFac$_{646-744}$ was computed for comparison to the F2 and F1 datasets.

As a means to assess the ability of the chemical shifts to exert a longer-range, global bias on the structures, 10% of the available chemical shifts for the GelFac$_{646-744}$ datasets were randomly selected and not used as restraints to form a test dataset. The per-residue RMSΔδ across the equilibrium ensemble (from 10 to 110ns) of each dataset was calculated and the values from the test dataset compared to the values used as restraints. Our results indicate that the RMSΔδ for the test dataset chemical shifts do not differ from those used as restraints (Figure 5.2). These findings show that the restraints have long range effects on the simulated structures that are not confined to single nuclei or residues for which restraints are applied. These effects have beneficial ramifications for simulations in which datasets exhibit clusters of residues with a paucity of data, such as residues 664-666 of the F1 dataset, which lack any chemical shift information but are flanked by residues with near-complete backbone chemical shifts.

Figure 5.2. Per-residue RMSΔδ across the restrained CS-MD of Gelation Factor domain 5 constructs. Red bars indicate data that were not used in the restraint datasets. Grey bars indicate data for which no experimental values are available. A) $GelFac_{644-744}$ F1; B) $GelFac_{644-744}$ F2.

5.3.2 Free energy landscapes of the different constructs suggest the possibility of multiple distinct populations

To identify significant populations within the restrained equilibrium ensembles, free energy landscapes (FELs) were plotted based on properties expected to be informative of the disrupted domain 5 fold. Such properties included number of β-sheet contacts, R_g, N-shift RMSΔδ and RMSD from starting structure. Of these, the most informative was the pairing of RMSΔδ and RMSD (Figure 5.3).

The FEL of $GelFac_{644\text{-}750}$ shows a single population, consistent with the expectation that this state should exhibit a single stable structure. By contrast, all the truncated states show more than one population, with at least one population having a lower RMSD than that observed for $GelFac_{644\text{-}750}$. This result may be related to the additional amino-terminal tag present in these constructs. The tag may lower the RMSD for those structures because, lacking CS restraints, it is retained in the position selected by the forcefield alone, just as during the equilibration phase prior to application of the chemical shift restraints. A more informative comparison to the $GelFac_{644\text{-}750}$ ensemble could be drawn by excluding the tag residues from the RMSD calculation.

The $GelFac_{646\text{-}744}$ F1 system contains three minima at increasing RMSD from the starting structure, consistent with the knowledge that the F1 dataset shows greater deviation from the full-domain than the F2 dataset. Indeed, the population with greatest RMSD is also the population with lowest RMSΔδ, suggesting that this population may the one favoured by the chemical shift restraints, but the lower-RMSD population may be the one favoured by the force field alone.

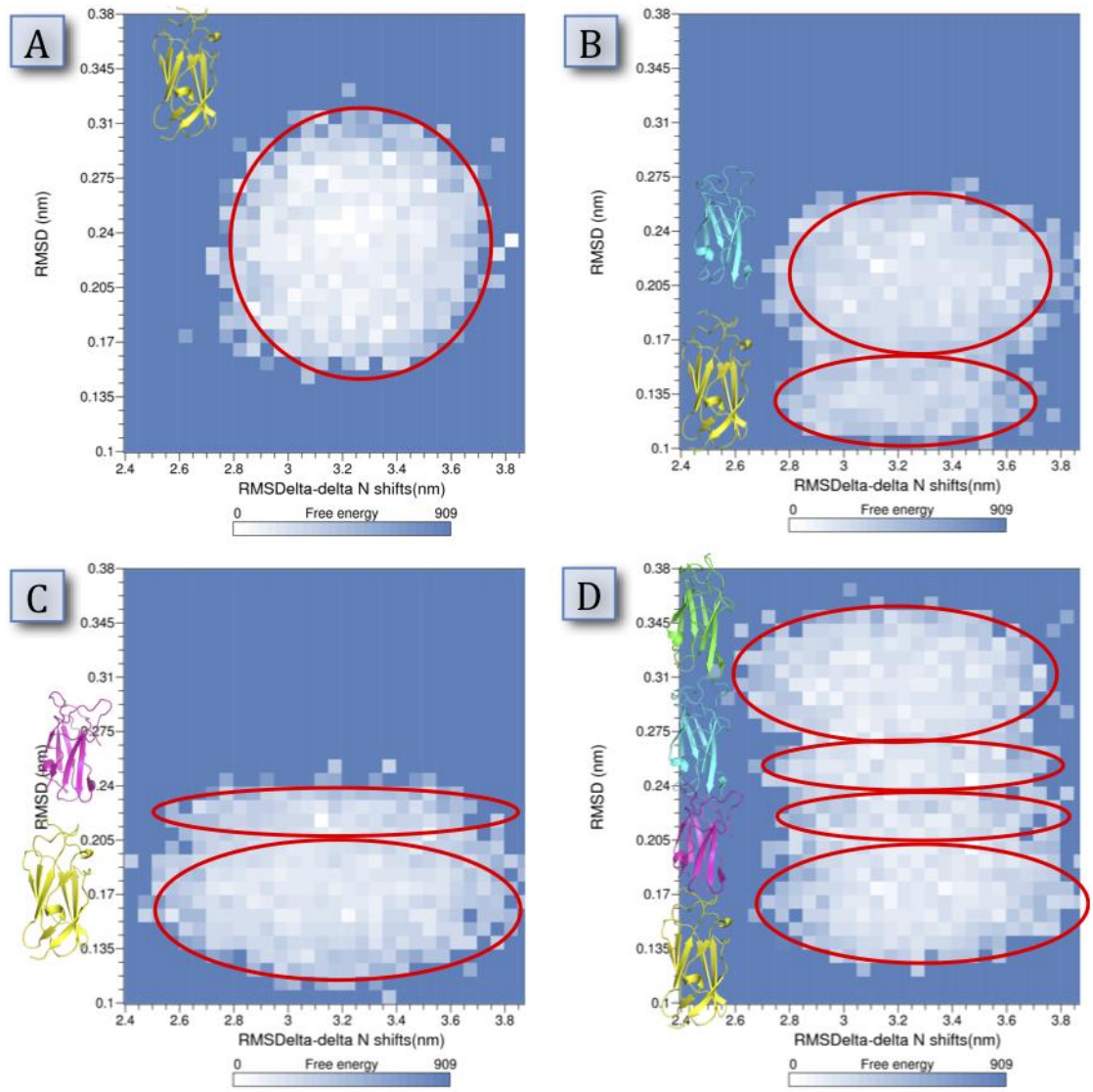

Figure 5.3. RMSΔδ-RMSD free energy landscapes of the portion of each CS-MD simulation considered to be at equilibrium, ie from 10ns onwards. A, GelFac$_{644-750}$; B, GelFac$_{646-746}$; C, GelFac$_{646-744}$ F2; D, GelFac$_{646-744}$ F1. Red lines denote the approximate perimeter of the different populations present within each ensemble adjacent to cartoon representations of the average structure of each population.

5.3.3 The unfolded state is not yet equilibrated within the timescale of the simulations

As demonstrated previously, the RMSΔδ values over time suggest that the various constructs remained relatively rigid during the course of the simulations. This was also the case for restrained simulations using the dataset of unfolded peaks for the GelFac$_{646-744}$ construct. However, this system was expected to be

significantly disrupted so as to unfold, but the small difference in RMSD from the starting structures observed (data not shown) indicated that the perturbation in the ensemble dynamics imposed by the use of chemical shift restraints is minimal under the conditions so far employed. Considering the highly diverse structures expected to be present in the ensemble adopted by an unfolded or disordered system, it is unlikely that this technique alone would sample sufficient configurational space over such a short simulation time. It may be possible, however, to discern structural properties about this state from backbone chemical shifts alone sufficiently to quantify secondary structural states (Vila et al., 2009). Furthermore, it may be possible to sample the FEL of the chemical shift restrained ensemble by use advanced sampling techniques, such as metadynamics, in which previously sampled states within specified degrees of freedom are disfavoured allowing sampling of a greater range in a shorter time (Barducci et al., 2011).

Furthermore, it is possible to monitor the perturbation of the systems over time by monitoring changes in the RMSD, as in Figure 5.4. These show steady increases in disruption to the structure over time in both $GelFac_{646\text{-}744}$ systems that never return to previous lower values. This suggests that these systems may not be fully equilibrated within the restrained forcefield and that further simulation with the same conditions, or re-simulation with higher energies, may allow sufficient disruption to the systems to allow equilibration.

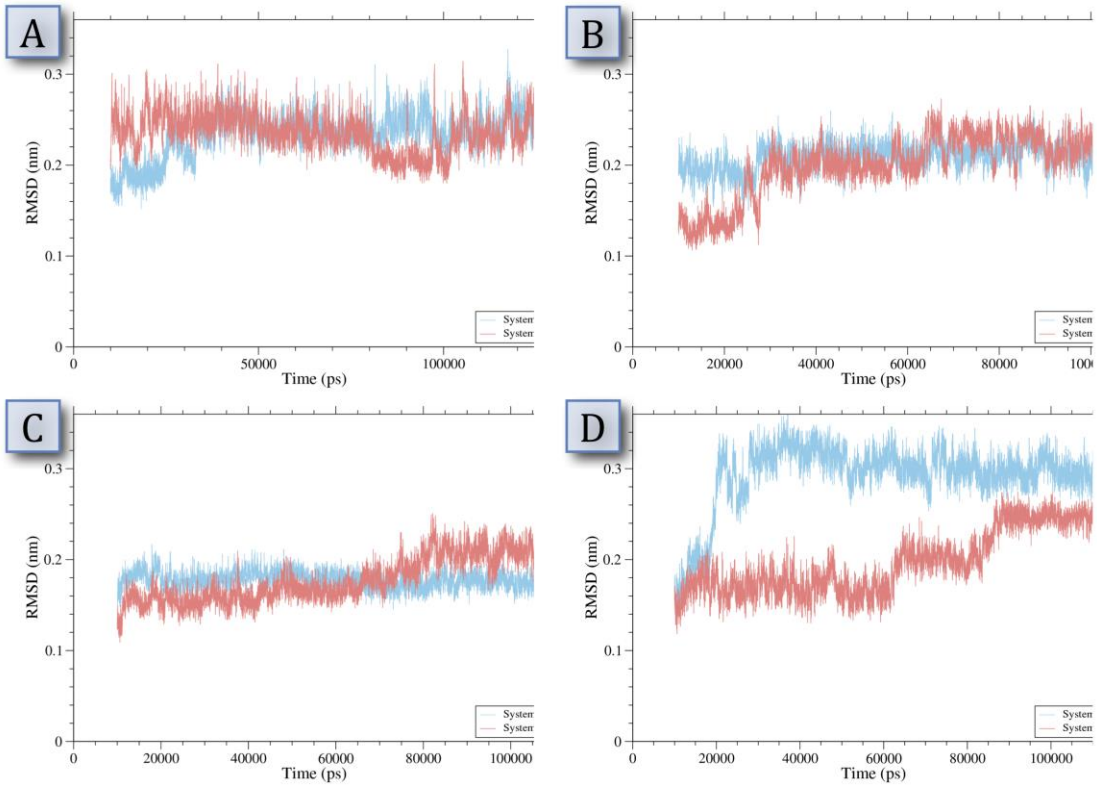

Figure 5.4. RMSD of structures under CS-MD versus the starting structure. Each simulation comprises two replicas ("systems") undergoing combined ensemble-restrained CS-MD. A, GelFac$_{644-750}$; B, GelFac$_{646-746}$; C, GelFac$_{646-744}$ F2; D, GelFac$_{646-744}$ F1.

5.3.4 The chemical shift restraints stabilise structure and recover inter-element order

The S^2 order parameters were calculated following the protocol described in section 2.6 for the backbone N-HN pairs across the period from 20ns onwards in each trajectory and compared to comparable simulations in which the chemical shifts were absent. As can be seen in Figure 5.5, both CSMD and MD simulations show substantial secondary-structure-like S^2 values of 0.8 or above.

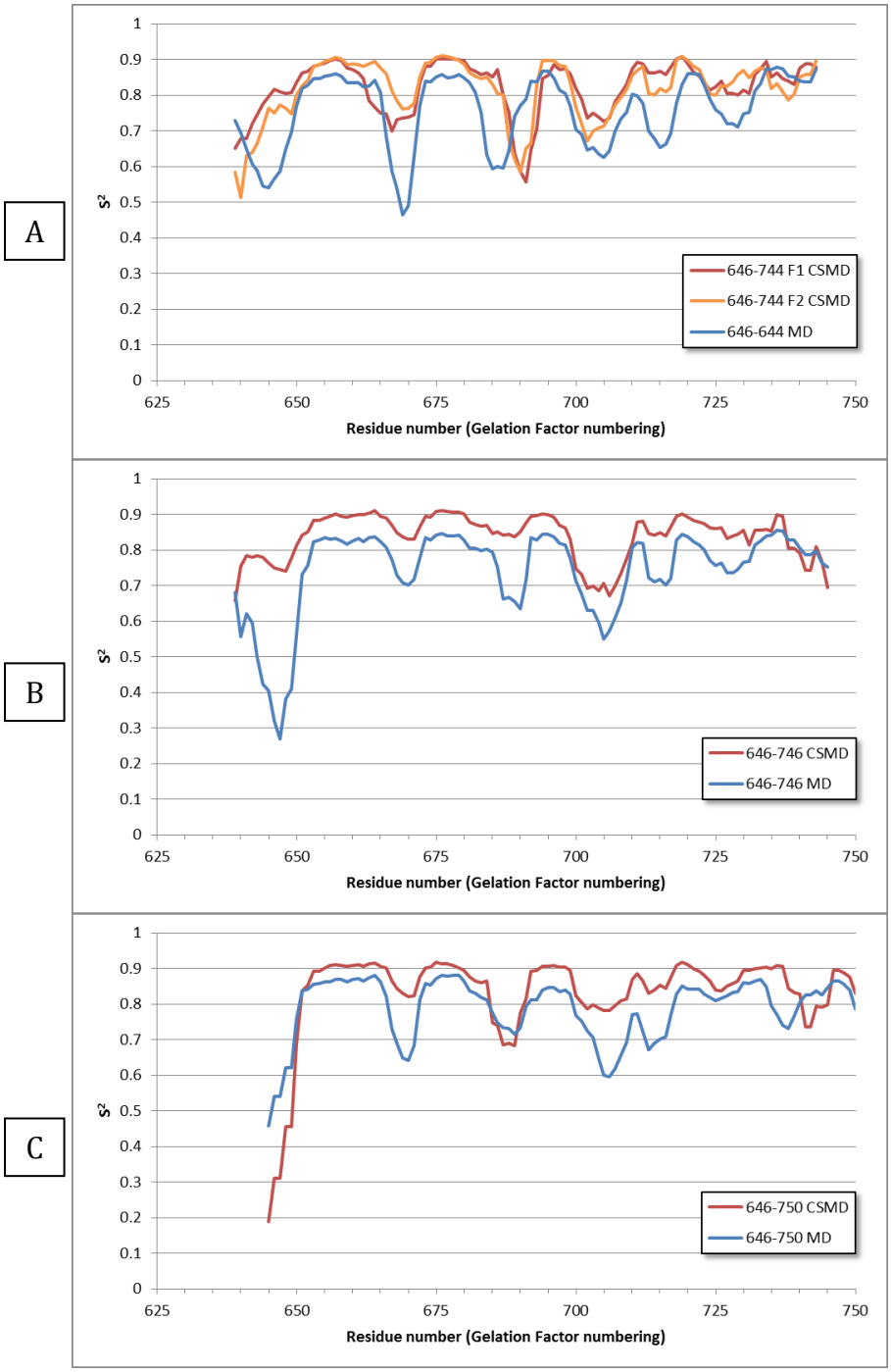

Figure 5.5 Per-residue S^2 order parameter for N-HN pairs CSMD and MD simulations of the three Gelation Factor truncation constructs. In each figure, the blue trace shows the S^2 value for MD simulations while the red/orange traces show that of the chemical-shift restrained equivalent. A) The CSMD trajectories for the two folded chemical shift datasets for the $GelFac_{646-744}$ construct, F1 and F2 are shown in red and orange respectively. B) The trajectory restrained by the $GelFac_{646-746}$ construct dataset is shown in red. C) The trajectory restrained by the $GelFac_{646-750}$ construct dataset is shown in red.

Evident from Figure 5.5 is that the CSMD trajectory exhibits greater order overall order than the MD equivalent, suggesting that a more-rigid, folded state is favoured in each case. Moreover, the nature and location of the troughs shows interesting properties. In each case, a trough can be seen at around residues 665 to 670, which corresponds to a turn-bend region between β sheets. While the MD trajectories each exhibit a pronounced dip in order, indicating that the MD force field does not suggest significant structural propensity, the chemical shift-restrained trajectories show only a minor dip that suggests retention of order. A similar pattern is observed around residues 715 to 720, which corresponds to another β sheet region; again the CSMD is accurately predicting a secondary structural element that MD alone did not. Another observation is that in the systems lacking the complete final sheet ($GelFac_{646-744}$ and $GelFac_{646-646}$), the MD-alone trajectories suggest loss of order in the prior sheet from residues 725-735. The CSMD trajectories, however, retain order in this region, suggesting that the chemical shifts indicate structure is retained in this region when the force field alone would predict otherwise.

5.3.5 Structural properties of the folded states are not yet equilibrated in the trajectory

As an independent means to assess the validity of the restrained ensembles, the hydrodynamic (Stokes) radius (R_H) was collected for the experimental samples.

The radius of gyration (R_g), which is proportional to R_H, suggested that even the most expanded structures in the ensembles fell short of the experimental value. For confirmation, the R_H was calculated for the most expanded structure in each ensemble using HydroPro under the conditions used for the experimental data (temperature 283K, solvent viscosity 1.307×10^{-3} Pa). As shown in Table 5.3, the R_H values from the restrained ensembles fall significantly below the experimental values. The restrained ensembles are therefore not adequately reproducing the state of the structures observed by NMR.

Construct	Experimental R_H (nm)	Peak CS-MD R_H (nm)
$GelFac_{644-750}$	2.20	1.918
$GelFac_{646-746}$	2.20	1.944
$GelFac_{646-744}$ F2	2.45	1.888
$GelFac_{646-744}$ F1	2.40	1.914

Table 5.3. Comparison of experimental hydrodynamic radii (R_H) versus peak R_H in the restrained MD ensemble.

However, that the R_H values are *lower* than expected suggests that higher-temperature MD, or better force fields, will produce more accurate ensembles. This is in agreement with the findings from the RMSΔδ values that suggests the MD ensembles are in a stable conformation that requires greater energy – either as higher temperature or greater restraint weighting – to sufficiently perturb the starting structures.

5.3.6 CHESHIRE unable to resolve precise structural differences

A potential concern with the current approach is that, by using the crystal structure as starting point, CS-MD may only give structures that are similar to the crystal structure and that do not necessarily represent the partially-folded states that the chemical shifts are expected to produce. It is therefore desirable that unbiased starting structures be used either in tandem (as an orthogonal validation) or as a substitute for the crystal-structure starting point.

Attempting to re-fold the structure from an entirely disordered state would likely be error-prone and time consuming. The CHESHIRE (chemical shift restraints) pipeline, however, allows *ab initio* protein structure determination by use of molecular fragment replacement in conjunction with 3PRED secondary structure prediction from chemical shifts and TOPOS database searching for fragments with similar sequence, secondary structure and chemical shifts.

The CHESHIRE pipeline involves production of a large number of low-resolution structures, selection and refinement of the best-scoring structures,

characterisation of the refined ensemble and, finally, selection of a very small number of the best structures for use either as a starting point for CS-MD, or for comparison to the ensembles produced by CS-MD. The CHESHIRE pipeline was attempted with the chemical shift data available and was unable to produce structures within 3Å backbone co-ordinate RMSD and so would be unable to identify precise local structural differences between the relatively similar chemical shift datasets.

5.4 Conclusion

To monitor the effect of chemical shift restraints on MD trajectories, a measurement has been developed – called RMSΔδ – that assesses the difference between the predicted chemical shift associated with each nuclei for structures within a simulation at any given time point, and the experimentally-derived chemical shifts that are being used as restraints. While signal from this metric is noisy – small local variations can affect the values within a range of ±0.2 ppm in for backbone nitrogen nuclei, for example – trends in this value – or window averages over 1-10ns can provide useful information as to whether the system is converging or diverging from the restraint values.

By monitoring the RMSΔδ for the range of systems and restraints available for differing Gelation Factor domain 5 truncations, it is clear in Figure 5.1 that the restraints are having a significant effect on the trajectory of the simulation, so the CS-MD is successfully restraining the simulation to conformations that more accurately reproduce the chemical shifts observed experimentally. While the simulations were initially observed to demonstrate a rapid decrease in RMSΔδ over the pico- to nano-second timescale during equilibration (data not shown), it only became evident during analysis of the production ensemble that a second, much slower equilibration event was still occurring over the nano- to micro-second timescale. It would therefore be of considerable value to repeat these simulations over longer timescales to obtain more accurate representations of the equilibrium properties of this system.

One concern about the systems for which data were available for Gelation Factor was that the considerable number of absent chemical shifts in the dataset would

prevent global structural changes from occurring in the system. To address this, the CD-MD simulations were conducted with a random 10% of the available shifts were excluded from the restraint dataset. As can be seen in Figure 5.2, the nuclei with absent shifts did not have RMSΔδ values significantly greater than those for which restraints were employed. Some nuclei that were present in a group of adjacent nuclei with chemical shifts unavailable did have a higher RMSΔδ than those for which restraints were employed, but they were not so much higher that they would cause concern. This information does suggest that regions with multiple adjacent nuclei for which restraints are unavailable may have local structures that are less representative of the empirically-observed conformations, but that the global structure would still be accurate.

Across the production ensemble of these truncation systems, Free Energy Landscapes of RMSΔδ versus RMSD demonstrated multiple distinct sub populations. IN particular, in concordance with the theory that the more truncated systems would be less stable (due to the lack of the final G-strand to provide hydrogen bonding to stabilise the structure), those systems with greater truncation exhibited a greater range of sub populations (Figure 5.3).

Per-residue analysis of the S^2 model free order parameter suggests that the unrestrained MD successfully identified stable secondary structural regions (Figure 5.5), although less stably so than the CS-MD systems. However, in loop regions between secondary structural elements, the unrestrained MD predicted significant disorder, while the CS-MD retained fairly ordered conformations. This suggests that the CS-MD is restraining these less-ordered in ranges of conformations that still have significant biases, presumably more like those adopted experimentally.

The $GelFac_{646\text{-}750}$ CS-MD system demonstrates a highly-ordered C-terminal tail, suggesting that it has adopted a stable, native-like fold. The $GelFac_{646\text{-}746}$ system, however, demonstrates a significant decrease in order at the C-terminal tail from residues 740 onwards – more so than the unrestrained MD – indicating that the lack of the final 4 residues confers a significant disruption to the stability of the G-strand. The $GelFac_{646\text{-}744}$ system, however, exhibits stability at its final residue,

suggesting that without an unstable ultimate G-strand present, the penultimate strand retains a stable fold.

Attempts to use the current state-of-the-art structure-from-chemical shifts CHESHIRE pipeline failed to reproduce native-like conformations, which is likely partly a consequence of the dynamic conformations expected from this system. CS-MD was clearly found to be a far more accurate way to reproduce structural information from chemical shifts.

While further work to equilibrate the CS-MD GelFac systems would greatly assist with inferences and predictions about the structural propensities of the constructs, the work here has made multiple predictions about the effect of sequential truncation of the C-terminal strand of Gelation Factor, based on successful application of chemical shift restraints. Furthermore, based on the protocols developed here, additional details could be derived about further truncations based on NMR chemical shifts alone with relatively little human time to that required for this first foray, and substantially less compared to comparable studies by crystallographic studies.

6 Chemical Shift Restrained Molecular Simulations of Gelation Factor Ribosome-Nascent Chain Complexes

6.1 Introduction

As previously described in section 1.2, the ability to understand the structural details of protein folding on the ribosome will provide a step toward a detailed understanding of the *in vivo* folding process.

In Chapter 3, computational techniques for the use of NMR chemical shift data as restraints in molecular dynamics simulations were outlined that rendered feasible the simulation the large systems involved in ribosome nascent chain complexes.

This chapter will describe the application of these methodological strategies to simulations of Gelation Factor derived nascent chains using experimentally-recorded NMR chemical shift data as a mimetic of a multi-domain protein during translation.

6.1.1 Systems investigated

As previously described in section 1.6.1, Gelation Factor provides a useful model for study of co-translational folding by NMR spectroscopy partly due to its sequence coding for multiple distinct Ig-like domains. Furthermore, NMR chemical shift data have been recorded and assigned for RNCs derived from domains 5 and 6 of Gelation Factor (Christensen et al., 2011). As depicted in Figure 6.1, of particular interest is a construct in which a whole domain is emerged from the ribosome but with a short linker such that the domain is able to fold but is not afforded significant freedom of motion relative to the ribosome. This construct may provide information about the folding competence of domains that are emerged from the ribosome but still attached. Also of interest would be a construct in which a whole domain is not fully emerged so that the properties of the nascent chain while not completely present can be investigated, such as whether folding intermediates are observed.

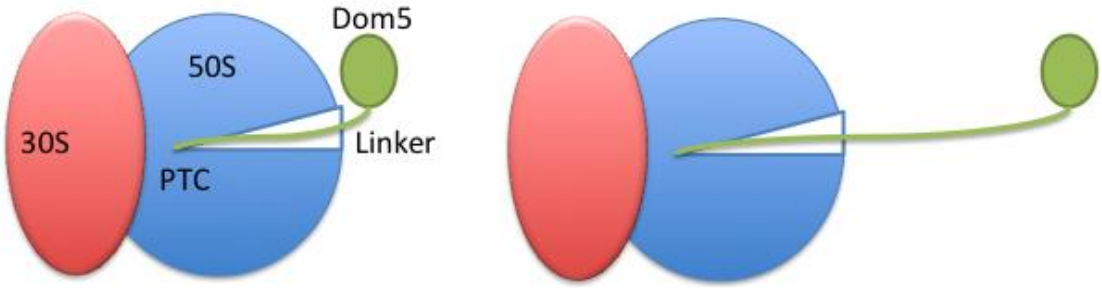

Figure 6.1 Schematic outlining two RNC systems of interest for study of cotranslation folding. On the left, a complete domain has emerged from the ribosomal exit tunnel but is bound to the PTC a linker amino acid sequence that is under stalled translation. On the right, the same domain is bound to the PTC by a much longer linker sequence. This linker contains the residues for another domain but is not fully emerged from the exit tunnel and so is not able to adopt a fully native fold.

To this end, two Gelation Factor constructs were chosen for RNC CS-MD. The first, denoted L47, consists of residues 646 to 750 of Gelation Factor with 47 downstream (C-terminal) residues connecting it to the PTC. The linker consists of 21 residues for the stalling sequence and vector-derived residues, and residues 751 to 770 of Gelation Factor. Based on the assumption that approximately 30 residues reside within the ribosome exit tunnel, this leaves just domain 5 and the interdomain linker between domains 5 and 6 of Gelation Factor external to the ribosome.

A second construct, termed L110, consists of residues 646 to 839 of Gelation Factor and the same 21 residues of stalling sequence and vector-derived residues as L47. This comprises the full sequence of domains 5 and 6 as well as the interdomain sequence. However, as the final 30 residues are expected to residue within the exit tunnel, at least the last 9 residues of domain 6 will be sterically inaccessible to the external nascent chain and thus unable to fold.

As seen in Figure 6.2, the L47 and L110 systems adequately demonstrate two of the states of interest for RNC systems and therefore provide suitable starting systems for the investigation of RNCs by CS-MD.

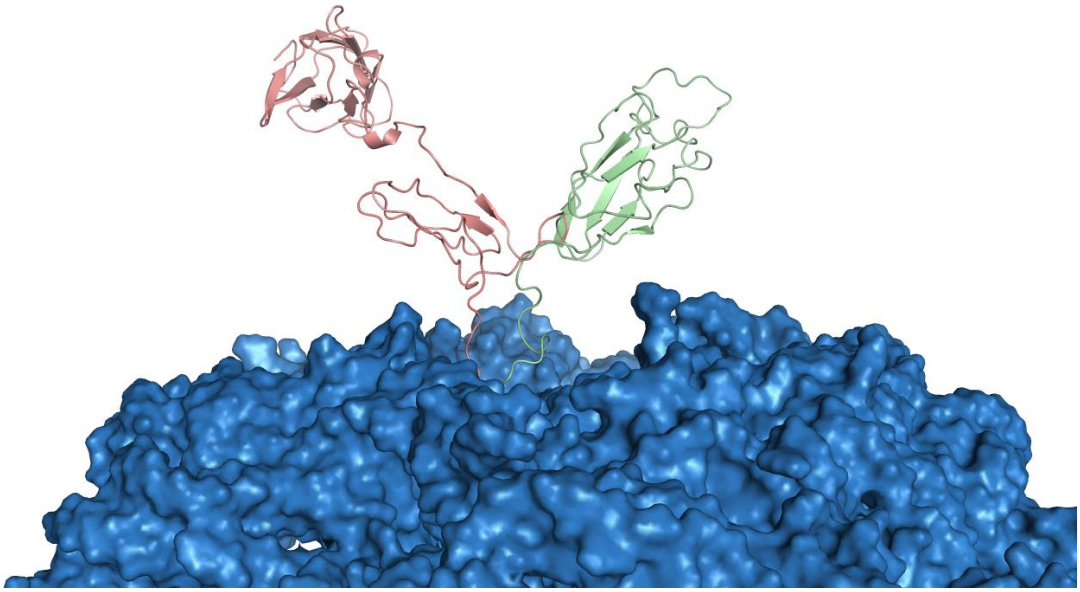

Figure 6.2 Representation of RNC systems of interest. In blue surface representation is the 50S subunit of the ribosome. The L47 system (right) is shown as a green cartoon, while the L110 system (left) is depicted as a puce cartoon. Domain 5 in both systems is observed as being entirely emerged from the ribosome, while domain 6 of L110 is still within the exit vestibule but able to form secondary structure. However, the translated residues of domain 6 in the L47 system are still predominantly within the exit tunnel.

6.1.2 Aims

The aim of this work was to demonstrate that CD-MD simulations can be performed on RNC structures; and to produce a set of ensembles of structures of the L47 and L110 Gelation Factor RNC constructs restrained by NMR chemical shifts. that demonstrate the ability of CS-MD to reproduce experimental data. In so doing, it was intended that such work form the basis for high-resolution structural investigation of the properties of nascent chains during translation.

6.2 Methods

6.2.1 Ribosome model preparation

A high resolution 50S ribosome structure was first selected from the Protein DataBank. The highest resolution and most complete structure available (PDB ID 2J01) was selected. Nucleotides corresponding to the PTC were identified by sequence alignment to published data. Using the Probe Accessible Distance

Calculator (Section **Error! Reference source not found.**) with a probe radius of .8Å, equivalent to the radius of a single amino acid, and a target grid volume of 10,000,000 voxels, a distance map to each accessible atom was calculated.

As discussed in section 6.2.2, the two nascent chain systems are of significantly different length and therefore range of motion. For the L47 construct, the maximum backbone length observed was 135Å while that of the L110 construct was 173Å. Using the map derived from PADC, two new ribosome models were created based on the 2J01 structure. For the L47 construct a structure was created that contained only those atoms that were within a nucleotide or amino acid residue that was within 150Å of the PTC, and for the L110 construct a cut-off of 200Å was used. These cut-offs were chosen such that the minimum number of atoms were retained while still allowing for extension of the constructs during MD simulation.

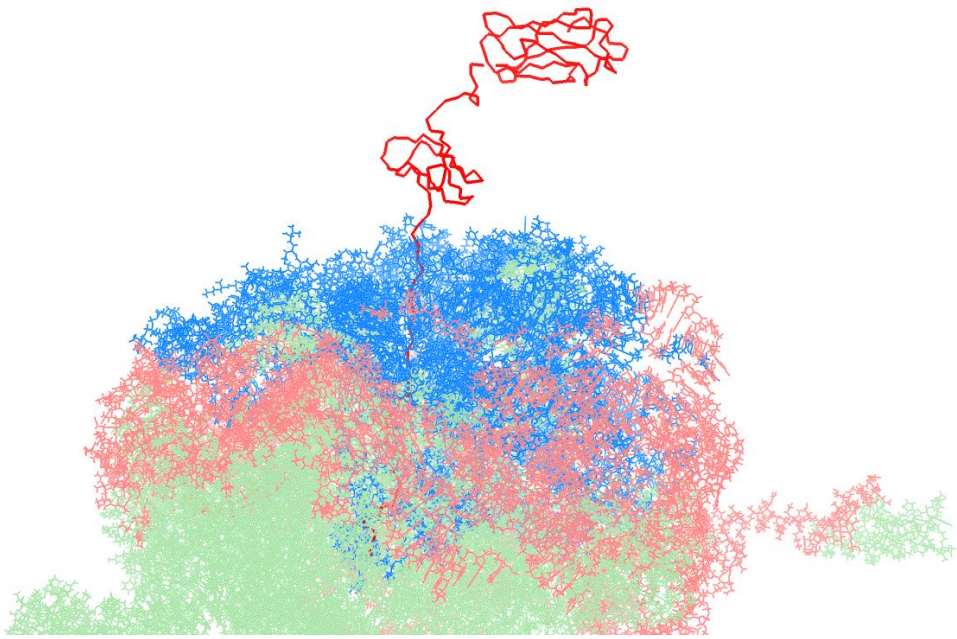

Figure 6.3 Portions of ribosome retained for different nascent chain systems. Shown are a multi-coloured ribosome 50S subunit and red nascent chain emerging from the exit tunnel. The blue portion of the ribosome represents the portion retained within the 150Å cull, used for simulations of the L47 system. In puce is the additional portion of the 50S subunit present within 200Å of the PTC used in simulations of the L110 system, while in green is the remaining portion of the 50S subunit not employed in either system.

6.2.2 Nascent chain model preparation

Unlike the ribosomal portion of the RNC, there are no high resolution structural models to use as starting structures for the simulations. Nascent chain preparation followed the following strategy.

1. Produce linear coarse-grained amino acid sequce threaded through the exit tunnel.
2. Energy-minimize the linear structure.
3. Perform high-temperature molecular dynamics simulation of coarse-grain structure to produce ensemble of structures that are sterically feasible. After equilibration at the high temperature, 128,000 steps were calculated, which corresponds to an ensemble of 128,000 coarse-grain structures
4. Select structures from the ensemble for CS-MD.
5. Construct all-atom model from coarse-grained structure.
6. Energy-minimize all-atom structure

Following this protocol a range of sterically-feasible energy-minimized all atom structures are available for subsequent refinement and processing with chemical shift restrained molecular dynamics. The selection of structures for subsequent CS-MD can be altered depending on the goals. For the purposes of this work, the most extended and least extended domain 6 structures were chosen as well as the structure with an extension closest to the ensemble mean.

Extension of domain 6 was measured by calculating the physical distance between the particles that correspond to residues 763 and 837 of Gelation Factor, which are the first and last residues of domain 6. This distance was calculated for each structure in the ensemble.

6.2.3 Reweighting protocol

As discussed in section (1.5.2), in order to coerce the simulation such that the ensemble average chemical shifts match those of the data employed as restraints, the energy function used by the simulation includes a term corresponding to deviation from the restraints with a variable multiplier term. This multiplier

term (α) may be changed to adjust the weighting in the energy function of the chemical shift term relative to the forcefield. Ie, having a low value of α causes the forcefield to have a higher influence on the energy function and therefore on the behaviour of the simulation. Conversely, a high value of α causes the simulation to adopt conformations that are in greater agreement with the chemical shifts regardless of the favourability of the forcefield.

When a system in a conformation that has very different predicted chemical shifts to the restraints, the energy term associated with chemical shift agreement will be very high. If the value of α is also high enough, the energy function will therefore impose very high forces on atoms within the simulation system. If the forces are sufficient, this will result in the system exploding and no meaningful information being extracted. It is therefore necessary to start with a simulation in which α is low and increase α at a rate that does not cause system explosions.

With small single-domain proteins, such reweighting of the chemical-shift term in the energy function may be conducted in a small number of increments (typically fewer than 5) until a maximum stable weighting is achieved. However, with a multi-domain system that includes unfolded regions and significant steric hindrance observed with RNCs, it was found that such a simple reweighting protocol was insufficient to prevent system explosions once the chemical shift term imposed a significant bias on the trajectory.

To overcome this, a more elaborate reweighting protocol was formalised:

1. Following equilibration within the force field alone (i.e. with no chemical shift restraints), CS-MD was conducted with a value of α of 0. This was conducted to ensure that the CS-MD implementation did not introduce unforeseen software errors that caused the simulation to fail. This simulation would proceed for 1ns of simulation time.
2. The final structures from step 1 would be used to conduct an identical simulation, however with α set at a low value (0.5 in the implementation used). At this value, the chemical shifts ought to have very little effect on the overall energy function. Any explosion at this stem would be a sign of

a significant problem in the starting structures, necessitating further analysis and reproduction of starting structures.

3. Following step 2, the simulation would continue, however with α increased by 0.2. If the simulation exploded, the new trajectory would be discarded, and the final structures re-used but with α at the last stable value.
4. Once the system had achieved 10ns simulation at a α value of 5.0 or higher, the final structures would be taken as the starting structures for a production simulation, and the value of α used as the appropriate weighting.

Following automation, this protocol allows for stable starting structures to be identified with minimal manual intervention.

6.3 Results

6.3.1 Extensive force field equilibration is required for the multidomain system with incomplete restraint data.

Unlike the single domain CS-MD systems, the RNC datasets were sparse, with many residues lacking any chemical shift restraints. Furthermore, the fixed ribosomal structure present in the system resulted in significant steric hindrance of residues close to the ribosome. As a result of these novel challenges, the approach to application of chemical shifts previously used for single domain systems resulted in frequent system explosions. To overcome these failures, a new protocol for introducing the restraints was developed. As outlined in 6.2.3, this protocol initially introduces the restraints as only a very small portion of the energy function and only increases the weighting of the restraints when the system is stable at the lower value. The new protocol is also automatable, allowing assessment of reproducibility and requiring only computing resource and not manual intervention to complete.

6.3.2 Domain 5 and Domain 6 structures are uncorrelated and unconnected

To assess whether folding in the nascent chain was localised to the sites of domains 5 and 6 and not occurring at the linker sequence between the domains, the distance between the terminal residue of domain 5 and the terminal residue of domain 6 in each structure of the L110 ensemble was calculated. This showed that in all structures, a distance of between 13.4Å and 28.3Å was always observed, implying a mostly-extended region in the approximately 9 inter-domain residues.

6.3.3 When emerged from ribosome, Domain 6 is adopts compact but unfolded structure

As described in section 6.2.2, the range of distances between the first and last residue of domain 6 was calculated for each of the structures within the L110 ensemble. This data were used to produce a distribution of extensions of domain 6, seen in Figure 6.4 Distribution of distances between first and last residue of domain 6 in the initial L110 coarse grain ensemble. This distance in the published crystal structure (PDB 1QFH) is 32.1Å. In the ensemble, the minimum distance was smaller, at 24.8Å but the maximum and mean distances were 65.4Å and 45.9Å respectively. The ensemble is clearly less compact than the native state as the terminal residues of the domain are further apart in the majority of the structures. Note, however, that some residues of domain 6 are still within the exit port of the ribosome and so are not able to sample compact structures with respect to the rest of domain 6.

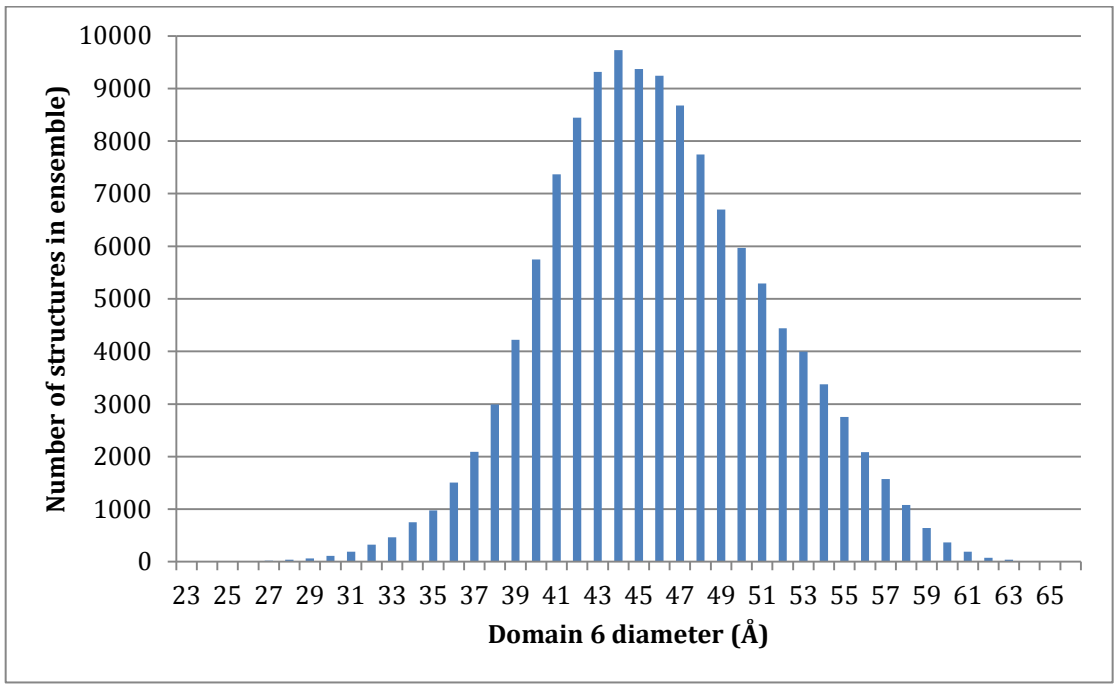

Figure 6.4 Distribution of distances between first and last residue of domain 6 in the initial L110 coarse grain ensemble.

However, as a comparison, a random walk simulation was conducted. This consisted of 128,000 independent simulations of a 74-step three-dimensional random walk, in which the direction of each step was chosen at random from those that did not involve any two points of the walk being closer than 3.5Å. This is similar to the coarse-grain simulation in that the particles in the coarse-grain simulation consisted of 3.5Å-diameter spheres that could not intersect. When the distribution of these random walks is plotted, as in Figure 6.5, the range of distances observed is much broader (3.5Å to 153Å), though the mean remains similar at 42.8Å.

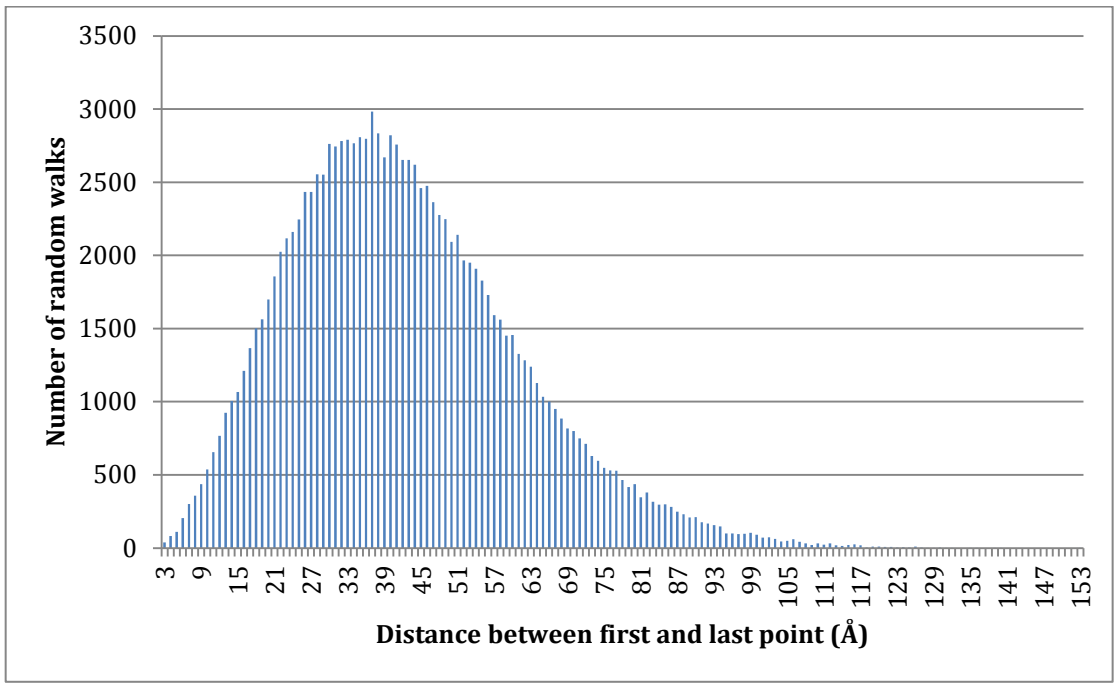

Figure 6.5 Distribution of end-to-end distances for 1280,000 random walks with 74 steps of 3.5Å length in which no to vertex may be within 3.5Å of another.

That the distribution of domain 6 sizes in the nascent chain ensemble remains narrow with a mean close to the native suggests that domain 6 retains a compact structure. Had any smaller distances been observed, this would have suggested that non compact structures in which the ends of the domain were permitted to come into close proximity were present. That no larger distances were observed shows that non-compact structures were disfavoured.

As the coarse-grain simulations incorporated high temperature dynamics to allow greater conformational sampling, this suggests that compact domain 6 structures are highly favourable, even while the whole domain is not completely accessible to fold.

6.3.4 The Steric properties of the ribosome prevent proper folding of domain 6 in the L110 system

In both the L47 and L110 ensembles, the entirety of domain 5 is fully outside of the ribosome, connected by a linker sufficient to span the entire exit tunnel and provide clearance from the exit port. It is also observed in both of these

ensembles that domain 5 forms a compact, native-like fold that forms close agreement with the native chemical shifts after refinement.

However, in the L110 ensemble, the linker is not sufficiently long to render all of the domain 6 residues outside of the ribosome. As a result of this, domain 6 is sterically impeded from adopting a native-like fold. Moreover, as seen in Figure 6.6, the folded portions of domain 6 are not forced close to the ribosome as if attempting to incorporate the remaining intra-ribosomal residues. Instead, in most structures in the ensemble, the residues that constitute the final β strand of the domain are extended, providing greater clearance from the ribosomal exit port.

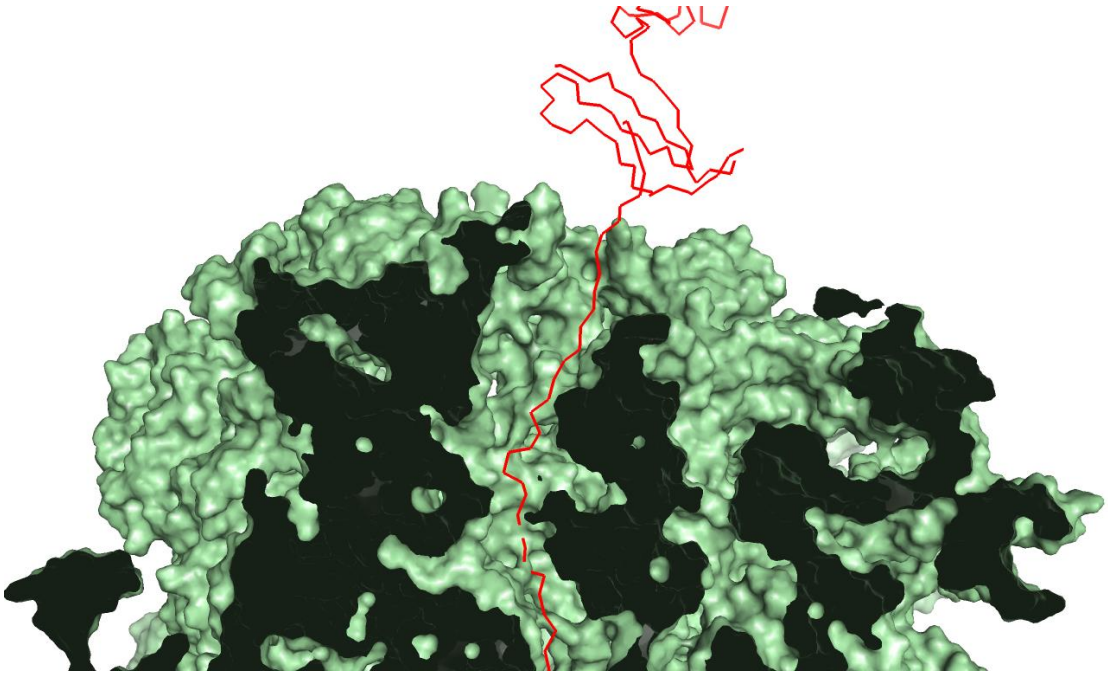

Figure 6.6 Example structure of the L110 system with a cross-section through the 50S subunit shown as a green solvent-accessible surface and the nascent chain as a red ribbon. A compact but unfolded domain 6 is visible at the exit port of the ribosome, but the dimensions of the exit port prevent domain 6 from being fully folded.

6.3.5 Domain 5 adopts native-like fold

Following equilibration with the chemical shift restraints and upon satisfaction of the chemical shifts, the domain 5 portion of both the L47 and L110 systems routinely adopts a native-like backbone fold. While extended sampling of the CS-MD system has not yet been completed, chemical shift agreement, where

available, was higher than that observed for smaller, single domain systems, with a RMSΔδ for the N^{15} shifts of between 3.8ppm and 4.2ppm. With equilibration, or with greater availability of chemical shifts, these values would be improved.

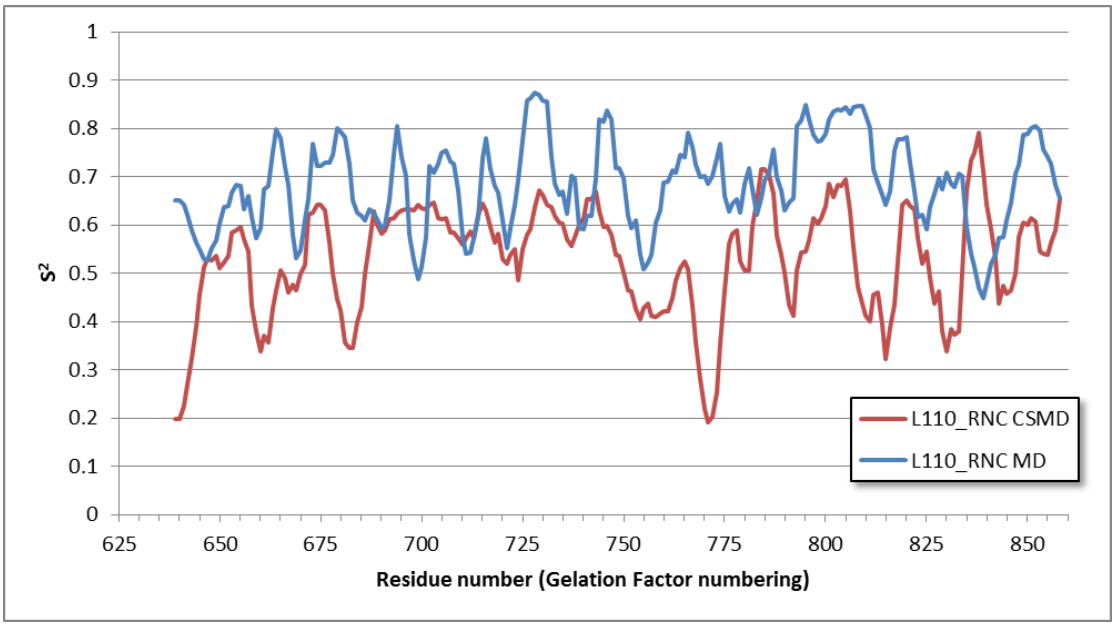

Figure 6.7 Per-residue N-HN S^2 order parameter of L110 RNC CSMD and MD systems. Calculations were performed as described in section 2.6.

As can be seen in Figure 6.7, the presence of the chemical shift restraints imposes a conformational bias on the L110 RNC system. Notably, at the inter-domain loop region (residues 747 to 763) a significant decrease in order is observed in the system under chemical shift restraint, while the more-ordered intra-domain regions remain more ordered.

6.4 Conclusion

As considered and simulated in Chapter 5, the RNC systems of Gelation Factor are significantly more difficult to study experimentally than the isolated protein, giving rise to much more spare chemical shift datasets. Consequently, the effect of the limited number of chemical shift restraints is more subtle, requiring more careful and thorough application of the restraints than for the isolated domain. To counter this issue, an automated method for performing this more intricate equilibration has been developed whereby the weighting of the chemical shifts within the force field of the CS-MD system is gradually increased, with the

system assessed for stability for a minimum of 1ns at each increase. This avoids the local minima caused by restraining only a select few nuclei from introducing unfeasible conformations that would lead to a system explosion.

While extended CS-MD trajectories were not obtainable during the present study, the only limitation to their collection is additional computer time. With data collection on a single node of a computer cluster such as the UCL Legion Research Computing service, calculation on the timescale of a month would provide simulation time of circa 100ns; highly-detailed dynamic structural information could be obtained within little time, little less of it requiring human intervention. Alternatively, use of structures taken from published isolated domains as the starting conditions for the simulation would decrease the equilibration time, albeit at the expense of biasing the initial conformation.

An additional approach that remains under investigation is the use of chemical shifts to screen an initial coarse grained ensemble for agreement. This involves the rapid generation of a large range of possible conformations, clustering to select representative samples, all-atom completion, energy minimisation and short time-scale (circa 1ns) CS-MD to reveal the best RMSΔδ at each of a broad range of conformations. The best structures could then be used as initial replicas for ensemble-averaged CS-MD. This will allow a large conformational space to be sampled, but with reduced computational time required for the restraints and without the bias or requirement introduced by using a known structure as the initial conformation.

The simulations conducted here predict that the domain 5 and 6 Gelation Factor system, L110, forms two compact domains that do not interact except for the imposition of their peptide linkage. With the presence of additional chemical shifts, it is likely that the structures would form more native-like folding of domain 5. However, the CS-MD simulations predict that domain 6 would not be able to form a native fold due to the restrictions imposed by the presence of the ribosome. Instead, without a longer C-terminal linker, this domain can only partially fold.

7 Concluding remarks

This thesis presents methods and strategies for the cell-free preparation of RNCs and for the use of NMR data in all-atom molecular dynamics simulations to reveal high-resolution details of the structural properties of proteins during biosynthesis on the ribosome. Recent advances in the *in vivo* preparation of RNCs (Christensen et al., 2011) have enabled the recording of limited NMR data, in particular chemical shift spectra with sharp peaks selectively recorded for spin systems within the nascent chain. While this provides a tantalising glimpse into the effects of the ribosome on the co-translational folding pathways of nascent chains, the nature of the macromolecular system limits samples to approximately 10μM concentration with sample lifetimes from hours to days prevent the recording of NOE spectra that would enable traditional methods of solving the structural properties to high resolution. However, developments in the use of chemical shifts to resolve high-resolution protein properties by restraining molecular dynamics simulations (Bundi and Wüthrich, 1979) provides a promising avenue for making greater use of the data that are available for RNC systems.

Chapter 3 outline the successful preparation of αSyn-RNCs by both *in vitro* and *in vivo* methods, as well as methodological enhancements that were necessary for the detection and quantification of samples. A sample required for NMR study would require approximately 300μl at 5-10μM concentration. Furthermore, the instability of ribosomes in concentrations above this concentration requires that almost all ribosomes present in the sample must be occupied by a nascent chain lest the signal be too week (if under-occupied) or the ribosomes prematurely degrade (if over-concentrated).

In vitro preparation of ^{15}N-labeled αSyn-RNCs was performed, with positive results. A number of factors to improve stability and occupancy of samples were explored, including plasmid quality reaction time, presence of protease inhibitors, labelled amino acid source and sucrose cushioning parameters. With these advances in preparation protocol, it was calculated that the necessary materials to provide NMR-scale samples could indeed be supplied using the

Roche Diagnostics Rapid Translation System RTS5000 cell-free protein biosynthesis kits. By modulation of the amino acid source, it is possible to precisely control the labelling of amino acids within the resultant RNC. It would therefore follow that it is possible to use selective ^{13}C labelling to further screen the NMR data to shown only specific residues of interest. The cost of this larger cell-free kit, however, was prohibitive for the studies undertaken here.

As described in Chapter 4, RNC systems impose many impediments to direct application of CSMD techniques. Principal among these obstacles is the enormous system complexity; while CSMD requires an all-atom simulation that is feasible for small protein systems (RNase A, a prototype CSMD system, contains approximately 2200 atoms), the 50S ribosomal subunit alone contains in excess of 168,000 atoms. Beyond the technical challenges of simulating such a large system, the sampling rate for such a system is prohibitively slow with simulations requiring many years to complete. It is demonstrated that adjustment of commonly-used simulation parameters (τ_T and ξ_{BD}) to increase the sampling rate of the RNC structure do not impart any perceptible improvement on the range of conformations of the nascent chain. Fortunately, as has been revealed in unrestrained simulations of RNC systems, simulating the ribosome as a dynamic system has no significant effect on the energy landscape of tethered nascent chains (Case, 1995, O'Brien et al., 2010). As a result, it is possible to ignore the dynamics of the ribosome and to remove atoms that have no interaction with the ribosome.

While a tremendous advantage to simulations of RNC systems, removal of non-interacting atoms from the ribosome is not a trivial task. In section 4.2.3, the task is to remove as many atoms as possible while ensuring that no atoms are removed that may interact with the ribosome. Naïve approaches, such as removal of solvent-inaccessible atoms or removal of atoms by eye both retain many more atoms than necessary (thus dramatically slowing the rate of simulation) and, in the case of the latter method, remove atoms that can interact in unintuitive ways with the nascent chain. Instead, Chapter 4 describes the development of a new technique – the Probe-Accessible Distance Calculator – that identifies the atoms that may interact with a particle of a specified radius

and the minimum through-*vacuo* distance that such a probe would have to traverse from any given atom to any accessible atom. Inspired by the level-set method of defining solvent accessible surface (Xu and Case, 2001), PADC treats a structure as a discrete three-dimensional grid with sub-ångström cell-widths. This discretises the three-dimensional co-ordinate space of an RNC system to any selected resolution, and uses a three-stage search process to 1) identify ribosome-occupied voxels, 2) identify those voxels that are accessible to a probe of any given radius and 3) calculate the shortest distance through accessible voxels to every probe-accessible ribosome-occupied voxel. These distances are then mapped back to the ribosomal atoms by which they are occupied.

By applying the PADC to a chose ribosome structure, using a probe radius equivalent to the radius of an amino acid and defining the origin at the PTC of the ribosome, it is possible to accurately identify all atoms that may interact with a nascent chain and how long a nascent chain would have to be to reach any point on the ribosome. Consequently, for any chosen nascent chain system, it is possible to rapidly produce ribosome models with complete amino acids that contain an atom that may interact with the nascent chain while removing unnecessary atoms. Depending on the nascent chain system, typically 4-20,000 atoms from the 50S subunit need to be retained to provide the ribosome interaction with the nascent chain. While the resulting system is still very large for all atom CSMD, it makes the simulation timescales feasible with typically 4 to 6 weeks required for production runs.

Furthermore, Chapter 4 demonstrates that even coarse-grained simulations of RNCs with a simple Lennard-Jones force field can be used to sample the broadest extent of conformations accessible to a nascent chain. Calculation of the per-residue S^2 model-free order parameter based on such a trajectory was used to make useful predictions as to the NMR-visibility of residues. Such predictions should be tested in further work, but could be of considerable use in designing constructs such that the minimal linker length is used that still enables regions of interest in the nascent chain to be NMR visible.

Where Chapter 4 describes the methods necessary to make CSMD of RNC systems feasible, Chapters 5 and 6 describe methods for the analysis of CSMD trajectories and the application of the techniques to actual CSMD of RNCs.

To monitor the effect of chemical shift restraints on MD trajectories, a measurement has been developed – called RMSΔδ – that assesses the difference between the predicted chemical shift associated with each nuclei for structures within a simulation at any given time point, and the experimentally-derived chemical shifts that are being used as restraints. While signal from this metric is noisy – small local variations can affect the values within a range of ±0.2 ppm in for backbone nitrogen nuclei, for example – trends in this value or window averages over 1-10ns can provide useful information as to whether the system is converging or diverging from the restraint values.

In Chapter 5, a CS-MD is applied to a mimetic for an RNC system: progressive C-terminal truncations of a folded Ig-like domain in isolation. The domain – taken from a protein known to exhibit co-translational folding – is well characterised with atomic-resolution structures and assigned NMR spectra available for comparison. By applying NMR restraints to the system, it was observed to affect the structural properties as measured by deviation from the chemical shift by RMSΔδ values. Moreover, by artificially excluding data from the restraints, the system was observed to coerce unrestrained residues to match the data. This implies that even in sparse datasets such as those recorded for RNC systems, even the limited chemical shift data available may have global effects on the system that leads it to conform to ensembles that resemble the unavailable data.

Across the production ensemble of these truncation systems, Free Energy Landscapes of RMSΔδ versus RMSD demonstrated multiple distinct sub populations. In particular, in concordance with the theory that the more truncated systems would be less stable (due to the lack of the final G-strand to provide hydrogen bonding to stabilise the structure), those systems with greater truncation exhibited a greater range of sub populations.

Per-residue analysis of the S^2 model free order parameter suggests that the unrestrained MD successfully identified stable secondary structural regions, although less stably so than the CS-MD systems. However, in loop regions

between secondary structural elements, the unrestrained MD predicted significant disorder, while the CS-MD retained fairly ordered conformations. This suggests that the CS-MD is restraining these less-ordered in ranges of conformations that still have significant biases, presumably more like those adopted experimentally.

The GelFac$_{646\text{-}750}$ CS-MD system demonstrates a highly-ordered C-terminal tail, suggesting that it has adopted a stable, native-like fold. The GelFac$_{646\text{-}746}$ system, however, demonstrates a significant decrease in order at the C-terminal tail from residues 740 onwards – more so than the unrestrained MD – indicating that the lack of the final 4 residues confers a significant disruption to the stability of the G-strand. The GelFac$_{646\text{-}744}$ system, however, exhibits stability at its final residue, suggesting that without an unstable ultimate G-strand present, the penultimate strand retains a stable fold.

Attempts to use the current state-of-the-art structure-from-chemical shifts CHESHIRE pipeline failed to reproduce native-like conformations, which is likely partly a consequence of the dynamic conformations expected from this system. CS-MD was clearly found to be a far more accurate way to reproduce structural information from chemical shifts.

In Chapter 6, the application of NMR chemical shifts from an RNC to a CS-MD system are described. By using optimised CS-MD procedure developed in Chapter 4, all-atom RNC systems are simulated with chemical shifts incorporated as restraints. Additional difficulties were observed and surmounted, in particular the challenge of appropriately equilibrating such a complex system with the sparse restraint dataset. To counter this issue, an automated method for performing this more intricate equilibration has been developed whereby the weighting of the chemical shifts within the force field of the CS-MD system is gradually increased, with the system assessed for stability for a minimum of 1ns at each increase. This avoids the local minima caused by restraining only a select few nuclei from introducing unfeasible conformations that would lead to a system explosion. While extended CS-MD trajectories were not obtainable during the present study, the only limitation to their collection is additional computer time.

While preliminary in nature, the results showed the ability to coerce the system into ensembles that more closely match the available data, suggesting that the approach is feasible and worthy of further exploration.

An additional approach that remains under investigation is the use of chemical shifts to screen an initial coarse grained ensemble for agreement (CG+CS-MD). This involves the rapid generation of a large range of possible conformations, clustering to select representative samples, all-atom completion, energy minimisation and short time-scale (circa 1ns) CS-MD to reveal the best RMSΔδ at each of a broad range of conformations. The best structures could then be used as initial replicas for ensemble-averaged CS-MD. This will allow a large conformational space to be sampled, but with reduced computational time required for the restraints and without the bias or requirement introduced by using a known structure as the initial conformation.

This thesis demonstrates that it now possible to obtain high-resolution ensemble conformations of RNCs that adhere to the chemical shift data that are beginning to be collected, thus providing high-resolution structures of proteins undergoing folding on the ribosome for the first time.

Presented here are solutions to all of the hurdles faced when attempting to conduct chemical shift restrained molecular dynamics investigations of RNCs. Specifically addressed are: methods for generation of initial nascent chain structures, methods for evaluation of those initial structures, methods for incorporating the ribosome (including simulation parameters and generation of reduced structures), methods for conducting all-atom molecular dynamics on those RNC structures, methods for incorporating the chemical shift restraints on those structures and methods for the analysis of RNC CS-MD trajectories. The work presented here has involved the creation of protocols and scripts that automate much of this work, leaving the researcher free from repetitive data-processing tasks and free to undertake more elaborate analyses.

Further work in this area is now able to rapidly produce CS-MD RNC simulations; to investigate the CG+CS-MD refinement protocol proposed to rapidly sample a broad range of conformations; and to analyse the trajectories to make a plethora of biologically-relevant predictions.

8 Appendices

8.1 Annotated αSyn-RNC sequence

```
T7 promoter          lacO
TAATACGACTCACTATAGG GGAATTGTGAGCGGATAACAATTCCCC

XbaI                T7_transl_en_RBS        NdeI
TCTAGA AATAATTTTGT TTAACTTTAAGAAGGAG ATATA CATATG

HisTag              NheI
CATCACCATCACCATCAT GCTAGC

Alpha-synculein
ATGGATGTAT TCATGAAAGG ACTTTCAAAG GCCAAGGAGG GAGTTGTGGC TGCTGCTGAG
AAAACCAAAC AGGGTGTGGC AGAAGCAGCA GGAAAGACAA AAGAGGGTGT TCTCTATGTA
GGCTCCAAAA CCAAGGAGGG AGTGGTGCAT GGTGTGGCAA CAGTGGCTGA GAAGACCAAA
GAGCAAGTGA CAAATGTTGG AGGAGCAGTG GTGACGGGTG TGACAGCAGT AGCCCAGAAG
ACAATGGAGG GAGCAGGGAG CATTGCAGCA GCCACTGGCT TTGTCAAAAA GGACCAGTTG
GGCAAGAATG AAGAAGGAGC CCCACAGGAA GGAATTCTGG AAGATATGCC TGTGGATCCT
GACAATGAGG CTTATGAAAT GCCTTCTGAG GAAGGGTATC AAGACTACGA ACCTGAAGCC

KpnI  SpeI    EcoRI
GGTAC CACTAGT GAATTC

SecM                                                    ***
TTCAGCACGCCCGTCTGGATAAGCCAGGCGCAAGGCATCCGTGCTGGCCCTTAA

NcoI                linker
CCATGGACCTAACAACAA TAAACCTTACTTCATTTTATTAACTCCGCAA

toeprint             HindI  linker
CGCGGGGCGTTTGAGATTTT AAGCTT TCTCTC

StreptoTag
GGATCGCATTTGGACTTCTGCCCAGGGTGGCCCCACGGTGGATCC

XhoI   6xHis
CTCGAG CACCACCACCACCACCAC TGAGATCCGG

T7 terminator
CTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCC
TTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTG
```

DNA sequence of the αSyn-pLDC17 vector at the RNC construct site. Nucleotides in blue are derived from the Novagen pET21b(+) vector backbone. Regions of interest – including restriction sites and protein segments – are annotated in italics. The Start codon of the nascent chain sequence is in bold. The stalling proline codon that forms a Pro-tRNA^{Pro} link at the ribosomal P-site is marked with asterisks.

8.2 MDG Media

Stock	Final Concentration (per flask)

50x salts (autoclave)	1x (5ml)
5%(w/v) L-aspartic acid pH 7.0 (autoclave)	0.2%(v/v) (10ml)
1M MgSO4 (filter sterilize)	2mM (0.5ml)
1000x trace metals (filter sterilize)	0.2x (50ul)
20% (w/v) glucose (autoclave)	0.4% (v/v) (5ml)
100mg/ml Ampicillin	100mg/ml (250µl)

50x salts: 1.25M Na_2HPO_4, 1.25M KH_2PO_4, 2.5M NH_4Cl, 0.25M Na_2SO_3
1000x trace metals: 50mM $FeCl_2$ [dissolved in 0.1M HCl], 20mM $CaCl_2$, 1mM each ($MnCl_2.4H_2O$, $ZnSO_4.7H_2O$), 2mM each ($CoCl_2.6H_2O$, $CuCl_2.2H_2O$, $NiCl_2.6H_2O$, $Na_2MoO_4.2H_2O$, $Na_2SeO_3.5H_2O$, H_3BO_3)

8.3 EM9 (Enhanced M9) Media

Stock	Final Concentration (per flask)
10x EM9 salts (autoclave)	1x (25ml)
20%(w/v) NH_4Cl (filter sterilise) pH7.5	1g/L (2.5ml)
20% (w/v) glucose (autoclave/filter sterilise)	0.4% (v/v) (5ml)
100x BME Vitamins	0.25x (0.625ml)
1M $CaCl_2$ (filter sterilise)	0.2mM (50µl)
1M $MgSO_4$ (filter sterilise)	5mM (1.25ml)
100mg/ml Ampicillin	100mg/ml (0.5ml)
1000x Trace metals	0.25x (63 µl)

EM9 salts: 71g/L Na_2HPO_4, 34g/L KH_2PO_4, 5.84g/L NaCl, pH 8.0-8.2

8.4 Tico Buffer

10mM Hepes, 30mM NH4Cl, 12mM MgCl2, 1mM BME, pH 7.5

9 References

ANFINSEN, C. B. (1973) Principles that Govern the Folding of Protein Chains. *Science,* **181**, 223-230.

ATIEH, Z., AUBERT-FRÉCON, M. & ALLOUCHE, A.-R. (2010) Rapid, Accurate and Simple Model to Predict NMR Chemical Shifts for Biological Molecules. *Journal of Physical Chemistry B,* **114**, 16388-16392.

BALDWIN, R. L. (1993) Pulsed H/D-exchange studies of folding intermediates. *Current Opinion in Structural Biology,* **3**, 84-91.

BAN, N., NISSEN, P., HANSEN, J., MOORE, P. B. & STEITZ, T. A. (2000) The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4Å Resolution. *Science,* **289**, 905-920.

BARDUCCI, A., BONOMI, M. & PARRINELLO, M. (2011) Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science,* **1**, 826-843.

BARTELS, A. L. & LEENDERS, K. L. (2009) Parkinson's disease: The syndrome, the pathogenesis and pathophysiology. *Cortex,* **45**, 915-921.

BASHAN, A. & YONATH, A. (2005) Ribosome crystallography: catalysis and evolution of peptide-bond formation, nascent chain elongation and its co-translational folding. *Biochemical Society Transactions,* **33**, 488–492.

BECKER, T., BHUSHAN, S., JARASCH, A., ARMACHE, J.-P., FUNES, S., JOSSINET, F., GUMBART, J., MIELKE, T., BERNINGHAUSEN, O., SCHULTEN, K., WESTHOF, E., GILMORE, R., MANDON, E. C. & BECKMANN, R. (2009) Structure of Monomeric Yeast and Mammalian Sec61 Complexes Interacting with the Translating Ribosome. *Science,* **326**, 1369-1373.

BEEMAN, D. (1976) Some multistep methods for use in molecular dynamics calculations. *Journal of Computational Physics,* **20**, 130-139.

BEN-SHEM, A., JENNER, L., YUSUPOVA, G. & YUSUPOV, M. (2010) Crystal Structure of the Eukaryotic Ribosome. *Science,* **330**, 1203-1209.

BEST, R. B. & HUMMER, G. (2009) Optimized Molecular Dynamics Force Fields Applied to the Helix–Coil Transition of Polypeptides. *The Journal of Physical Chemistry B,* **113**, 9004-9015.

BHUSHAN, S., HOFFMANN, T., SEIDELT, B., FRAUENFELD, J., MIELKE, T., BERNINGHAUSEN, O., WILSON, D. N. & BECKMANN, R. (2011) SecM-Stalled Ribosomes Adopt an Altered Geometry at the Peptidyl Transferase Center. *PLoS Biol,* **9**, e1000581.

BONOMI, M., BRANDUARDI, D., BUSSI, G., CAMILLONI, C., PROVASI, D., RAITERI, P., DONADIO, D., MARINELLI, F., PIETRUCCI, F., BROGLIA, R. A. & PARRINELLO, M. (2009) PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications,* **180**, 1961-1972.

BUNDI, A. & WÜTHRICH, K. (1979) 1H-nmr parameters of the common amino acid residues measured in aqueous solutions of the linear tetrapeptides H-Gly-Gly-X-L-Ala-OH. *Biopolymers,* **18**, 285-297.

BUNGARTZ, H.-J., ZIMMER, S., BUCHHOLZ, M. & PFLÜGER, D. (2014) *Chapter 13: Molecular Dynamics.* in BORWEIN, J. M. & HOLDEN, H. (Eds.) *Modeling and Simulation: An application-oriented introduction* Springer-Verlag Berlin Heidelberg,

BUSSI, G., DONADIO, D. & PARRINELLO, M. (2007) Canonical sampling through velocity rescaling. *The Journal of Chemical Physics,* **126**, 014101-7.

CABRITA, L. D., DOBSON, C. M. & CHRISTODOULOU, J. (2010) Early Nascent Chain Folding Events on the Ribosome. *Israel Journal of Chemistry,* **50**, 99-108.

CAMILLONI, C., DE SIMONE, A., VRANKEN, W. F. & VENDRUSCOLO, M. (2012a) Determination of Secondary Structure Populations in Disordered States of Proteins Using Nuclear Magnetic Resonance Chemical Shifts. *Biochemistry,* **51**, 2224-2231.

CAMILLONI, C., ROBUSTELLI, P., SIMONE, A. D., CAVALLI, A. & VENDRUSCOLO, M. (2012b) Characterization of the Conformational Equilibrium between the Two Major Substates of RNase A Using NMR Chemical Shifts. *Journal of the American Chemical Society,* **134**, 3968-3971.

CAN, T., CHEN, C.-I. & WANG, Y.-F. (2006) Efficient molecular surface generation using level-set methods. *Journal of Molecular Graphics and Modelling,* **25**, 442-454.

CASE, D. (1995) Calibration of ring-current effects in proteins and nucleic acids. *Journal of Biomolecular NMR,* **6**, 341-346.

CAVALLI, A., SALVATELLA, X., DOBSON, C. M. & VENDRUSCOLO, M. (2007) Protein structure determination from NMR chemical shifts. *Proceedings of the National Academy of Sciences,* **104**, 9615-9620.

CAVANAGH, J., FAIRBROTHER, W. J., ARTHUR G. PALMER, I., RANCE, M. & SKELTON, N. J. (2007) *Chapter 5 Relaxation and Dynamic Processes.* in *Protein NMR Spectroscopy: Principles and Practice Second Edition* Elsevier Academic Press, London.

CHAMARY, J. V., PARMLEY, J. L. & HURST, L. D. (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet,* **7**, 98-108.

CHANDRA, S., GALLARDO, G., FERNÁNDEZ-CHACÓN, R., SCHLÜTER, O. M. & SÜDHOF, T. C. (2005) [alpha]-Synuclein Cooperates with CSP[alpha] in Preventing Neurodegeneration. *Cell,* **123**, 383-396.

CHARETTE, S. J., LAMBERT, H., NADEAU, P. J. & LANDRY, J. (2009) Protein quantification by chemiluminescent Western blotting: Elimination of the antibody factor by dilution series and calibration curve. *Journal of Immunological Methods,* **353**, 148-150.

CHEN, Q., THORPE, J. & KELLER, J. N. (2005) {alpha}-Synuclein Alters Proteasome Function, Protein Synthesis, and Stationary Phase Viability. *J. Biol. Chem.,* **280**, 30009-30017.

CHRISTENSEN, A. S., SAUER, S. P. A. & JENSEN, J. H. (2011) Definitive Benchmark Study of Ring Current Effects on Amide Proton Chemical Shifts. *Journal of Chemical Theory and Computation,* **7**, 2078-2084.

CLARK, P. L. (2004) Protein folding in the cell: reshaping the folding funnel. *Trends in Biochemical Sciences,* **29**, 527-534.

CLARK, P. L. & KING, J. (2001) A Newly Synthesized, Ribosome-bound Polypeptide Chain Adopts Conformations Dissimilar from Early in VitroRefolding Intermediates. *Journal of Biological Chemistry,* **276**, 25411-25420.

CLAYTON, D. F. & GEORGE, J. M. (1999) Synucleins in synaptic plasticity and neurodegenerative disorders. *Journal of Neuroscience Research,* **58**, 120-129.

COOPER, A. A., GITLER, A. D., CASHIKAR, A., HAYNES, C. M., HILL, K. J., BHULLAR, B., LIU, K., XU, K., STRATHEARN, K. E., LIU, F., CAO, S., CALDWELL, K. A., CALDWELL, G. A., MARSISCHKY, G., KOLODNER, R. D., LABAER, J., ROCHET, J.-C., BONINI, N. M. &

LINDQUIST, S. (2006) {alpha}-Synuclein Blocks ER-Golgi Traffic and Rab1 Rescues Neuron Loss in Parkinson's Models. *Science,* **313**, 324-328.

DEDMON, M. M., LINDORFF-LARSEN, K., CHRISTODOULOU, J., VENDRUSCOLO, M. & DOBSON, C. M. (2004) Mapping Long-Range Interactions in α-Synuclein using Spin-Label NMR and Ensemble Molecular Dynamics Simulations. *Journal of the American Chemical Society,* **127**, 476-477.

DER-SARKISSIAN, A., JAO, C. C., CHEN, J. & LANGEN, R. (2003) Structural Organization of {alpha}-Synuclein Fibrils Studied by Site-directed Spin Labeling. *J. Biol. Chem.,* **278**, 37530-37535.

DOBSON, C. M., EVANS, P. A. & RADFORD, S. E. (1994) Understanding how proteins fold: the lysozyme story so far. *Trends in Biochemical Sciences,* **19**, 31-37.

DONG, H., NILSSON, L. & KURLAND, C. G. (1996) Co-variation of tRNA Abundance and Codon Usage in *Escherichia coli* at Different Growth Rates. *Journal of Molecular Biology,* **260**, 649-663.

DULEBOHN, D., CHOY, J., SUNDERMEIER, T., OKAN, N. & KARZAI, A. W. (2007) Trans-Translation: The tmRNA-Mediated Surveillance Mechanism for Ribosome Rescue, Directed Protein Degradation, and Nonstop mRNA Decay. *Biochemistry,* **46**, 4681-4693.

EICHMANN, C., PREISSLER, S., RIEK, R. & DEUERLING, E. (2010) Cotranslational structure acquisition of nascent polypeptides monitored by NMR spectroscopy. *Proceedings of the National Academy of Sciences*.

ELIEZER, D., KUTLUAY, E., BUSSELL JR, R. & BROWNE, G. (2001) Conformational properties of [alpha]-synuclein in its free and lipid-associated states. *Journal of Molecular Biology,* **307**, 1061-1073.

ELLIS, J. P., BAKKE, C. K., KIRCHDOERFER, R. N., JUNGBAUER, L. M. & CAVAGNERO, S. (2008) Chain Dynamics of Nascent Polypeptides Emerging from the Ribosome. *ACS Chemical Biology,* **3**, 555-566.

ENGLANDER, S. W. & MAYNE, L. (1992) Protein Folding Studied Using Hydrogen-Exchange Labeling and Two-Dimensional NMR. *Annual Review of Biophysics and Biomolecular Structure,* **21**, 243-265.

ETCHELLS, S. A. & HARTL, F. U. (2004) The dynamic tunnel. *Nat Struct Mol Biol,* **11**, 391-392.

EVANS, M. S., UGRINOV, K. G., FRESE, M.-A. & CLARK, P. L. (2005) Homogeneous stalled ribosome nascent chain complexes produced in vivo or in vitro. *Nat Meth,* **2**, 757-762.

FEDOROV, A. N. & BALDWIN, T. O. (1995) Contribution of cotranslational folding to the rate of formation of native protein structure. *Proceedings of the National Academy of Sciences,* **92**, 1227-1231.

FEIG, M., KARANICOLAS, J. & CHARLES L. BROOKS, I. (2009) *MMTSB Tool Set, MMTSB NIH Research Resource, The Scripps Research Institute.*

FERSHT, A. R. (1995) Characterizing transition states in protein folding: an essential step in the puzzle. *Current Opinion in Structural Biology,* **5**, 79-84.

FRENKEL, D. & SMIT, B. (2002) *Understanding Molecular Simulation From Algorithms to Applications,* London, Academic Press.

FUCINI, P., RENNER, C., HERBERHOLD, C., NOEGEL, A. A. & HOLAK, T. A. (1997) The repeating segments of the F-actin cross-linking gelation factor (ABP-120) have an immunoglobulin-like fold. *Nat Struct Mol Biol,* **4**, 223-230.

FULLE, S. & GOHLKE, H. (2009) Statics of the Ribosomal Exit Tunnel: Implications for Cotranslational Peptide Folding, Elongation Regulation, and Antibiotics Binding. *Journal of Molecular Biology,* **387**, 502-517.

GILBERT, R. J. C., FUCINI, P., CONNELL, S., FULLER, S. D., NIERHAUS, K. H., ROBINSON, C. V., DOBSON, C. M. & STUART, D. I. (2004) Three-Dimensional Structures of Translating Ribosomes by Cryo-EM. *Molecular Cell,* **14**, 57-66.

GOEDERT, M. (2001) Alpha-synuclein and neurodegenerative diseases. *Nat Rev Neurosci,* **2**, 492-501.

GOLDBERG, M. E., SEMISOTNOV, G. V., FRIGUET, B., KUWAJIMA, K., PTITSYN, O. B. & SUGAI, S. (1990) An early immunoreactive folding intermediate of the tryptophan synthase β_2 subunit is a 'molten globule'. *FEBS Letters,* **263**, 51-56.

GRANT, S. G., JESSEE, J., BLOOM, F. R. & HANAHAN, D. (1990) Differential plasmid rescue from transgenic mouse DNAs into Escherichia coli methylation-restriction mutants. *Proceedings of the National Academy of Sciences,* **87**, 4645-4649.

GUVENCH, O. & MACKERELL, A., JR. (2008) *Comparison of Protein Force Fields for Molecular Dynamics Simulations.* in KUKOL, A. (Ed.) *Molecular Modeling of Proteins 63-88.* Humana Press,

HAIGH, C. W. & MALLION, R. B. (1979) Ring current theories in nuclear magnetic resonance. *Progress in Nuclear Magnetic Resonance Spectroscopy,* **13**, 303-344.

HAN, B., LIU, Y., GINZINGER, S. & WISHART, D. (2011) SHIFTX2: significantly improved protein chemical shift prediction. *Journal of Biomolecular NMR,* **50**, 43-57.

HARTL, F. U. & HAYER-HARTL, M. (2009) Converging concepts of protein folding in vitro and in vivo. *Nat Struct Mol Biol,* **16**, 574-581.

HENRY, E. R. & SZABO, A. (1985) Influence of vibrational motion on solid state line shapes and NMR relaxation. *The Journal of Chemical Physics,* **82**, 4753-4761.

HESS, B., BEKKER, H., BERENDSEN, H. J. C. & FRAAIJE, J. G. E. M. (1997) LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry,* **18**, 1463-1472.

HOCKNEY, R. W., GOEL, S. P. & EASTWOOD, J. W. (1974) Quiet high-resolution computer models of a plasma. *Journal of Computational Physics,* **14**, 148-158.

HSU, S.-T. D., FUCINI, P., CABRITA, L. D., LAUNAY, H. L. N., DOBSON, C. M. & CHRISTODOULOU, J. (2007) Structure and dynamics of a ribosome-bound nascent chain by NMR spectroscopy. *Proceedings of the National Academy of Sciences,* **104**, 16516-16521.

HSU, S. T. D., CABRITA, L. D., CHRISTODOULOU, J. & DOBSON, C. M. (2009a) 1H, 15N and 13C assignments of domain 5 of Dictyostelium discoideum gelation factor (ABP-120) in its native and 8M urea-denatured states. *Biomolecular NMR Assignments,* **3**, 29-31.

HSU, S. T. D., CABRITA, L. D., FUCINI, P., CHRISTODOULOU, J. & DOBSON, C. M. (2009b) Probing side-chain dynamics of a ribosome-bound nascent chain using methyl NMR spectroscopy. *Journal of the American Chemical Society,* **131**, 8366-8367.

HSU, S. T. D., CABRITA, L. D., FUCINI, P., DOBSON, C. M. & CHRISTODOULOU, J. (2009c) Structure, Dynamics and Folding of an Immunoglobulin Domain of the

Gelation Factor (ABP-120) from Dictyostelium discoideum. *Journal of Molecular Biology,* **388**, 865-879.

HUT, P., MAKINO, J. & MCMILLAN, S. (1995) Building a Better Leapfrog. *Astrophysical Journal Letters,* **443**, L93-L96.

IVANOV, I. G., SARAFFOVA, A. A. & ABOUHAIDAR, M. G. (1997) Unusual effect of clusters of rare arginine (AGG) codons on the expression of human interferon α1 gene in Escherichia Coli. *The International Journal of Biochemistry & Cell Biology,* **29**, 659-666.

JARASCH, A., DZIUK, P., BECKER, T., ARMACHE, J.-P., HAUSER, A., WILSON, D. N. & BECKMANN, R. (2011) The DARC site: a database of aligned ribosomal complexes. *Nucleic Acids Research*.

JOHNSON, C. E. & BOVEY, F. A. (1958) Calculation of Nuclear Magnetic Resonance Spectra of Aromatic Hydrocarbons. *The Journal of Chemical Physics,* **29**, 1012-1014.

JULIÁN, P., KONEVEGA, A. L., SCHERES, S. H. W., LÁZARO, M., GIL, D., WINTERMEYER, W., RODNINA, M. V. & VALLE, M. (2008) Structure of ratcheted ribosomes with tRNAs in hybrid states. *Proceedings of the National Academy of Sciences,* **105**, 16924-16927.

KAISER, C. M., CHANG, H.-C., AGASHE, V. R., LAKSHMIPATHY, S. K., ETCHELLS, S. A., HAYER-HARTL, M., HARTL, F. U. & BARRAL, J. M. (2006) Real-time observation of trigger factor function on translating ribosomes. *Nature,* **444**, 455-460.

KAISER, C. M., GOLDMAN, D. H., CHODERA, J. D., TINOCO, I. & BUSTAMANTE, C. (2011) The Ribosome Modulates Nascent Protein Folding. *Science,* **334**, 1723-1727.

KATRANIDIS, A., GRANGE, W., SCHLESINGER, R., CHOLI-PAPADOPOULOU, T., BRÜGGEMANN, D., HEGNER, M. & BÜLDT, G. (2011) Force measurements of the disruption of the nascent polypeptide chain from the ribosome by optical tweezers. *FEBS Letters,* **585**, 1859-1863.

KELKAR, D. A., KHUSHOO, A., YANG, Z. & SKACH, W. R. (2012) Kinetic Analysis of Ribosome-bound Fluorescent Proteins Reveals an Early, Stable, Cotranslational Folding Intermediate. *Journal of Biological Chemistry,* **287**, 2568-2578.

KHUSHOO, A., YANG, Z., JOHNSON, A. E. & SKACH, W. R. (2011) Ligand-Driven Vectorial Folding of Ribosome-Bound Human CFTR NBD1. *Molecular Cell,* **41**, 682-692.

KIMCHI-SARFATY, C., OH, J. M., KIM, I.-W., SAUNA, Z. E., CALCAGNO, A. M., AMBUDKAR, S. V. & GOTTESMAN, M. M. (2007) A "Silent" Polymorphism in the MDR1 Gene Changes Substrate Specificity. *Science,* **315**, 525-528.

KOHLHOFF, K. J., ROBUSTELLI, P., CAVALLI, A., SALVATELLA, X. & VENDRUSCOLO, M. (2009) Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *Journal of the American Chemical Society,* **131**, 13894-13895.

KOMAR, A. A. (2009) A pause for thought along the co-translational folding pathway. *Trends in Biochemical Sciences,* **34**, 16-24.

KRIVOV, G. G., SHAPOVALOV, M. V. & DUNBRACK, R. L. (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Structure, Function, and Bioinformatics,* **77**, 778-795.

KRUGER, R., KUHN, W., MULLER, T., WOITALLA, D., GRAEBER, M., KOSEL, S., PRZUNTEK, H., EPPLEN, J. T., SCHOLS, L. & RIESS, O. (1998) AlaSOPro mutation in the gene

encoding [alpha]-synuclein in Parkinson's disease. *Nat Genet,* **18**, 106-108.

KUHAR, I., VAN PUTTEN, J. P. M., ŽGUR-BERTOK, D., GAASTRA, W. & JORDI, B. J. A. M. (2001) Codon-usage based regulation of colicin K synthesis by the stress alarmone ppGpp. *Molecular Microbiology,* **41**, 207-216.

LAKSHMIPATHY, S. K., TOMIC, S., KAISER, C. M., CHANG, H.-C., GENEVAUX, P., GEORGOPOULOS, C., BARRAL, J. M., JOHNSON, A. E., HARTL, F. U. & ETCHELLS, S. A. (2007) Identification of Nascent Chain Interaction Sites on Trigger Factor. *Journal of Biological Chemistry,* **282**, 12186-12193.

LEVITT, M. H. (2008) *Spin Dynamics: Basics of Nuclear Magnetic Resonance Second Edition,* Chichester, John Wiley & Sons Ltd.

LIPARI, G. & SZABO, A. (1982) Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *Journal of the American Chemical Society,* **104**, 4546-4559.

LOKSZTEJN, A., SCHOLL, Z. & MARSZALEK, P. E. (2012) Atomic force microscopy captures folded ribosome bound nascent chains. *Chemical Communications,* **48**, 11727-11729.

LOVELL, S. C., DAVIS, I. W., ARENDALL, W. B., DE BAKKER, P. I. W., WORD, J. M., PRISANT, M. G., RICHARDSON, J. S. & RICHARDSON, D. C. (2003) Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. *Proteins: Structure, Function, and Bioinformatics,* **50**, 437-450.

MATSUURA, T., YANAGIDA, H., USHIODA, J., URABE, I. & YOMO, T. (2007) Nascent chain, mRNA, and ribosome complexes generated by a pure translation system. *Biochemical and Biophysical Research Communications,* **352**, 372-377.

MCCOY, A. J., FUCINI, P., NOEGEL, A. A. & STEWART, M. (1999) Structural basis for dimerization of the Dictyostelium gelation factor (ABP120) rod. *Nat Struct Mol Biol,* **6**, 836-841.

MEILER, J. (2003) PROSHIFT: Protein chemical shift prediction using artificial neural networks. *Journal of Biomolecular NMR,* **26**, 25-37.

MITRA, K., SCHAFFITZEL, C., SHAIKH, T., TAMA, F., JENNI, S., BROOKS, C. L., BAN, N. & FRANK, J. (2005) Structure of the E. coli protein-conducting channel bound to a translating ribosome. *Nature,* **438**, 318-324.

MOYNA, G., ZAUHAR, R. J., WILLIAMS, H. J., NACHMAN, R. J. & SCOTT, A. I. (1998) Comparison of Ring Current Methods for Use in Molecular Modeling Refinement of NMR Derived Three-Dimensional Structures. *Journal of Chemical Information and Computer Sciences,* **38**, 702-709.

MUNRO, J. B., SANBONMATSU, K. Y., SPAHN, C. M. T. & BLANCHARD, S. C. (2009) Navigating the ribosome's metastable energy landscape. *Trends in Biochemical Sciences,* **34**, 390-400.

MUTO, H., NAKATOGAWA, H. & ITO, K. (2006) Genetically Encoded but Nonpolypeptide Prolyl-tRNA Functions in the A Site for SecM-Mediated Ribosomal Stall. *Molecular Cell,* **22**, 545-552.

NAKATOGAWA, H. & ITO, K. (2002) The Ribosomal Exit Tunnel Functions as a Discriminating Gate. *Cell,* **108**, 629-636.

NEAL, S., NIP, A., ZHANG, H. & WISHART, D. (2003) Rapid and accurate calculation of protein ^{1}H, ^{13}C and ^{15}N chemical shifts. *Journal of Biomolecular NMR,* **26**, 215-240.

NICOLA, A. V., CHEN, W. & HELENIUS, A. (1999) Co-translational folding of an alphavirus capsid protein in the cytosol of living cells. *Nat Cell Biol,* **1**, 341-345.

NOEGEL, A. A., RAPP, S., LOTTSPEICH, F., SCHLEICHER, M. & STEWART, M. (1989) The Dictyostelium gelation factor shares a putative actin binding site with alpha-actinins and dystrophin and also has a rod domain containing six 100-residue motifs that appear to have a cross-beta conformation. *The Journal of Cell Biology,* **109**, 607-618.

O'BRIEN, E. P., HSU, S.-T. D., CHRISTODOULOU, J., VENDRUSCOLO, M. & DOBSON, C. M. (2010) Transient Tertiary Structure Formation within the Ribosome Exit Port. *Journal of the American Chemical Society,* **132**, 16928-16937.

OGLE, J. M., BRODERSEN, D. E., CLEMONS, W. M., TARRY, M. J., CARTER, A. P. & RAMAKRISHNAN, V. (2001) Recognition of Cognate Transfer RNA by the 30S Ribosomal Subunit. *Science,* **292**, 897-902.

OLDFIELD, E. (2002) Chemical Shifts in Amino Acids, Peptides and Proteins: From Quantum Chemistry to Drug Design. *Annual Review of Physical Chemistry,* **53**, 349-378.

ONUFRIEV, A., BASHFORD, D. & CASE, D. A. (2004) Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Structure, Function, and Bioinformatics,* **55**, 383-394.

ORTEGA, A., AMORÓS, D. & GARCÍA DE LA TORRE, J. (2011) Prediction of Hydrodynamic and Other Solution Properties of Rigid Proteins from Atomic- and Residue-Level Models. *Biophysical Journal,* **101**, 892-898.

OSAPAY, K. & CASE, D. A. (1991) A new analysis of proton chemical shifts in proteins. *Journal of the American Chemical Society,* **113**, 9436-9444.

PECH, M., SPRETER, T., BECKMANN, R. & BEATRIX, B. (2010) Dual Binding Mode of the Nascent Polypeptide-associated Complex Reveals a Novel Universal Adapter Site on the Ribosome. *Journal of Biological Chemistry,* **285**, 19679-19687.

PETRONE, P. M., SNOW, C. D., LUCENT, D. & PANDE, V. S. (2008) Side-chain recognition and gating in the ribosome exit tunnel. *Proceedings of the National Academy of Sciences,* **105**, 16549-16554.

POLYMEROPOULOS, M. H., LAVEDAN, C., LEROY, E., IDE, S. E., DEHEJIA, A., DUTRA, A., PIKE, B., ROOT, H., RUBENSTEIN, J., BOYER, R., STENROOS, E. S., CHANDRASEKHARAPPA, S., ATHANASSIADOU, A., PAPAPETROPOULOS, T., JOHNSON, W. G., LAZZARINI, A. M., DUVOISIN, R. C., DI IORIO, G., GOLBE, L. I. & NUSSBAUM, R. L. (1997) Mutation in the {alpha}-Synuclein Gene Identified in Families with Parkinson's Disease. *Science,* **276**, 2045-2047.

POPLE, J. A. (1958) Molecular orbital theory of aromatic ring currents. *Molecular Physics,* **1**, 175-180.

PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T. & FLANNERY, B. P. (2007) *Chapter 17: Integration of Ordinary Differential Equations.* in *Numerical Recipes: The Art of Scientific Computing 901-910.* 3rd ed. Cambridge University Press, Cambridge.

PRONK, S., PÁLL, S., SCHULZ, R., LARSSON, P., BJELKMAR, P., APOSTOLOV, R., SHIRTS, M. R., SMITH, J. C., KASSON, P. M., VAN DER SPOEL, D., HESS, B. & LINDAHL, E. (2013) GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*.

PURVIS, I. J., BETTANY, A. J. E., SANTIAGO, T. C., COGGINS, J. R., DUNCAN, K., EASON, R. & BROWN, A. J. P. (1987) The efficiency of folding of some proteins is

increased by controlled rates of translation in vivo: A hypothesis. *Journal of Molecular Biology,* **193**, 413-417.

QIU, L., PABIT, S. A., ROITBERG, A. E. & HAGEN, S. J. (2002) Smaller and Faster: The 20-Residue Trp-Cage Protein Folds in 4μs. *Journal of the American Chemical Society,* **124**, 12952-12953.

RAMAKRISHNAN, V. (2002) Ribosome Structure and the Mechanism of Translation. *Cell,* **108**, 557.

RIVERS, R. C., KUMITA, J. R., TARTAGLIA, G. G., DEDMON, M. M., PAWAR, A., VENDRUSCOLO, M., DOBSON, C. M. & CHRISTODOULOU, J. (2008) Molecular determinants of the aggregation behavior of alpha- and beta-synuclein. *Protein Science,* **17**, 887-898.

ROBUSTELLI, P., KOHLHOFF, K., CAVALLI, A. & VENDRUSCOLO, M. (2010) Using NMR Chemical Shifts as Structural Restraints in Molecular Dynamics Simulations of Proteins. *Structure,* **18**, 923-933.

ROSATO, A., ARAMINI, J. M., ARROWSMITH, C., BAGARIA, A., BAKER, D., CAVALLI, A., DORELEIJERS, J. F., ELETSKY, A., GIACHETTI, A., GUERRY, P., GUTMANAS, A., G¸NTERT, P., HE, Y., HERRMANN, T., HUANG, Y. J., JARAVINE, V., JONKER, H. R. A., KENNEDY, M. A., LANGE, O. F., LIU, G., MALLIAVIN, T. E., MANI, R., MAO, B., MONTELIONE, G. T., NILGES, M., ROSSI, P., VAN†DER†SCHOT, G., SCHWALBE, H., SZYPERSKI, T. A., VENDRUSCOLO, M., VERNON, R., VRANKEN, W. F., DE†VRIES, S., VUISTER, G. W., WU, B., YANG, Y. & BONVIN, A. M. J. J. (2012) Blind Testing of Routine, Fully Automated Determination of Protein Structures from NMR Data. *Structure (London, England : 1993),* **20**, 227-236.

RUDOLPH, R. & LILIE, H. (1996) In vitro folding of inclusion body proteins. *The FASEB Journal,* **10**, 49-56.

RUTH, R. D. (1983) A Canonical Integration Technique. *Nuclear Science, IEEE Transactions on,* **30**, 2669-2671.

RUTKOWSKA, A., BEERBAUM, M., RAJAGOPALAN, N., FIAUX, J., SCHMIEDER, P., KRAMER, G. N., OSCHKINAT, H. & BUKAU, B. (2009) Large-scale purification of ribosome-nascent chain complexes for biochemical and structural studies. *FEBS Letters,* **583**, 2407-2413.

SCHAFFITZEL, C. & BAN, N. (2007) Generation of ribosome nascent chain complexes for structural and functional studies. *Journal of Structural Biology,* **158**, 463-471.

SCHANDA, P., KUPČE, Ē. & BRUTSCHER, B. (2005) SOFAST-HMQC Experiments for Recording Two-dimensional Deteronuclear Correlation Spectra of Proteins within a Few Seconds. *Journal of Biomolecular NMR,* **33**, 199-211.

SCHIPPERS, J. H. M. & MUELLER-ROEBER, B. (2010) Ribosomal composition and control of leaf development. *Plant Science,* **179**, 307-315.

SCHMEING, T. M. & RAMAKRISHNAN, V. (2009) What recent ribosome structures have revealed about the mechanism of translation. *Nature,* **461**, 1234-1242.

SCHOFIELD, P. (1973) Computer simulation studies of the liquid state. *Computer Physics Communications,* **5**, 17-23.

SCHUWIRTH, B. S., BOROVINSKAYA, M. A., HAU, C. W., ZHANG, W., VILA-SANJURJO, A., HOLTON, J. M. & CATE, J. H. D. (2005) Structures of the Bacterial Ribosome at 3.5 Å Resolution. *Science,* **310**, 827-834.

SCHWAIGER, I., KARDINAL, A., SCHLEICHER, M., NOEGEL, A. A. & RIEF, M. (2004) A mechanical unfolding intermediate in an actin-crosslinking protein. *Nat Struct Mol Biol,* **11**, 81-85.

SCHWAIGER, I., SCHLEICHER, M., NOEGEL, A. A. & RIEF, M. (2005) The folding pathway of a fast-folding immunoglobulin domain revealed by single-molecule mechanical experiments. *EMBO Rep,* **6**, 46-51.

SEIDELT, B., INNIS, C. A., WILSON, D. N., GARTMANN, M., ARMACHE, J.-P., VILLA, E., TRABUCO, L. G., BECKER, T., MIELKE, T., SCHULTEN, K., STEITZ, T. A. & BECKMANN, R. (2009) Structural Insight into Nascent Polypeptide Chain-Mediated Translational Stalling. *Science,* **326**, 1412-1415.

SELMER, M., DUNHAM, C. M., MURPHY, F. V., WEIXLBAUMER, A., PETRY, S., KELLEY, A. C., WEIR, J. R. & RAMAKRISHNAN, V. (2006) Structure of the 70S Ribosome Complexed with mRNA and tRNA. *Science,* **313**, 1935-1942.

SHEN, Y. & BAX, A. (2010) SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *Journal of Biomolecular NMR,* **48**, 13-22.

SHEN, Y., VERNON, R., BAKER, D. & BAX, A. (2009) De novo protein structure generation from incomplete chemical shift assignments. *Journal of Biomolecular NMR,* **43**, 63-78.

SOSNICK, T. R. & BARRICK, D. (2011) The folding of single domain proteins - have we reached a consensus? *Current Opinion in Structural Biology,* **21**, 12-24.

SPILLANTINI, M. G., CROWTHER, R. A., JAKES, R., HASEGAWA, M. & GOEDERT, M. (1998) Î±-Synuclein in filamentous inclusions of Lewy bodies from Parkinson's disease and dementia with Lewy bodies. *Proceedings of the National Academy of Sciences of the United States of America,* **95**, 6469-6473.

STUDIER, F. W. & MOFFATT, B. A. (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *Journal of Molecular Biology,* **189**, 113-130.

SWOPE, W. C., ANDERSEN, H. C., BERENS, P. H. & WILSON, K. R. (1982) A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics,* **76**, 637-649.

TAKAHASHI, S., IIDA, M., FURUSAWA, H., SHIMIZU, Y., UEDA, T. & OKAHATA, Y. (2009) Real-Time Monitoring of Cell-Free Translation on a Quartz-Crystal Microbalance. *Journal of the American Chemical Society,* **131**, 9326-9332.

TSAI, C.-J., SAUNA, Z. E., KIMCHI-SARFATY, C., AMBUDKAR, S. V., GOTTESMAN, M. M. & NUSSINOV, R. (2008) Synonymous Mutations and Ribosome Stalling Can Lead to Altered Folding Pathways and Distinct Minima. *Journal of Molecular Biology,* **383**, 281-291.

TSALKOVA, T., ODOM, O. W., KRAMER, G. & HARDESTY, B. (1998) Different conformations of nascent peptides on ribosomes. *Journal of Molecular Biology,* **278**, 713-723.

UVERSKY, V. N. (2007) Neuropathology, biochemistry, and biophysics of α-synuclein aggregation. *Journal of Neurochemistry,* **103**, 17-37.

VAN DEN BERG, B., ELLIS, R. J. & DOBSON, C. M. (1999) Effects of macromolecular crowding on protein folding and aggregation. *EMBO J,* **18**, 6927-6933.

VÁZQUEZ-LASLOP, N., RAMU, H., KLEPACKI, D., KANNAN, K. & MANKIN, A. S. (2010) The key function of a conserved and modified rRNA residue in the ribosomal response to the nascent peptide. *EMBO Journal,* **29**, 3108-3117.

VERLET, L. (1967) Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review,* **159**, 98-103.

VILA, J. A., ARNAUTOVA, Y. A., MARTIN, O. A. & SCHERAGA, H. A. (2009) Quantum-mechanics-derived 13CŒ± chemical shift server (CheShift) for protein structure validation. *Proceedings of the National Academy of Sciences,* **106**, 16972-16977.

VILLALI, J. & KERN, D. (2010) Choreographing an enzyme's dance. *Current Opinion in Chemical Biology,* **14**, 636-643.

WAUDBY, C. A., LAUNAY, H. L. N., CABRITA, L. D. & CHRISTODOULOU, J. (2009) Protein folding on the ribosome studied using NMR spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy*.

WHITFORD, P. C., GEGGIER, P., ALTMAN, R. B., BLANCHARD, S. C., ONUCHIC, J. N. & SANBONMATSU, K. Y. (2010) Accommodation of aminoacyl-tRNA into the ribosome involves reversible excursions along multiple pathways. *RNA,* **16**, 1196-1204.

WISHART, D. S., ARNDT, D., BERJANSKII, M., TANG, P., ZHOU, J. & LIN, G. (2008) CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. *Nucleic Acids Research,* **36**, W496-W502.

WISHART, D. S. & CASE, D. A. (2002) *Use of chemical shifts in macromolecular structure determination.* in THOMAS L. JAMES, V. D. U. S. (Ed.) *Methods in Enzymology 3-34.* Academic Press,

XU, X.-P. & CASE, D. (2001) Automated prediction of 15N, 13C α, 13C β and 13C′ chemical shifts in proteins using a density functional database. *Journal of Biomolecular NMR,* **21**, 321-333.

YOKOTA, O., TERADA, S., ISHIZU, H., UJIKE, H., ISHIHARA, T., NAKASHIMA, H., YASUDA, M., KITAMURA, Y., UÉDA, K., CHECLER, F. & KURODA, S. (2002) NACP/a-Synuclein, NAC, and ß-amyloid pathology of familial Alzheimer's disease with the

E184D presenilin-1 mutation: a clinicopathological study of two autopsy cases. *Acta Neuropathologica,* **104**, 637-648.

YUSUPOV, M. M., YUSUPOVA, G. Z., BAUCOM, A., LIEBERMAN, K., EARNEST, T. N., CATE, J. H. D. & NOLLER, H. F. (2001) Crystal Structure of the Ribosome at 5.5 √Ö Resolution. *Science,* **292**, 883-896.

ZARRANZ, J. J., ALEGRE, J., GÓMEZ-ESTEBAN, J. C., LEZCANO, E., ROS, R., AMPUERO, I., VIDAL, L., HOENICKA, J., RODRIGUEZ, O., ATARÉS, B., LLORENS, V., TORTOSA, E. G., SER, T. D., MUÑOZ, D. G. & YEBENES, J. G. D. (2004) The new mutation, E46K, of alpha-synuclein causes parkinson and Lewy body dementia. *Annals of Neurology,* **55**, 164-173.

ZAVIALOV, A. V., BUCKINGHAM, R. H. & EHRENBERG, M. N. (2001) A Posttermination Ribosomal Complex Is the Guanine Nucleotide Exchange Factor for Peptide Release Factor RF3. *Cell,* **107**, 115-124.

ZHANG, G., HUBALEWSKA, M. & IGNATOVA, Z. (2009) Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature Structural & Molecular Biology,* **16**, 274-280.