Chapter 1

Introduction

1.1 OVERVIEW

Proteins play diverse roles in living cells and constitute more than half of their dry weight. No living part of any organism is completely devoid of protein. Proteins have been found to participate in almost every aspect of cellular activity and perform their respective tasks with incredible speed and accuracy.

A central dogma of cell biology is that genetic information propagates through DNA, RNA and protein. Genetic information coded in DNA is first transferred to RNA by transcription and then to protein via translation. In the translation process, twenty different amino acids, each with distinct chemical characteristics, are charged on respective tRNAs, which then deliver these individual building blocks to ribosomes bound to mRNA. The whole machinery, with the help of numerous cofactors, links up the individual amino acids into a polymeric peptide chain. The variety of polypeptide lengths and compositions allows for enormous versatility in the chemical properties of different proteins, and it presumably explains why evolution has selected proteins rather than RNA molecules to catalyze most cellular reactions, though RNA has been speculated to be life's most ancient molecule.

Although a minimal model for protein synthesis was put forth in the early 1960s (Watson, 1964), a detailed mechanistic understanding of translation has not become possible until recently. The publication of the crystal structures of the 50S and 30S ribosomal subunits and the intact 70S ribosome in the last a couple of years has greatly advanced our knowledge about this vital biosynthesis process at the atomic level (reviewed by Ramakrishnan, 2002). Translation proceeds through three steps: initiation, elongation and termination, with numerous factors involved in each step (reviewed by Green and Noller, 1997). Initiation and elongation have been extensively studied ever since the deciphering of the genetic code nearly fifty years ago (for reviews see, Kozak, 1999; Pestova and Hellen, 2000; Rodnina *et al.,* 1999). However, the last step, termination, had remained out of reach for many years. The publication of the structures of several release factors (RFs) in the last few years provides insights into the important reactions in the termination process (reviewed by Connell and Nierhaus, 2000). A tripeptide within bacterial RFs for the stop codon recognition, which is referred to as a tripeptide "anticodon" or a tripeptide discriminator, has been recently identified (reviewed by Nakamura and Ito, 2002).

All studies support the proposal that the process of termination begins when a stop codon on mRNA is encountered in the A site. A stop codon is absolutely required for releasing a polypeptide from the translational complex (reviewed by Kisselev and Buckingham, 2000). However, what if mRNAs are truncated at their 3'-ends such that no in-frame stop codon is available? This situation will lead to the synthesis of a faulty protein. Although prokaryotes and

eukaryotes use a similar strategy for mRNA surveillance (Bhardwaj and Williams, 2002), the problems caused by a missing termination codon seem more deleterious in prokaryotes than in eukaryotes. In eukaryotes, transcription and translation are separated in space and time; transcription occurs in the nucleus, and translation happens in the cytoplasm. The initiation of translation in eukaryotes is regulated by many *cis*-regulatory elements including the 3'-UTR and the poly-A tail of the message, which can control gene expression by directly enhancing translation initiation (reviewed by Scorilas, 2002) as well as by affecting the localization and stability of mRNAs (reviewed by Decker and Parker, 1995). Therefore, the truncation of stop codons in eukaryotic mRNAs, which leads to the missing of the 3'-UTR and the poly-A tail, will block the export (from the nucleus to the cytoplasm) and the localization of mRNAs, destabilize messages, and eventually repress translation initiation. In contrast, prokaryotic translation is usually coupled with transcription (pre-mRNA splicing and editing are not needed), with fewer protein factors required and less extensive control used for the initiation. This leaves much room for multiple ribosomes stalling on a problematic mRNA. To deal with this circumstance, prokaryotes have developed a unique system, the tmRNA-SmpB quality control system, which was identified about ten years ago and has been characterized through various biochemical methods. The work described within this dissertation focuses on the structural properties of this remarkable RNP complex, and reports the first atomic resolution structure of one of its protein components, SmpB.

1.2 THE TMRNA-SMPB QUALITY-CONTROL SYSTEM IN PROKARYOTES

Due to various reasons, broken or improperly folded proteins are present in all living cells and may lead to a variety of deleterious consequences. To cope with these problematic proteins or selectively destroy those normal but unwanted intracellular proteins, different organisms have developed different strategies. A protein termed ubiquitin (Ub) is present in all eukaryotic cells and plays a vital role in cell metabolism. Ubiquitin is a highly conserved, ~76-amino acid protein that targets eukaryotic proteins for degradation by covalently attaching to certain amino acids (reviewed by Hershko and Ciechanover, 1998). This process is the major controlled proteolysis pathway in eukaryotic cells and is responsible for the regular turnover of a wide variety of proteins and for the regulation of eukaryotic messenger RNA synthesis (reviewed by Conaway *et al.*, 2002).

Similar to the ubiquitin system in eukaryotes, in prokaryotes there is also a protein tagging system. But, instead of covalently attaching a separate signal protein to selected proteins, this tagging system translationally adds a short C-terminal peptide tag encoded by a separate message, tmRNA, to a nascent polypeptide by a so-called *trans*-translation mechanism (Keiler *et al.*, 1996). tmRNA was originally called SsrA RNA or 10Sa RNA, which was first discovered in *E. coli* when a 10S RNA fraction (Ray *et al.*, 1979) was resolved into two species, the 10Sa RNA (or SsrA, small stable RNA A) and the 10Sb RNA (the catalytic subunit of ribonuclease P) (Gurevitz *et al.*, 1983; Subbarao and Apirion, 1989). Sequence comparison and phylogenetic analysis suggested a tRNA-like structure in *Mycobacterium tuberculosis* 10Sa RNA (Tyagi and

Kinger, 1992). This finding was confirmed later in *E. coli* and other species (Komine *et al.*, 1994; Ushida *et al.*, 1994). It was found that the well-conserved 5' end and the 3' end of 10Sa RNA are arranged to a common alanine tRNA-like structure containing an amino acid-acceptor stem and a TΨC-stem/loop. 10Sa RNA is aminoacylatable with alanine *in vitro* and binds to the 70S ribosome *in vivo* (Ushida *et al.*, 1994). The discovery of an attached C-terminal peptide sequence encoded by 10Sa RNA on a recombinant protein murine interleukin-6 (IL-6) unveiled the remarkable mRNA function of 10Sa RNA (Tu *et al.*, 1995). From then on, a new name, tmRNA (transfer-messenger RNA), has become popular for this type of unique RNA (Figure 1.1).

A mature tmRNA typically comprises ~360 nucleotides and is generated from its precursor by three tRNA-specific processing enzymes: ribonuclease P (Komine *et al.*, 1994), ribonuclease III (Srivastava *et al.*, 1992) and ribonuclease E (Lin-Chao *et al.*, 1999). Although tmRNA exists as a single piece in most species, two-piece tmRNA has been identified in some species (Keiler *et al.*, 2000). Many early studies focusing on the biological roles of tmRNA showed that tmRNA is not essential for cell growth, but the presence of tmRNA provides intrinsic advantages for cell survival. Disruption of the 10Sa gene has been observed to lead to several subtle phenotypes, including temperature-sensitive growth, reduced motility, inability to support growth of λ*imm*P22 hybrid phage, induction of Alp protease activity and enhanced activity of several repressor proteins (Kirby *et al.*, 1994; Komine *et al.*, 1994; Retallack and Friedman, 1995;

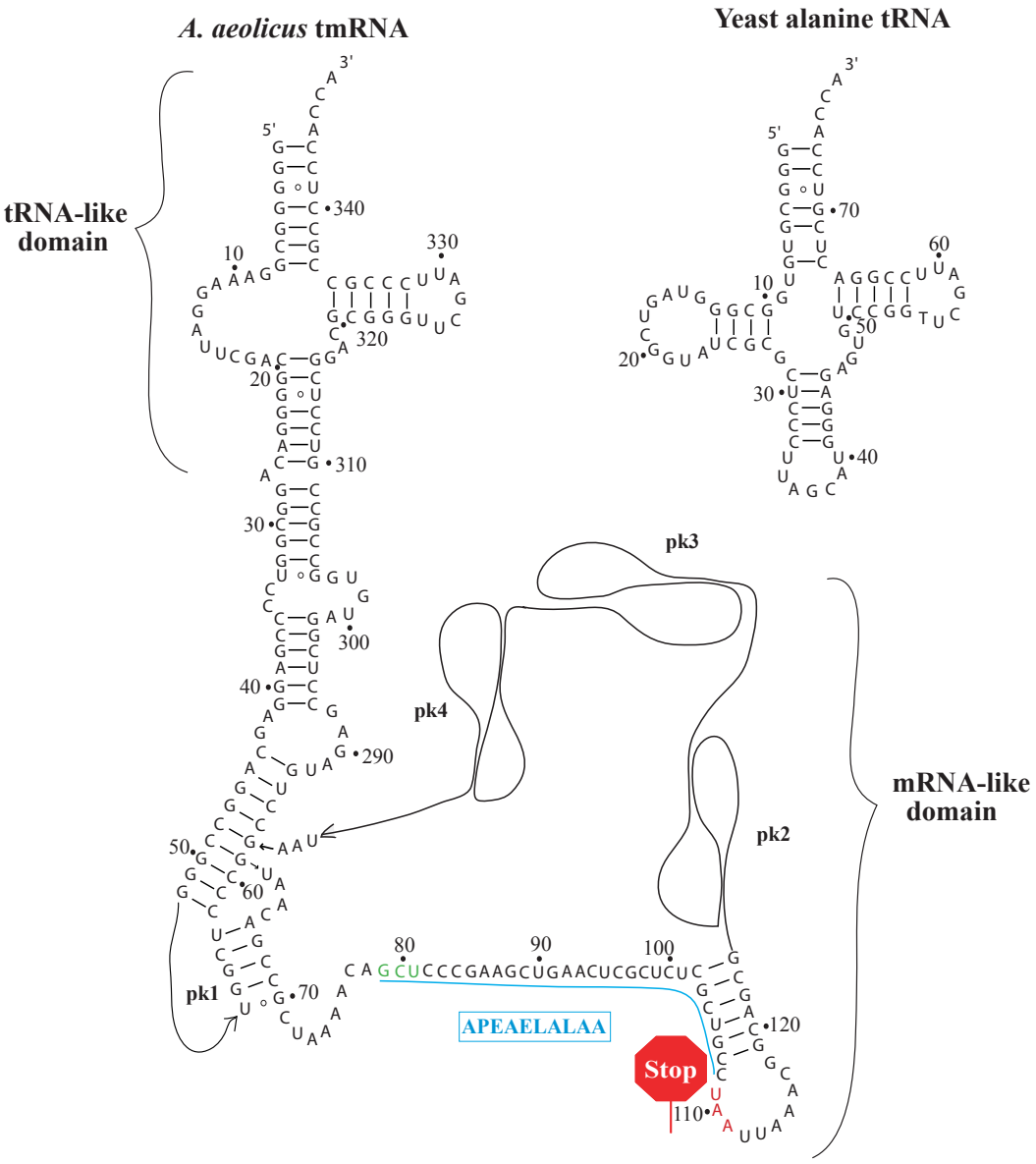

Figure 1.1 The secondary structure of *A. aeolicus* tmRNA. tRNA-like domain and coding sequence shown were derived from the *E. coli* standard tmRNA(Williams and Bartel, 1996) based on phylogenetic analysis and sequence comparison. Underlined are the coding sequence of the tmRNA-like domain; boxed are the predicted proteolysis tag sequence; colored in green is the resume codon. The three pseudoknots (pk2, pk3 and pk4) in the mRNA-like domain are sketched in single lines. Note that the strongly conserved triplet UA(A/G) is not present in *A. aeolicus* tmRNA two nucleotides upstream of the resume codon.

Retallack *et al*., 1994). However, the mechanism of the C-terminal tagging pathway and its physiological roles in cells remained obscure at that time.

The explicit biological roles of tmRNA and the mechanism of how it functions in cells were first reported by Keiler *et al.* (1996). Their studies indicated that the added non-polar destabilizing tag is essential for the degradation of the tagged proteins by C-terminal specific proteases present in both the cytoplasm and periplasm of *E. coli,* and this tagging process occurs when the mRNA templates lack stop codons (Keiler *et al.*, 1996). Recent studies support the proposal that tmRNA has not only dual activities (tRNA and mRNA), but also dual functions: one is to rescue ribosomes stalled on mRNA templates; the other is to remove proteins resulting from defective mRNAs. The release of stalled ribosomes has been found to be the primary function of the tmRNA-SmpB quality control system based on the studies of how charging of tmRNA and tagged protein degradation influence λ*imm*P22 growth in *E. coli* (Withey and Friedman, 1999). In eukaryotes, there is also a process to release the stalled ribosomes, which is called the Ski7p/exosome-mediated nonstop decay, but the mechanism of it is completely different from that of the *trans*-translation process in prokaryotes (Frischmeyer *et al.*, 2002; van Hoof *et al.*, 2002).

The *trans*-translation model, which explains how tmRNA functions in cells, is shown in Figure 1.2. In this process, first, the $tmRNA^{Ala}$-SmpB complex (SmpB: small protein B, an essential protein component) recognizes a stalled ribosome on an incomplete or untranslatable mRNA and enters its empty A-site. Second, the alanine charged on the tRNA-like domain of tmRNA is transferred to

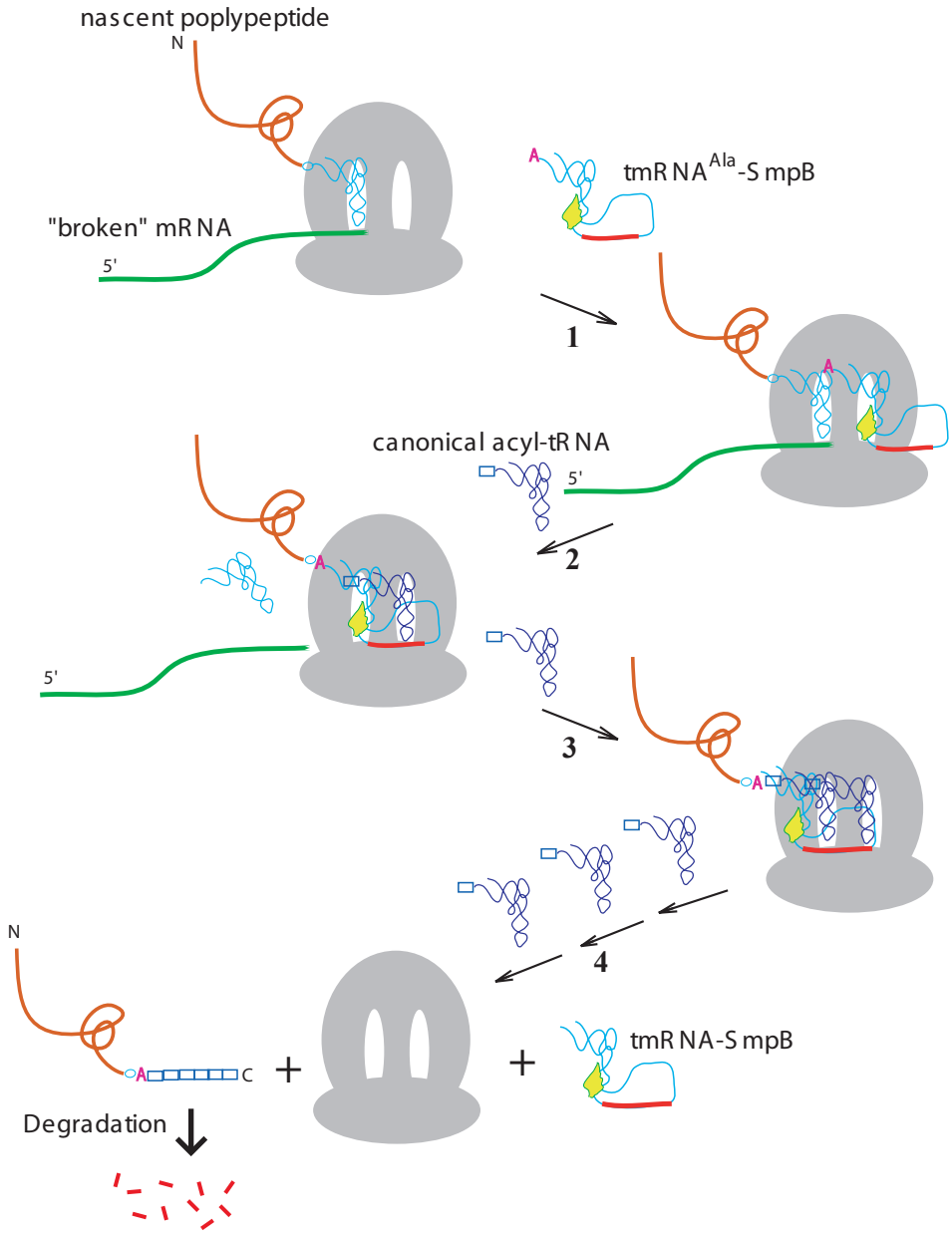

Figure 1.2 The *trans*-translation model of the tmRNA-SmpB quality control system. (1) tmRNA^{Ala}-SmpB complex recognizes and binds to the empty A-site of the stalled ribosome caused by the missing of in-frame stop-codon on the mRNA. (2) Ribosome switches from the "broken" mRNA to the coding sequence of tmRNA (colored in red). (3) Regular translation resumes on tmRNA. (4) Elongation and termination result in a released ribosome, a tmRNA-SmpB complex, and a tagged poplypeptide that will be directed for degradation.

the nascent polypeptide chain and the original mRNA template is released from the ribosome in a message-switching event. Third, regular translation resumes on the internal ORF of tmRNA and elongation proceeds until a stop codon is reached. Finally, translation terminates and the ribosome is recycled along with the release of the uncharged tmRNA-SmpB complex and the C-terminal tagged protein, which is targeted for degradation by a variety of ATP-dependent cellular proteases (Gottesman *et al.*, 1998; Wiegert and Schumann, 2001). Multiple sites in the tag sequence recognized by diverse proteases have been found. These proteases work in concert to modulate proteolysis (Flynn *et al.*, 2001).

In the original model, it was postulated that *trans*-translation serves as a mechanism for destroying truncated proteins produced from damaged mRNAs (Keiler *et al.*, 1996). Recent available data indicate that this *trans*-translation process also occurs in other cases including internal rare codons (Roche and Sauer, 1999), inefficient termination codons (Hayes *et al.*, 2002a and 2002b), interference from translation of downstream reading frames (Roche and Sauer, 2001), and lack of polypeptide release factors. All these situations cause the same result - ribosomes stall or pause on a message, which appears to be necessary and sufficient for the initiation of the *trans*-translation activity. Other than tagging polypeptides resulting from defective messages, the tmRNA-SmpB quality control system has been found to target natural proteins such as the lacI mRNA encoding Lac repressor (LacI) for degradation to assist cells to adapt to lactose availability by supporting a rapid induction of lac operon expression (Abo *et al.*, 2000).

A highly conserved triplet UA(A/G) (normally recognized as a stop codon by release factor-1) has been found to be present two nucleotides upstream of the resume codon in most tmRNAs (Williams *et al.,* 1999). Published data suggest that tmRNA interacts with the tRNA that decodes the resume codon prior to entering the ribosome (Gillet and Felden, 2001).

Small protein B (SmpB) is a unique, highly conserved RNA-binding protein present in all eubacteria sequenced so far. It consists of ~160 amino acids and has no significant homology to any other known proteins except for SmpB proteins from different species. This protein has been found to be an essential component of the tmRNA quality-control system. The SmpB-deletion mutant has the same phenotypes as the ssrA-defective mutant (Karzai *et al.,* 1999). Earlier work suggests that SmpB binds specifically to SsrA RNA with a high affinity and is required for stable association of SsrA with ribosomes *in vivo* (Karzai *et al.,* 1999). However, recent studies found that SmpB binds both tmRNA and tRNA with a similar affinity and the acceptor-T arm constitutes the primary SmpB binding site in both tmRNA and tRNA (Wower *et al.,* 2002).

Studies by several labs reveal that the tmRNA-SmpB quality control system is a large ribonucleoprotein complex that contains multiple associated proteins including ribosomal protein S1, EF-Tu, phosphoribosyl pyrophosphate synthase, RNase R and YfbG in addition to SsrA RNA and SmpB (Karzai and Sauer, 2001; Wower *et al.,* 2001), but SmpB is the only protein component known to date to be fully dedicated to the function of this prokaryotic tagging system. Karzai and Sauer (2000) proposed that RNase R might degrade the

mRNAs released in the *trans*-translation process. However, recent studies investigating the fate of incomplete mRNA upon tmRNA action demonstrate that RNase R is not responsible for the increased decay of incomplete mRNA over complete mRNA (Bhardwaj and Williams, 2002). Their work suggests that tmRNA helps prokaryotic cells remove defective mRNA in a similar manner to the nonsense-mediated decay (NMD) in eukaryotes.

In summary, the tmRNA-SmpB system is a unique quality control system preserved throughout the bacterial kingdom. Although it was first identified only ten years ago, research on the function of tmRNA has rapidly progressed in the last several years. The *trans*-translation model has survived diverse experimental tests and become firmly established. It has become clear that the tmRNA-SmpB system has the dual functions of releasing stalled ribosomes from defective mRNAs and of targeting proteins for subsequent degradation by multiple ATP-dependent proteases. Both of these functions play an important role in cell metabolism, but the release of stalled ribosomes is believed to be the primary function of *trans*-translation. However, it should be noticed that phenotype studies of deletion and/or disruption of tmRNA/SmpB genes have been carried out only on three out of the nineteen bacterial phyla (reviewed by Gillet and Felden, 2001b), and although a general model of tmRNA function has been proposed (Keiler *et al*., 1996; Karzai *et al*., 2000), the structure-based mechanism of the *trans*-translation activity and its function in cell growth still remain obscure. The structural model of SmpB described in this dissertation lifts a corner of the

covering blanket and provides the basis for further structural characterization of the whole complex.

1.3 STRUCTURAL STUDY OF PROTEINS BY NMR

Nuclear magnetic resonance (NMR) spectroscopy is one of the techniques that allow structural characterization of molecules at the atomic level. NMR techniques take advantage of the uneven distribution of the two magnetic spin states of nuclei that have a non-zero spin quantum number (I). The nuclei commonly used in biochemical NMR studies usually have a spin quantum number of ½, including ^{1}H, ^{13}C, ^{15}N, ^{19}F and ^{31}P. These nuclei have a single unpaired spin and therefore possess a net nuclear magnetic moment. NMR is concerned with the transitions between the two energy levels (m = ½ and -½) of a nucleus in an applied external magnetic field. The transition energy is proportional to the strength of the magnetic field (Figure 1.3).

Initially, NMR was performed by applying monochromatic radiation with continuous frequency to a sample and locating its absorption maxima. This method is called continuous wave (CW) NMR. A similar strategy is still in common use in optical spectroscopy such as ultra-violet (UV) absorption measurement. However, the inherent insensitivity of NMR spectroscopy, which is rooted in the small population difference between the low and high energy states of non-zero spin nuclei in a magnetic field, inevitably leads to a low signal-to-noise ratio. One way to improve the signal-to-noise ratio is signal averaging, which makes genuine signals stand out of the background by repeating the same

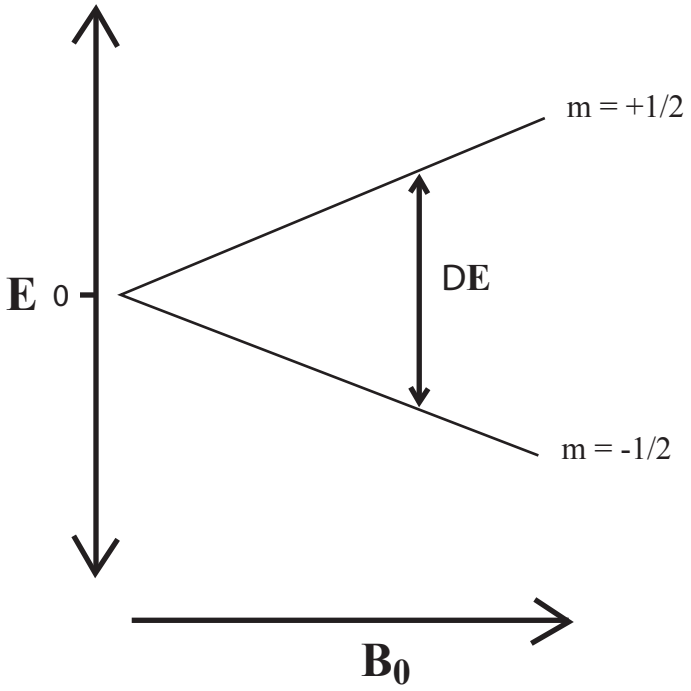

Figure 1.3 Illustration of the difference between the two energy levels of a nucleus (m = ± 1/2) in an external magnetic field. E is relative energy, DE is the transition energy, m is the magnetic spin quantum number, and B_0 is the applied magnetic field with an arrow indicating the magnetic strength increasing. The transition energy is proportional to the magnetic field strength.

experiment several times with the trade-off of the experiment speed. For recording a single 1D spectrum of a small molecule, this signal-averaging CW NMR spectroscopy is feasible. However, when the targets become biological macromolecules such as proteins or nucleic acids, the time taken using this NMR method will turn out to be intolerably long, because small frequency intervals and multidimensional spectra have to be used in order to discriminate between closely spaced signals of more severe overlap. To overcome this difficulty, a new NMR spectroscopy was developed, which allows elucidating the solution structures of molecules with a moderate size within reasonable time. This new NMR approach, which is called pulsed FT NMR (Fourier-transform Nuclear Magnetic Resonance), applies an impulse containing all characteristic frequencies needed to a sample and then measures all frequencies simultaneously instead of one after another as in CW NMR. In this way, tremendous savings in experimental time can be gained. For example, in a CW NMR experiment measuring signals over 5000 Hz spectral width with a sweep rate of 5 Hz/second requires 1000 seconds; in contrast, in a pulse NMR experiment, only 1 second is required because of the capability of measuring all frequencies in one time. The advantage of the pulse NMR method is self-evident for multidimensional data recording.

NMR spectroscopy is an important analytical technique for the determination of protein structures at the atomic resolution and currently provides the only alternative to X-ray crystallography for obtaining this information. Since the publication of the first complete solution structure of the protein BUSI IIA (Williamson *et al.* 1985), high resolution multidimensional NMR has been widely

used for determining the three-dimensional structures of biological macromolecules including proteins, nucleic acids and carbohydrates. However, the transformation of experimental data to structure is less direct in NMR spectroscopy than in X-ray crystallography. Structural studies of molecules by NMR techniques depend on successful resonance assignments and the use of nuclear Overhauser effect (NOE), which exploits the correlations between the intensity of NOE signals and the distance between protons. In general, distance measurements are limited to interacting nuclei that are within 5-6Å: the shorter the distance, the stronger the NOE signal. Figure 1.4 illustrates the common procedure for structure determination of proteins by NMR.

As in X-ray crystallography, NMR requires a milligram scale quantity of very pure protein. Before a structure determination is launched, a qualitative inspection of the 1D NMR spectrum is usually carried out to assess if the protein is unfolded, folded or aggregated (Figure 1.5). Chemical shifts of similar groups of protons show up in similar regions of the spectrum, and different proton types typically have distinct chemical shifts (Figure 1.6). 2D COSY, TOCSY and NOESY spectra can provide further information about the compactness of the structure and the approximate relative ratio of helices, strands and random coils in the structure. The latter can be of help in making decisions as to whether it is necessary or not to perform proteolysis analysis (see section 3.2). Combinations of triple resonance experiments performed using ^{13}C and ^{15}N labeled samples, such as HN(CO)CACB, HNCA, and HNCACB, can provide the unambiguous

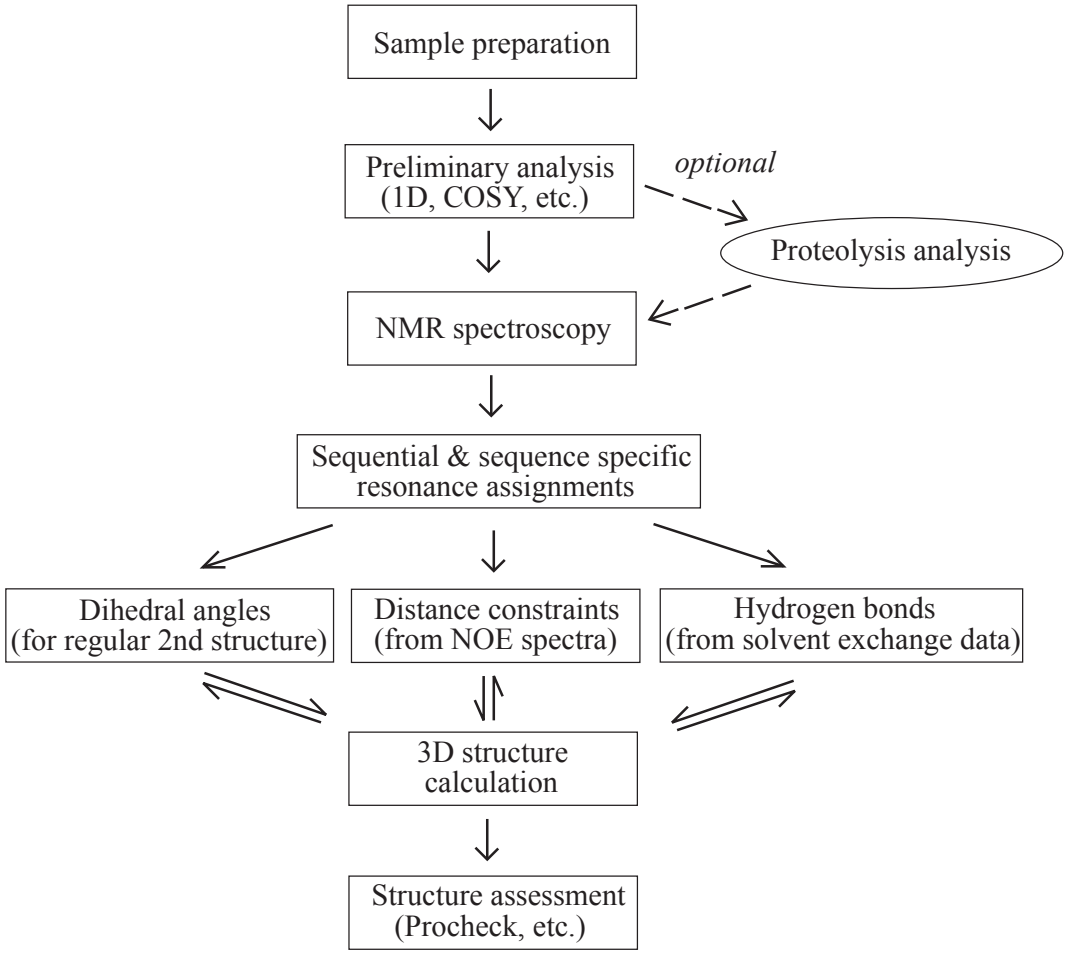

Figure 1.4 Illustration of the structural studies of macromolecules by modern NMR techniques.

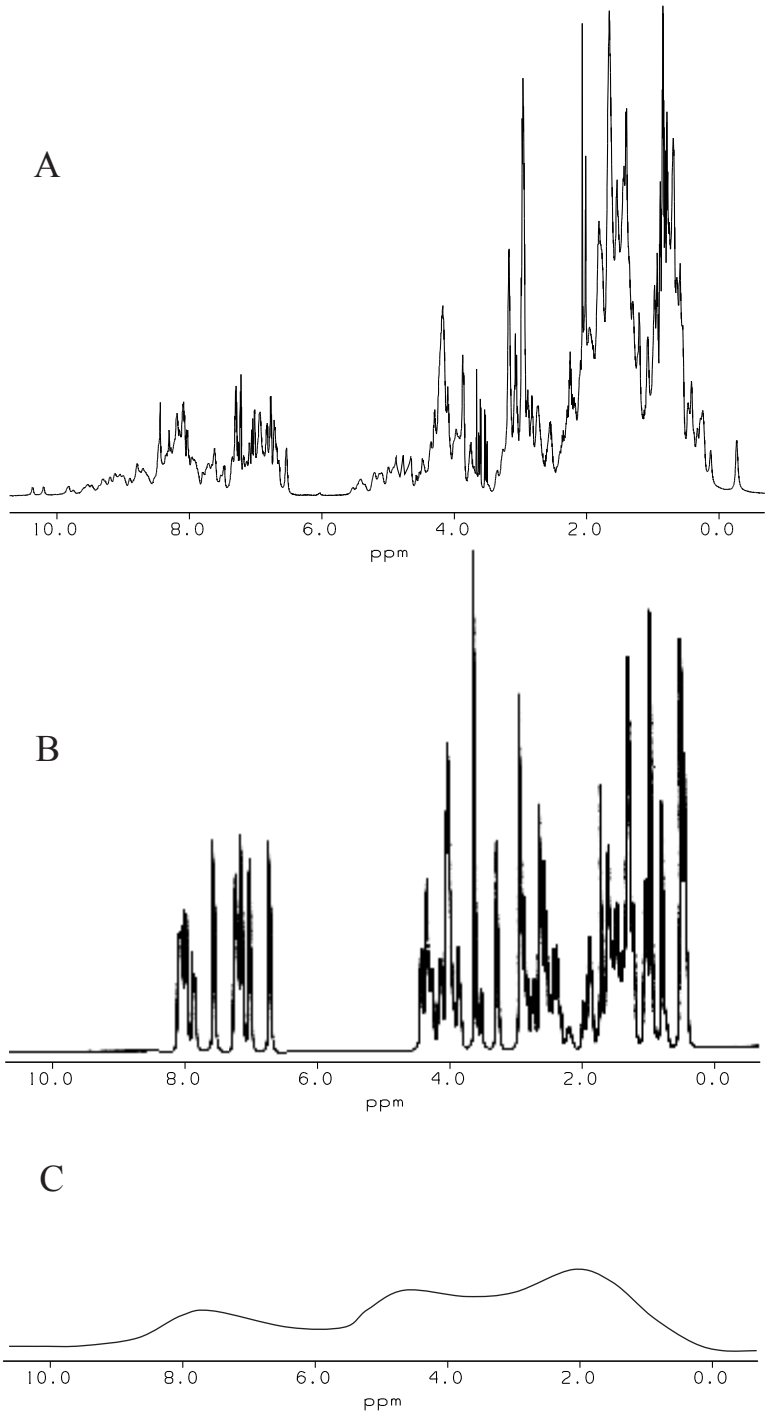

Figure 1.5 Assessment of protein folding/aggregation properties by a qualitative inspection of a 1D NMR spectrum. (A) Sharp, dispersed chemical shifts indicate that the protein is folded but not aggretated. (B) Clustered sharp lines are typical of random coils: all protons have chemical shifts similar to that of free amino acids. (C) Very broad resonances are the sign for proteins with high molecular weight or aggregations.

sequential assignment of protein backbone resonances so as to establish sequential connectivities between amino acids (Ikura *et al.,* 1990). Experiments such as HCCH-TOCSY allow complete side-chain assignment (Cavanagh *et al.*, 1996), which is of particular importance for residues constituting the hydrophobic core of the structure. The finding of the correlation between the observed chemical shifts of C^{α} and C^{β} and the backbone torsion angles of regular secondary structures (Spera and Bax, 1991) provides an additional way for identifying structure segments and for verifying the final structure models.

Over the past decade, a vast number of proteins (along with other bio-molecules) with increasing sizes have been structurally characterized using multidimensional NMR spectroscopy and advanced isotope-labeling techniques. Up to July 23, 2002, 2857 out of 18294 structures deposited in Protein Data Bank were solved by NMR techniques. However, due to the inherent insensitivity of NMR spectroscopy, a molecular weight of 25 kDa was, until recently, the upper limit that can be possibly tackled. The newly developed TROSY experiments make it possible to detect and assign the resonances originating from proteins with molecular weights greater than 30 kDa (Pervushin *et al.,* 1997; Pervushin *et al.,* 1998; Salzmann *et al.,* 1998). In addition, dipolar couplings provide a powerful tool for determining the relative orientation of domains in multi-domain proteins (Tolman *et al.,* 1995).

Besides the size limit, a traditional NMR structure determination process consists of a number of time-consuming steps, including data collection, resonance assignment, NOE signal interpretation, and structure calculation. In

many cases, resonance assignment and NOE signal interpretation are still carried out manually and empirically. The NMR structure determination of a protein with modest molecular weight may take a few months to several years, typically. The time-consuming and non-automated processes pose great challenges to the application of NMR spectroscopy in the upcoming high-throughput structural genomics activities, which aim to assign a structure to each of the remaining sequences (about two-thirds) of all genome sequences without structure available yet (Mittl and Grutter, 2001). However, substantial advances made in the last several years promise to solve this dilemma. The time needed for data acquisition can be reduced by enhancing the signal-to-noise ratio of spectrometers or by defining structure determination strategies that require only a minimal data set. The development of 900 MHz and higher field NMR spectrometers provides a direct solution to increase the sensitivity of NMR spectrometers. New cryo-probes and pre-amplification circuits have been introduced to significantly increase the signal-to-noise ratio. A suite of nine reduced-dimensionality ^{13}C, ^{15}N, ^{1}H-triple-resonance NMR experiments has been developed for rapid and complete protein resonance assignments (Szyperski *et al.,* 2002). In combination with residue-specific deuteration/protonation, data sets may be recorded within two or three weeks (Medek *et al.,* 2000). In addition, newly developed approaches for automated peak picking and ambiguous NOE assignment of multidimensional NMR spectra with strong overlap promise to substantially reduce the time required for resonance assignments and NOE data interpretation (Koradi *et al.,* 1998; Nilges *et al.,* 1997; Orekhov *et al.,* 2001). Encouragingly, Atkinson and

Saudek (2002) have successfully performed the direct structure determination of a 76-amino acid protein without specific resonance assignments, although it is largely not feasible for large molecules at the present time. All these advances in NMR instrumentation and implementation forebode that the generation of suitable biological specimens, rather than data acquisition and analysis, will soon become the limiting factor of NMR structure determination of molecules of interest.

Although it has limitations compared to X-ray crystallography, high-resolution NMR spectroscopy has its own advantages. First of all, NMR is capable of characterizing proteins that may not produce crystals suitable for investigation by X-ray diffraction. For example, producing crystals of membrane proteins is still a challenge (only ~1% of the structures deposited in the current PDB are classified as membrane proteins). The successful use of TROSY-based NMR experiments on membrane proteins (Fernandez *et al.*, 2001) opens a new door to study structures, functions and dynamics of integral membrane proteins. Secondly, NMR structures can be used as the search models for solving structures of protein crystals for which it is difficult to find suitable heavy atom derivatives (for reviews see Chen, 2001; Chen *et al.*, 2000). Thirdly, NMR can be used to study the dynamics of molecules and measure the picosecond to nanosecond timescale backbone and sidechain fluctuations in solution. Fourthly, NMR can provide an important technique for selecting well-behaved proteins and optimizing conditions for structure determination whether by NMR or X-ray crystallography. Fifthly, NMR can provide a means of identifying small ligands as well as macromolecular partners that may be essential for proper folding and

function. Sixthly, developments and implementations of new NMR technologies and strategies, such as the fully optimized package (RNAPack) for high-resolution RNA solution structure determination, make NMR spectroscopy an attractive and rapid structural tool and allows integration of atomic resolution structural information into biochemical studies of large RNA systems (Lukavsky and Puglisi, 2001). Lastly, NMR chemical shift mapping has become a popular method to identify the interface of a complex without knowing the structure of the partner molecules (see section 1.4).

In summary, NMR techniques play an important role in the characterization of biological macromolecules. Though it has its own limitations, enormous versatility and rapid developments of high-resolution NMR promise its vast merits and diverse utilities in many aspects of scientific studies.

1.4 NMR PERTURBATION

The chemical shift of each atom in a molecule directly reflects the local chemical environment of that atom. It is sensitive to small perturbations in local geometry and conformation, the anisotropy of nearby magnetic fields and local electrostatics (Oldfield, 1995), and has therefore proven to be a useful tool for monitoring the effects of ligand binding and conformational changes that occur within biological macromolecules.

Previous studies demonstrate that, by checking the changes of chemical shifts caused by ligand binding, it is possible to identify the residues involved in mutual interactions in a complex (Penkett *et al.*, 2000; Williamson *et al.*, 1997).

Especially, when the 3D structure of the biological macromolecule (usually protein) is known, mapping of the residues that undergo a binding-dependent chemical shift change or intensity perturbation onto the structure allows locating the binding site(s) on the protein (Brazin *et al.,* 2000; Emerson *et al.,* 1995). This approach is very attracting in that it bypasses the need for structure determination and resonance assignment of the ligand, and is now a popular method for studying protein-ligand interactions in solution. On one hand, chemical shift mapping analysis can provide the candidate residues for further site-directed mutagenesis studies aimed at identifying the residues responsible for the recognition and interaction (Osborne *et al.,* 1997; Shekhtman *et al.,* 2001); on the other hand, the combination of docking algorithms with NMR chemical shift perturbation analysis will provide an alternative way for investigating macromolecular protein complexes that requires less experimental time, effort and resources, and has potential applications in the large-scale structural genomics (Morelli *et al.,* 2000, 2001).

In the NMR perturbation method, only the component(s) of the complex to be examined needs to be labeled with an NMR active isotope, such as ^{15}N, ^{13}C and ^{2}H. NMR perturbation results are usually examined by detecting chemical shift changes in a simple ^{15}N-^{1}H HSQC or HMQC spectrum of the uniformly labeled protein (or other molecules) as a function of the added unlabeled ligand(s). Sometimes, NMR perturbation is also studied by comparing the structures of free and ligand-bound states of a protein (Moy *et al.,* 2000). The latter is more accurate but takes longer time, since the structure of the ligand-bound state has to

be determined also. The unlabeled ligand can be protein (Peterson and Gettins, 2001), DNA (Foster *et al.,* 1998), RNA (Hinck *et al.,* 1997), protein plus DNA (Cai *et al.,* 2001), DNA plus RNA (Katahira *et al.,* 2001), small organic molecules (Suzanne *et al.,* 1998), oligosaccharide (Jain *et al.,* 2001), etc. Among these, protein-protein interactions have been the most broadly and successfully studied.

Although it has been reported that the ligand-binding sites defined by NMR perturbation were in full agreement with that determined independently by X-ray crystallography and mutational analysis (Huang *et al.,* 1998; Song and Markley, 2001), interfaces identified using this approach, however, are not always identical to those revealed using X-ray crystallography, especially for large (>50kDa) protein-protein complexes (Gouda *et al.,* 1998). An alternative NMR method that uses a cross-saturation phenomenon in combination with TROSY detection in an optimally deuterium labeled system has been developed to more precisely define the interaction sites of large protein-protein complexes (Takahashi *et al.,* 2000).

NMR perturbation has become a popular approach for studying complex interactions in solution, but it is noteworthy that results obtained by this technique are not conclusive. On one hand, perturbations can be broadly classified as being caused by a direct interaction with the ligand, by structural changes that occur within the protein upon ligand binding, or both. Because of the latter effect, chemical shift perturbations must be interpreted carefully because ligand-induced structural changes are not necessarily related to sites of direct complex contact

(Hinck *et al.,* 1997), and differentiation of these two scenarios is very difficult at the current stage of analysis because the correlation between the chemical shifts and the protein structure are not yet absolutely understood (Nagata, *et al.,* 1999). On the other hand, the high concentrations of biological macromolecules required for these experiments (~1 mM) raise questions concerning the possibility for non-specific interactions being detected, thereby compromising the information obtained (Rajagopal *et al.*, 1997).

1.5 CONCLUDING REMARKS

In this dissertation, the NMR structure of small protein B from *Aquifex aeolicus* is presented. It is the first atomic resolution structure of this type of unique RNA-binding protein. The structure sheds light on the function of this indispensable component in the tmRNA-SmpB quality control system that is preserved in all known prokaryotes, and constitutes the basis for further analysis of the functionalities of the complex. NMR perturbation studies of different RNA/protein complexes provide some information regarding the protein-RNA interactions. Preliminary crystallographic studies of the protein crystals constitute a basis for solving the protein structure by X-ray crystallography and for investigating the structure and function of the complex.

End of Chapter 1

Chapter 2

Cloning, expression and Purification of *A. aeolicus of* SmpB

2.1 CLONING

Obtaining milligram amounts of very pure protein sample was essential for all subsequent structural studies, especially the crystallization trials and the NMR spectroscopy experiments. Thanks to the availability of the modern recombinant DNA cloning technologies, obtaining milligram scale of proteins is now routine in most cases.

E. coli expression systems that are not tightly regulated may not be suited for the overexpression of this exogenous protein, because the *E. coli* SmpB protein is expected to be expressed at a very low level and the introduction of a high level of homologous protein could be highly toxic to the cells. The system that was used for these studies is easily inducible and tightly regulated to prevent basal level expression of the protein. *Aquifex aeolicus* SmpB was chosen for these structural studies because this thermophilic version of the protein is more stable and thus increases the chance for the successful structural characterization. Studies of three-dimensional structures require that the target protein is in the native state and resist unfolding and degradation for as long as it takes to record

NMR spectra at relatively high temperature (usually 30°C or 40°C) or to grow crystals and perform diffraction experiments.

A. aeolicus genomic DNA was kindly provided by Dr. Robert Huber, who and colleagues determined the complete genome sequence of the hyperthermophilic bacterium *Aquifex aeolicus* (Deckert *et al.,* 1998). The coding sequence of the *A. aeolicus smpb* gene was amplified by PCR to introduce an *Nde* I site at the 5' end and a *Bam*H I site at the 3' end of the coding sequence. The sequences of the two primers used for PCR were P19: 5'-GAGGGAGCATATGGGCAAAAGC-3' and P14: 5'-TCTCGGATCCTCA GAGGTGTATTTTACCTTTAAAC-3', with *Nde* I and *BamH* I sites underlined respectively. The amplified DNA sequence was purified after digestion with both *Nde* I and *Bam*H I endonucleases to generate sticky ends. Meanwhile, purified pET-9a plasmid vector (Novagen), which carries the "plain" T7 promoter and a kanamycin-resistant gene, was also cut by both endonucleases, and the target fragment was purified and examined on an agrose gel. The concentrations of both the cloning fragment and the cut-vector were determined by UV absorbance.

Ligation was carried out by mixing ten fold excess of cloning fragment with one fold of cut-vector plus ligation buffer and DNA ligase, and then incubating at 25°C for 2 hours to generate the SmpB-overexpressing recombinant plasmid pSB (pET-9a *A. aeolicus* small protein B). The recombinant plasmid pSB was transformed into *E. coli* strain DH5α, a cloning host giving good plasmid yields. The amplified plasmid was extracted and underwent DNA sequencing to verify the cloned sequence.

2.2 EXPRESSION

After the cloned sequence was confirmed to be correct, the recombinant plasmid pSB was transformed into the *E. coli* expression strain BL21 (DE3) (Studier *et al.,* 1990), which is good for high-level protein expression and easy induction. The expression was induced by 0.4 mM isopropyl β-D-thiogalactoside (IPTG) when the cell culture, which had been grown in Luria broth containing 20 µg/ml kanamycin, reached an OD of ~0.4-0.6 (λ = 600nm). After the cells continued growing at 37°C for 3-5 more hours, they were ready for harvesting.

2.3 PURIFICATION

Though a good strategy for obtaining the purified protein was finally found, a lot of effort had been made and many difficulties had been encountered in establishing the final purification protocols. After having the protein successfully overexpressed, the first strategy thought of for purification was to take advantage of its presumably thermostable property. Boiling tests proved that *A. aeolicus* SmpB was truly heat-resistant and it was possible to remove most of the impurities by a single step of boiling. However, three disadvantages of this approach caused it to be discarded. First, boiling experiments were inconsistent. The boiling results could be affected dramatically by even a small change in buffer components, salt concentrations or buffer pH. Second, no matter how good the boiling results were, at least 40% of the target protein was always lost in the pellet after boiling and spinning (Figure 2.1). Third, it was hard to know what changes boiling could bring to the tertiary structure of the protein.

Although boiling was not a good method for large-scale purification and structure studies of this protein, it was used to purify enough sample for protein sequencing. The N-terminal twenty amino acids determined by the protein sequencing experiment turned out to be the same as that of the native *A. aeolicus* SmpB except for the missing of the first methionine, which might have been post-translationally cleaved off.

Next, a typical protein purification procedure was tested. After the cell lysate was cleaned by the addition of 0.1% polyethylenimine (PEI), both anion and cation exchange columns were explored for purification. The calculated isoelectric point (pI) of the protein is 9.92. In buffer with pH 7.5, the protein was expected to bind to a cation exchanger. However, the behavior of the protein was different from expectations. It bound only the anion exchanger and was always in the flow-through of any cation exchange columns. In addition, the binding to the anion exchanger seemed irregular, since the protein came off the column randomly during gradient elution. It was impossible to wash off the protein in a single peak and get it purified.

Problems described above were finally solved by adding extra polyethylenimine, up to 1% (w/v). The final purification strategy was as follows. After induction was done, *E. coli* cells were harvested by centrifugation at 6,000 rpm for 15 minutes, 4 °C. Pelleted cells were stored at -80 °C overnight. On the next day, frozen cells were thoroughly resuspended in 3-5 volumes of lysis buffer (50 mM Tris-HCl, pH 7.5, 100 mM NaCl, 1 mM EDTA, 1 mM DTT, 0.1% (v/v) Triton X-100), i.e., 3-5 ml buffer was used for 1 gram (wet weight) of cells. The

resuspended cells were disrupted by incubation on ice in the presence of about 3 mg/ml of lysozyme for approximately 20 minutes or until the solution turned viscous while being stirred, followed by sonication at 4 °C using a pulse sequence of 20 seconds on, 60 seconds off, until its viscosity decreased dramatically (4-6 times, typically). The broken cells were centrifuged at 22,000 × g for 25 minutes under 4°C. After the volume of the resulting supernatant was determined and recorded, the supernatant was carefully transferred into a fit beaker (an aliquot of 20 µl was saved for subsequent analysis on an SDS-PAGE gel). One-tenth volume of 10% (w/v) polyethylenimine (pH 8.0) was gradually added to the cell lysate within 10 minutes while being stirred on ice. Some egg-drop like precipitate typically appeared at this point. The mixture was kept on ice for 10 more minutes followed by being spun at 20,000 × g for 15 minutes at 4 °C. The supernatant was carefully poured into a fit beaker; the volume was determined and recorded (an aliquot of 20 µl was saved for subsequent analysis on an SDS-PAGE gel). A two-step ammonium sulfate (AS) precipitation was then carried out to concentrate and preliminarily purify the expressed target protein. In the first step, powdered AS was slowly added to the PEI-treated supernatant to bring the AS to 35% saturation. The mixture was centrifuged at 20,000 × g for 20 minutes under 4 °C. The supernatant was kept. This step removed cell debris and most impurities with lower solubility in AS but left most SmpB in the supernatant. In the second step, more powdered AS was gradually added to the above supernatant to bring the AS to 65% saturation. The mixture was centrifuged at 20,000 × g for 20 minutes at 4 °C. The pellet was kept after this step. This step should precipitate

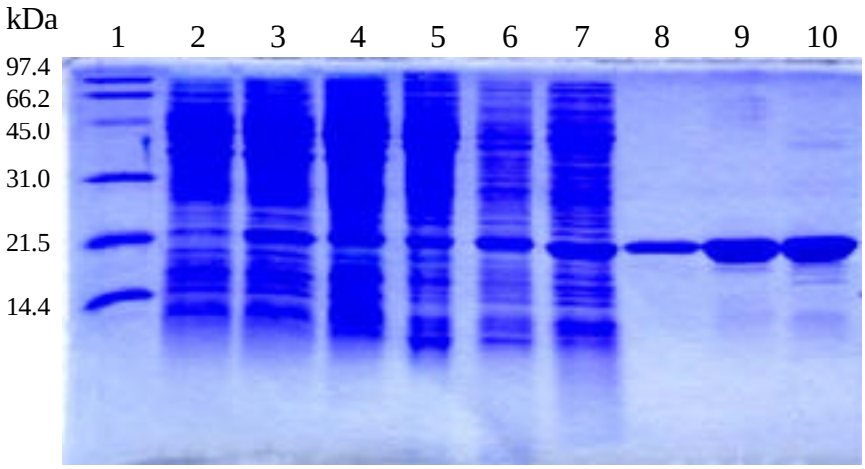

Figure 2.2 Purification results of *A. aeolicus* on SDS-PAGE 1. Low range molecular weight standard (BioRad), 2. Non-induced cells, 3. Induced cells, 4. Cell lysate, 5. 1% polyethylenimine supernatant, 6. 35-65% AS precipitate, 7. Dialysis supernatant, 8-10. Three different batches of S-15 elute (21-24% buffer B). The concentrations of the stacking and resolving gels were 3% and 16%, respectively.

out >95% of the SmpB along with much of the impurities. The resulting pellet was resuspended in 3-4 volumes of column running buffer A (50 mM NaH_2PO_4/Na_2HPO_4, pH 7.5, 100 mM NaCl). The resuspension was then dialyzed against 30-50 volumes of buffer A for 4 hours or longer. The dialyzed solution was spun at 20,000 × g for 10 minutes under 4 °C. The resulting supernatant was loaded onto a strong anion exchanger column, Unosphere Q (BioRad). The flow-through from this column was collected (all of the SmpB should be in this fraction) and then loaded onto a cation exchange column S-15. After the loading was complete, the column was washed with 1.5 × bed volume of buffer A to remove all proteins that could not bind to the column. The target protein was eluted by a gradient of increasing salt concentration. The protein came off at 21-24% of buffer B (50 mM NaH_2PO_4/Na_2HPO_4, pH 7.5, 500 mM NaCl). By this purification protocol, the final product was above 95% pure as judged on an SDS-PAGE gel (Figure 2.2). A preliminary 1D NMR spectrum of the purified protein suggested that the protein is properly folded and it exists predominately as a monomer in solution (Figure 1.6).

End of Chapter 2

Chapter 3

Proteolysis studies of *A. aeolicus* small protein B

3.1 INTRODUCTION

Proteolysis refers to the cleavage of peptide bonds in a protein or polypeptide chain by enzymes or chemical reagents. It is involved in a variety of physiological processes and plays vital roles in maintaining the living form of a cell, including removal of undesirable proteins, activation or maturation of the formerly inactive protein, regulation of central functions in the complex networks of stress response, differentiation, cell cycle, apoptosis and immune response, and so on. *In vitro* controlled proteolysis has been extensively explored and often employed to characterize the properties of different proteins.

3.1.1 Protease, proteinase and peptidase

To avoid confusion and misunderstanding of the terms, the International Union of Biochemistry and Molecular Biology (IUBMB, 1984) has recommended to use the term *peptidase* for the subset of peptide bond hydrolases (Subclass E.C 3.4.). The widely used term *protease* refers to the same enzyme as *peptidase*. *Peptidases* contain two groups of enzymes: the endopeptidases and the

exopeptidases; the former cleave peptide bonds at sites within the protein and the latter remove amino acids sequentially from either N or C-terminus. The term *proteinase* is also used as a synonym word for *endopeptidase*. The modern scheme of nomenclature is shown in Figure 3.1.

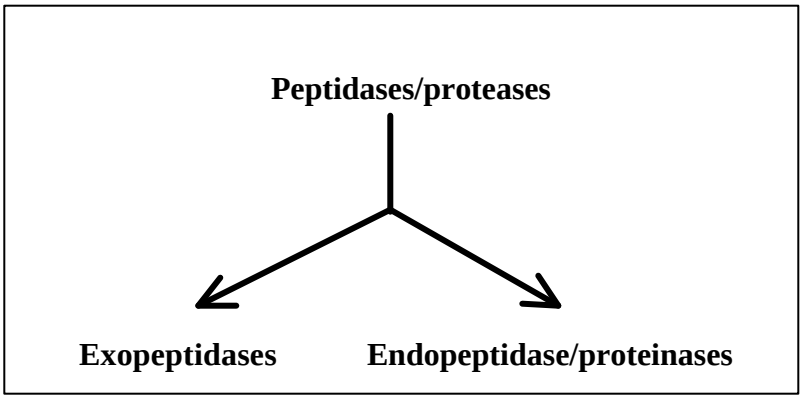

Figure 3.1 Relationship of the nomenclatures for proteases and peptidases.

3.1.2 Classification of proteinases

Proteinases are classified according to their catalytic mechanisms. Four mechanistic classes have been recognized by the IUBMB. They are serine proteinases, cysteine proteinases, aspartyl proteinases and metallo proteinases (Rawlings and Barrett, 1993). The catalytic reaction of serine and cysteine

proteinases involves the formation of covalent intermediates; in contrast, catalysis of aspartyl and metallo proteinases does not involve a covalent intermediate although a tetrahedral intermediate does exist. Table 3.1 lists the recognition sites for selected peptidases and chemicals.

In addition to these four mechanistic classes, there is a section of the enzyme nomenclature that is allocated for proteases of unidentified catalytic mechanism. This indicates that the catalytic mechanism has not been identified but the possibility remains that novel types of proteases do exist.

3.1.3 Catalytic mechanism of serine proteinases

The class of serine proteinases is one of the most commonly used endopeptidases for *in vitro* limited proteolysis because of their large-scale availability, easy handling and high specificity.

This class comprises two different families, the chymotrypsin family including the mammalian enzymes such as chymotrypsin, trypsin and elastase, and the substilisin family including the bacterial enzymes such as subtilisin (Rawlings and Barrett, 1994). The general 3D structure is different in the two families but they have the same active site geometry and then catalysis proceeds via the same mechanism. The catalytic reaction can be arbitrarily divided into two steps. The first step is the formation of an acyl enzyme intermediate between the substrate and the essential serine. This covalent tetrahedral intermediate is formed through the nucleophilic attack of the active site serine on the carbonyl carbon atom of the scissile peptide bond. During the second step (deacylation), the acyl-

enzyme intermediate is hydrolyzed by a water molecule to release the peptide and to restore the Ser-hydroxyl of the enzyme. The reversal of acylation leads to the formation of the second tetrahedral transition state intermediate. A water molecule is the attacking nucleophile instead of the serine residue. Finally, the histidine residue provides a general base and accepts the OH group of the reactive serine to yield the reaction's carboxyl product and the active enzyme (Voet *et al.,* 1998).

The two most frequently used proteases in this class are chymotrypsin and trypsin. The high homology and similar three-dimensional structures between them suggests that they might have arisen via duplications of an ancestral proteinase gene followed by the divergent evolution of the resulting enzymes (Voet *et al.,* 1998). However, the two enzymes have distinct specificities: chymotrypsin has a large shallow pocket lined with hydrophobic residues to accommodate the large hydrophobic side chains of phenylalanine, tyrosine and tryptophan, and so catalyses the cleavage of peptides and esters of these amino acids. Trypsin has a deep narrow pocket with an aspartate residue at the bottom of the pocket, and this aspartic acid forms a salt bridge with the positively charged group at the end of the substrate lysine and arginine side chains, on which this enzyme acts (Figure 3.2).

3.2 WHY PERFORM A PROTEOLYSIS EXPERIMENT

Controlled limited proteolysis has been widely used to study the properties of proteins, including (1) analysis of protein primary structure (Hsieh *et al.,*

1996), (2) identification of independent subdomains in proteins (Arbuckle *et al.*, 2001; Herrera *et al.*, 1993; Negishi *et al.*, 1995; Sekiguchi and Shuman 1997; Webb *et al.*, 1995), (3) prediction of folding and structure of proteins (Aceto *et al.*, 1998; Ehrlich *et al.*, 1994; Greasley *et al.*, 1993) or equilibrium intermediates (Webb *et al.*, 1997), (4) obtaining evidence for protein conformational changes induced by the binding of ligand (Egelund *et al.*, 2001; Mabjeesh and Kanner, 1993; Moldoveanu *et al.*, 2001; Yang *et al.*, 2000), point mutation (Medvedeva *et al.*, 1999) or change of solvent (Nicot *et al.*, 1993), (5) probing of protein stability and dynamics (Endo *et al.*, 1985; Iakoucheva *et al.*, 2001; Nakazawa *et al.*, 1993; Sassoon *et al.*, 2001), (6) study of protein structure-function relationship in the absence of three-dimensional structural data (Aceto *et al.*,1995; Arima *et al.*, 1998), (7) detection of functional change upon treatment by chemical reagents (Newton and Williams, 1993), (8) characterization of the mechanisms of protein degradation (Moradian-Oldak *et al.*, 2001), (9) solubilization of membrane binding proteins from the membrane (Hooper and Turner, 2000).

Two observations accounted for the initial motivation of performing controlled proteolysis experiments on *A. aeolicus* SmpB. First, crystallization trials had been exhaustively explored after the *A. aeolicus* small protein B was purified but no positive results were obtained. One possible explanation for the lack of crystals is that the protein might contain some unstructured long loop region(s) and/or disordered terminus (termini). The recently solved large ribosomal subunit 3D atomic structure (Ban *et al.*, 2000) indicated that many ribosomal proteins contain unstructured extensions protruding from the globular

bodies, which is coincidently happening on most of the ribosomal proteins that had been resisted crystallization in isolation. For example, protein L2 was eventually crystallized and its structure was finally solved only after its disordered extensions were removed (Nakagawa *et al.*, 1999). Secondly, preliminary two-dimensional NMR data suggested that part of the protein might be unstructured. Shown in Figure 3.3 is the "finger-print" region of the 2D COSY spectrum of the full-length *A. aeolicus* SmpB. As can be seen, the majority of the resonances are well dispersed, which is typical to a folded structure. However, there are quite a few resonances crowding in the center of the spectrum where the resonances of random coils are generally located.

3.3 RESULTS AND DISCUSSION

A series of commercially available proteinases were used to carry out the limited proteolysis experiments under controlled conditions. However, only the proteolytic action of endopeptidase trypsin was thoroughly examined, because on one hand there are more potential tryptic sites (residues Arg and Lys) in SmpB, a basic RNA-binding protein, and on the other hand trypsin is highly specific and gives more reproducible results. Figure 3.4 schematically shows the influence of residues around the cleavage site on tryptic proteolysis. There are no significant inhibition sites present in *A. aeolicus* SmpB sequence except for Arg134.

As shown in Figure 3.5, a stable shortened product was generated after the full-length SmpB (10mg/ml) was incubated at 37°C for 15 minutes in the presence of 0.01-0.1 mg/ml of trypsin. However, after the digestion product was

applied onto the cation exchange column S-15, two fractions were eluted by an increasing-concentration salt gradient (buffers and protocols used were the same as those for the purification of *A. aeolicus* SmpB, see section 2.3) as shown in Figure 3.6. Mass spectrometric analysis revealed the identities of these two fractions (Figure 3.7). The final results indicated that trypsin chopped off the C-terminal twenty-three amino acids as well as the two from the N-terminal of *A. aeolicus* SmpB. Careful examination of the cleavage products suggested that there are multiple trypsin accessible sites in the C-terminal twenty-three amino acid tail: Figure 3.6A lane 5 shows at least three extra bands between the full-length band and the final product band, and there is no lower band corresponding to the intact twenty-three amino acid polypeptide sequence in lane 3 and 4. These observations imply that the C-terminal tail is easily accessible for trypsin and may be relatively unstructured. This conclusion is consistent with the earlier suspicion based on the 2D spectrum. Actually, the 2D COSY spectrum of the trypsin digestion product (Gly1-Arg133) does miss only the overlapping resonances in the center (Figure 3.8).

Prolonged treatment with trypsin (Figure 3.9) suggested that (1) the core fragment (residues 3-133) was a compact, proteolytically stable domain. Examination of the primary sequence of the protein and the cleavage efficiency of trypsin (Figure 3.4) reveals that all the cleavage sites in SmpB are easily accessible (at least in principle) to trypsin. Therefore, the stability of this core fragment was not due to the presence of sequences that are not favorable for trypsin cleavage.

(2) The cleavage site on the N-terminal was more resistant to proteolysis than those at the C-terminal of *A. aeolicus* SmpB.

End of Chapter 3

Chapter 4*[1]

NMR Structure of *A. aeolicus* small protein B

4.1 PREPARATION OF ISOTOPE-LABELED SAMPLES

4.1.1 Why bother with isotope labeling

Successful structure determinations by NMR rely on nearly complete resonance assignments as well as a sufficient number of accurate distances from nuclear Overhauser effect (NOE) data. Structure determination of a protein with molecule weight higher than 10 kDa (MW of full-length SmpB is ~18 kDa; MW of 1-133 fragment is ~15 kDa) may become extremely difficult because of the severe resonance overlap in the conventional 1D and 2D homonuclear NMR spectra. This overlap problem inevitably leads to the ambiguous assignments of many proton resonances, and thus makes the structure determination almost certainly impossible. Thanks to the uniform-isotope labeling and the development of triple-resonance multidimensional NMR experiments, at present NMR characterization of proteins and protein/protein or protein/RNA complexes with molecular weight up to 25-30 kDa are, in many cases, possible (Kanelis *et al.*, 2001).

---

*[1]This chapter is modified from Dong, G., Nowakowski, J. & Hoffman, D.W. (2002) *EMBO J.* **21:** 1845-1854.

Multidimensional heteronuclear NMR experiments can easily make the resonances become dispersed and thus be readily resolved by the addition of a heteronuclear dimension. ^{15}N and ^{13}C are the two NMR-active isotopes routinely incorporated into proteins for use in NMR structure determinations.

4.1.2 Isotope labeling of *A. aeolicus* SmpB

Samples of SmpB enriched in ^{15}N and/or ^{13}C were prepared by growing cells in a modified version of M19 minimal media (Sambrook *et al.*, 1989) supplemented with 1 g/l of ^{15}N-ammonium chloride and/or 1 g/l of ^{13}C-glucose (Cambridge Isotope Laboratories) as the sole source of nitrogen and/or carbon (Table 4.1). Induction and purification protocols were the same as those for the unlabeled *A. aeolicus* SmpB (see section 2.2 and 2.3). Fragment 1-133 was generated by incubating of ~10mg/ml of purified SmpB (50 mM Na-PO4, pH 8.0, 100 mM NaCl) at 37°C for 2 minutes in the presence of 0.1 mg/ml of trypsin, followed by cation exchange chromatography (UNOsphere S-15, Bio-Rad). The tryptic proteolysis reaction was terminated by the addition of 15 µg/ml of phenylmethyl sulfonyl fluoride (PMSF).

4.2 NMR SPECTROSCOPY

The NMR data collection was performed on a 500 MHz Varian Inova spectrometer equipped with a triple-resonance probe and *z*-axis pulsed-field gradient. The 1-2 mM SmpB protein samples were prepared in 20 mM sodium

phosphate buffer (pH 6.0) containing 80 mM NaCl, 20 mM NaN_3 and 90% H_2O/10% D_2O or 100% D_2O. The temperature of the samples was maintained at 30 or 40°C. The water signal was suppressed by selective presaturation during the relaxation delay. Homonuclear COSY, TOCSY and NOESY experiments were carried out using unlabeled samples. Three-dimensional heteronuclear HNCA (Muhandiram and Kay, 1994), HNCO (Grzesiek and Bax, 1992), HNCACB (Muhandiram and Kay, 1994) and HN(CO)CACB (Muhandiram and Kay, 1994) spectra were collected using ^{15}N- and ^{13}C-enrinched protein samples. These heteronuclear spectra correlated the backbone protons to the N, C^{α}, C^{O} and C^{β} signals of the same and adjacent amino acid residues and were used for subsequent backbone resonance assignments. Side chain ^{1}H resonance assignments were obtained using two-dimensional DQF-COSY and TOCSY spectra, and three-dimensional ^{15}N-^{1}H-^{1}H HMQC-TOCSY and ^{1}H-^{1}H-^{13}C HCCH-TOCSY (Kay *et al.*, 1993) spectra. NOE cross peaks were identified in two-dimensional ^{1}H-^{1}H NOESY spectra, a three-dimensional ^{15}N-^{1}H-^{1}H HSQC-NOESY spectrum, and a three-dimensional ^{13}C-edited ^{1}H-^{1}H HSQC-NOESY (Pascal *et al.*, 1994) spectrum. All heteronuclear NMR spectra showed a good dispersion of cross-peaks essential for high-quality structure determination. The ^{13}C-edited ^{1}H-^{1}H NOESY spectrum was collected (in 90% H_2O/10%D_2O solvent), so that NOE peaks between amide and side-chain protons could be resolved by the chemical shift of a side-chain ^{13}C nucleus.

All spectral data were processed on a Silicon Graphics workstation using the program FELIX (Hare Research). ^{1}H, ^{15}N, and ^{13}C chemical shifts are

referenced as recommended by Wishart *et al.* (1995), with proton chemical shifts referenced to internal 2,2-dimethyl-2-silapentane-5-sulfonate (DSS) at 0 ppm. The 0 ppm ^{13}C and ^{15}N reference frequencies were determined by multiplying the 0 ppm ^{1}H reference frequency by 0.251 449 530 and 0.101 329 118, respectively.

4.3 RESONANCE ASSIGNMENT

Sequence-specific and nearly complete assignments of the resonances are the prerequisite for solving protein structures by NMR techniques. To successfully perform the assignments on a large protein, such as *A. aeolicus* SmpB, 3D NMR spectra are required in order to resolve the overlapping resonances.

Sequential assignment was carried out using several $^{1}H/^{15}N/^{13}C$-3D heteronulcear spectra. The HNCOCACB spectrum, which correlates the amide proton (NH) of an amino acid to the C^{α} and C^{β} of the residue ahead of it, was employed to assign the C^{α}-1 and C^{β}-1 chemical shifts of each spin system. Based on the correlation between the amide proton (NH) and the C^{α} of the same amino acid in the HNCA spectrum, the C^{α} chemical shift of each amino acid was assigned. The C^{β} chemical shift of each spin system was lastly assigned using the HNCACB spectrum that correlates the amide proton (NH) to the C^{α} and C^{β} of the same amino acid. By this procedure, the chemical shifts of C^{α}, C^{β}, C^{α}-1 and C^{β}-1 of each detectable resonance were assigned. These results were then used to establish the sequential resonance assignment by examination of the matched corresponding carbon chemical shifts.

Sequence specific assignments were completed with the help the characteristic C^{α} and/or C^{β} chemical shifts of the so-called "index" amino acids including serine/threonine (C^{α}, C^{β} ~ 60-70 ppm), alanine (C^{β} ~ 20 ppm) and glycine (C^{α} ~ 45 ppm, no C^{β}). Although it is theoretically possible to completely assign the backbone nuclei using a set of 3D heteronuclear spectra, an unambiguous assignment of all backbone nuclei is rarely accomplished due to the resonance degeneracy problem in backbone nitrogen and carbonyl carbon chemical shifts of residues near the termini and in some internal flexible regions. For *A. aeolicus* SmpB (residues 1-133), 123 out of the possibly detected 128 spin systems (>96%) were unambiguously assigned. Those unassigned amino acids are not conserved, and are thus presumably not in functionally important regions of the protein structure.

After the sequence specific backbone assignments were finished, side chain proton assignments of individual amino acids were performed using the 3D TOCSY and HCCH-TOCSY spectra; the latter is particularly important for the complete side chain assignments of the conserved hydrophobic amino acids that constitute the hydrophobic core of a protein.

4.4 STRUCTURE DETERMINATION

4.4.1 Restraint assignment strategies

To determine the structure of *A. aeolicus* SmpB, the hybrid distance geometry-simulated annealing and energy minimization protocols within the CNS program suite (Brünger *et al.*, 1998) were employed. The goal was to identify the

full range of structures that are consistent with the distance and angle constraints derived from the NMR data, while having reasonable molecular geometry, consistent with a minimum value of the CNS energy function. Distance constraints were derived from the intensities of cross peaks within the multidimensional NOESY spectra. Whenever possible, NOE cross peaks were identified in spectra with relatively short (60 ms) mixing times, to minimize the effects of spin diffusion on the structure calculation. The NOE cross-peaks were qualitatively categorized as strong, medium, weak and very weak and used to assign upper distance limits of 3.2, 3.6, 4.2 and 4.5 Å, respectively. NOE cross peaks obtained from the ^{15}N-edited or ^{13}C-edited three-dimensional NOESY spectra collected with longer mixing times (120 ms), or in two-dimensional homonuclear spectra with mixing times of up to 160 ms, were assigned to inter-proton distance bounds as follows: strong and medium peaks (≤5.5 Å), weak and very weak peaks (≤7.0 Å). These more generous distance ranges were used so as to minimize errors due to the influence of spin diffusion on peak intensities.

Pseudo-atom correction for unassigned stereo partners and magnetically equivalent protons was applied to eliminate the errors caused by the ambiguous assignments of equivalent protons. The practical strategies were, (1) when NOEs involving methyl protons of valine and leucine were not stereospecifically assigned, distances were measured from the center of the two methyl groups, and 2 Å was added to the inter-proton distance; (2) for NOEs involving other methyl protons, distances were measured from the center of the methyl group and 1 Å was added to the inter-proton distance; (3) for NOEs involving methylene protons

with no stereospecific assignment, distances were measured from the center of the methylene group and 0.7 Å was added to the inter-proton distance; (4) for NOEs involving delta and epsilon protons on tyrosine and phenylalanine rings that were not uniquely assigned, distances were measured from the central point between the two delta (or epsilon) protons, and 2.5 Å were added to the inter-proton distance.

Backbone torsion angle restraints for phi and psi were only included for regions of regular β-strand or α-helix structure that were clearly identified by characteristic NOE cross peak patterns, ^{13}C chemical shifts and slowly exchanging amide protons. In these cases, phi and psi were restricted to –120° ± 25° and 150° ± 25°, respectively, for β-strands, and –60° ± 25° and –60° ± 25°, respectively, for α-helices. Hydrogen bond restraints were implemented using distance bounds to reinforce canonical secondary structures and were only included for regions of regular β-strand or α-helix structure identified based on characteristic NOE patterns and chemical shift indices, where the amide protons were substantially protected (Figure 4.1) from solvent exchange (in D_2O buffer after two hours at 30°C). Distance ranges involving these hydrogen bonds (N-H•••O-C) were set as follows: N-H•••O-C, 4.35 ± 0.2 ? ; N-H•••O, 3.14 ± 0.2 ? ; H•••O-C, 3.38 ± 0.2 ?; H•••O, 2.15 ± 0.2 ? .

4.4.2 Structure calculation

Structure calculations were performed using the hybrid distance geometry/simulated annealing method in the CNS program (version 1.0) (Brünger*et al.,* 1998). The structure calculation proceeded in two stages. First, ten diverse starting structures were generated by subjecting a random coil model to the CNS simulated annealing protocol using only the dihedral angle constraints. In the second stage, the ten structures generated above were used as starting models for one hundred runs of the simulated annealing protocol in the presence of all identified NMR constraints. Most of the simulated annealing runs resulted in similar structures with similar energies. From this finalset of refined models, a set of 20 structures were selected that satisfy the following criteria: (1) their CNS energy term was at or very near the minimum value obtained; (2) there were no inter-proton distance constraint violations of >0.6 Å; and (3) no torsion angle constraint violations exceeded 2.5°. These 20 structure models are a fair representation of the full range of structures that satisfy the NMR-derived restraints while having reasonable molecular geometry, as defined by the CNS energy function. Structural statistics (Table 4.2) were calculated with the assistance of the program PROCHECK-NMR (Laskowski *et al.,* 1996).

4.5 RESULTS AND DISCUSSION

4.5.1 Structure of *A. aeolicus* SmpB

4.5.1.1 Structure of the core fragment

NMR determination was first implemented to the trypsin-generated major fragment (residues 1-133), which has a presumably compact structure based on the proteolysis studies. This fragment was found to be very suited for NMR structural characterization. 123 out of the 128 possibly detectable residues (total 133 amino acids minus 4 proline residues and the amino acid on the N-terminus) were identified and unambiguously assigned on the ^{15}N-^{1}H correlated HSQC spectrum (Figure 4.2), and all four proline amino acids were also assigned by examination of other spectra. The structure of this trypsin-resistant core was determined from constraints derived from NMR data, specifically distance constraints derived from observed NOE intensities, and torsion angle and hydrogen bond constraints for the regions identified as having regular β-sheet or helical secondary structure. The structure turned out to be quite well defined by the NMR data, with >2000 unambiguous distance constraints (559 of which were long range) being derived from the nuclear Overhauser effect (NOE) spectra. Nearly complete chemical shift assignments were obtained for the ^{1}H, ^{13}C and ^{15}N nuclei. Of particular significance, nearly complete resonance assignments were obtained for, along with others, all of the leucine, isoleucine, valine, proline and alanine side chains; assignments of NOE cross peaks derived from these side chains were critical in accurately defining the hydrophobic core of the protein.

Figure 4.3 shows the superposition of a set of structures that are equally consistent with the NMR-derived constraints. These structures fairly represent the full range of conformations of the protein that are consistent with the NMR data, and are all quite similar (Figure 4.4). Structural statistics for residues 1-133 of SmpB are summarized in Table 4.2. Coordinates of ten conformers for residues 1-133 of the SmpB structure with the associated restraint list have been deposited in the Protein Data Bank (access code 1K8H).

The NMR studies revealed seven β strands (designated β1, β2, …, β7) and three regular α helices (designated α1, α2 and α3) in the core of the SmpB protein. Six of the strands (β2-β7) form a closed antiparallel β barrel with strand β1 antiparallel to β6 but being occluded from the barrel (Figure 4.1). The lengths of the β strands vary significantly. Strands β2, β6 and β7 comprise eleven residues each and are connected by an extensive network of hydrogen bonds to form a large sheet; strands β1, β3, β4 and β5 are relatively short, consisting of four to six residues each. The regular networks of hydrogen bonds in strands β2 and β6 are both interrupted by a single-residue bulge, at residues Leu22 and Leu109, respectively (Figure 4.1). These bulges are well defined by the NMR data. The bulge structures contribute to the overall twist of the β-sheet as well as the maintenance of the closed β barrel structure (bulge at Leu109) or the possible connection between strands β1 and β2 (bulge at Leu22). Remarkably, the residues of strands β2 and β6 near each bulge are well conserved among various species (Figure 4.5), suggesting that the bulges are an important structural feature maintained in the evolution process. The three α helices are, the one-turn helix α1

which links strands β1 and β2, the two-and-a-half-turn helix α2 which links strands β2 and β3, and the three-and-a-half-turn helix α3 which links strands β5 and β6. Each of these three helices contains a conserved hydrophobic face that contacts conserved hydrophobic residues on the external surface of the β-barrel. The most extensive region of non-regular secondary structure is located between residues 64 and 82 (designated loop 2), which connects strands β4 and β5. There is another short flexible region between residues 43 and 49, designated loop 1, which links helix α1 and strand β3. Ribbon diagrams depicting the fold of SmpB are shown in Figure 4.6.

4.5.1.2 Structure of the C-terminal tail and its relationship with the core

NMR data provide evidence that the structure of the trypsin-resistant core of *A. aeolicus* SmpB (residues 1–133) is independent of the hydrophilic C-terminal "tail" of residues 134–156. Triple-resonance spectra were acquired for the full-length and truncated version of the protein, and used to compare the chemical shifts of the ^{1}H, ^{13}C and ^{15}N nuclei. Table 4.3 shows the chemical shift differences between the full-length and truncated version (residues 1-133) of the protein for backbone nitrogen (N), amide proton (NH), backbone alpha carbon (C^{α}) and backbone beta carbon (C^{β}) resonances (shown in ppm from data with a proton frequency of 500 MHz). For the region of the molecule where the resonance assignments are essentially complete (residues 3–130), chemical shift comparison revealed two regions with substantial changes: one is around residue His71 and the other is around residue His88. These changes can be attributed to

the histidine residues whose side chain has a pKa of 6.0, the same as the pH of the used buffer. Small variations of the buffer for different data sets were expected, which might have significant effects on the ionization of histidine side chain and therefore induce the changes of their chemical shifts. The reason why there is no significant chemical shift change for residue His77 might be due to the lack of hydrophilic residues next to it, or the side chain of this histidine is shielded. In fact, significant chemical shift changes of His88 had been frequently noticed among other data sets (not shown).

In conclusion, there are no significant chemical shift changes that can be attributed to the removal of the C-terminal tail. The resonances of the nuclei within the C-terminal tail were identified by comparing the spectra of the full-length and truncated versions of the protein; however, it was impossible to make sequence-specific assignments for the majority of these tail residues due to the severe overlap of their chemical shifts near the random coil values. Overall, the NMR data support a model where the globular trypsin-resistant core of SmpB is structurally independent of the C-terminal tail, which most likely has a random coil structure, at least when the purified SmpB is isolated in solution.

4.5.2 Structural neighbors of *A. aeolicus* SmpB

Primary sequence search by BLAST (available at the NCBI web site), which explores all available sequence databases, indicated that *A. aeolicus* SmpB has no significant homology with other known proteins other than SmpB proteins from different species. However, the three-dimensional structure solved by multi-

dimensional NMR spectroscopy here provides an additional means for comparing SmpB with other proteins. In favorable cases, comparing 3D structures may reveal biologically interesting similarities and distant evolutionary relationships that are not detectable by comparing sequences.

Searches using the programs DALI (Holm and Sander, 1997) and VAST (available at the NCBI web site) failed to find any structures within the Protein Data Bank that have significant similarity to SmpB, based on comparison of the coordinates of backbone atoms. However, a more qualitative inspection of the overall topology of SmpB revealed the presence of an oligonucleotide-binding (OB) fold (Murzin, 1993) within the structure of its β-sheet. In this respect, SmpB is similar to several other RNA-binding proteins that are known to be associated with translation (Draper and Reynaldo, 1999), including ribosomal protein S17 and the prokaryotic translation initiation factor IF1 (Figure 4.7). Like SmpB, proteins IF1 and S17 each contain substantial RNA binding surfaces (Figure 4.8). IF1 binds to the 16S RNA in a transient manner at the ribosomal A-site (Carter *et al.*, 2001), while S17 is a more permanent part of the ribosomal small subunit with extensive contacts to the 16S RNA (Wimberly *et al.*, 2000). The similarity between SmpB and the N-terminal domain of aspartyl tRNA synthetase (DRS) is even more notable due to the presence of a helix that is in a position analogous to helix α3 in SmpB and its more likely β-barrel tertiary fold (Figure 4.8). This tRNA synthetase domain directly contacts the anticodon loop when in complex with the tRNA (Cavarelli *et al.*, 1993). These apparent subtle similarities between the structures of IF1, S17, DRS and SmpB, combined with their close association

with the RNA components of the translational apparatus, suggest that they may be the divergent functional versions of a common ancestor or be linked by an evolutionary relationship. Functionally, SmpB may be more like a ribosomal protein than a transient RNA-binding protein, such as a tRNA synthetase or an initiation factor, since the available evidence suggests it is an integral part of the tmRNA ribonucleoprotein complex (Karzai and Sauer, 2001).

In addition to SmpB, the structures of other proteins that contact RNA and are associated with the process of translation have been found to contain six-stranded closed β-barrels; these include domain III of EF-Tu (Nissen *et al.,* 1995) and domains within translation initiation factor IF2 (Meunier *et al.,* 2000; Roll-Mecak *et al.,* 2000). However, the connectivity of the strands within the β-barrels of the EF-Tu and the IF2 domains differs from that in SmpB, providing evidence against a close evolutionary relationship between these proteins.

4.5.3 Surface charge distribution and predicted function

Amino acids are conserved among similar proteins from widely divergent species for reasons related to either structure or function; residues within a protein core are often conserved for structural reasons, while conserved residues on a protein surface are often important for function. When the amino acid sequences of SmpB from several species are aligned (Figure 4.5) and this alignment is compared with the structural model, it becomes possible to identify and distinguish those amino acids that are most likely to be essential for either structural or functional purposes. Hydrophobic residues within the core of the

protein, as well as conserved residues in turns, are likely to have essential structural roles; these residues are boxed in Figure 4.5. Well-conserved amino acids on the surface of the protein were also identified; these residues are most likely to make essential contacts with the tmRNA or other RNA or protein components of the translational apparatus, as part of the mechanism of ribosome rescue and tagging of peptides for degradation.

Conserved residues cover a substantial fraction of the surface of SmpB. This is particularly apparent when the protein is viewed facing strands β3, β4 and β5 of the β-barrel (Figure 4.9a). The conserved surface residues are clustered into two regions of the protein surface. One of these regions consists of a broad surface centered on the solvent-exposed side of the strands β4 and β5, and contains residues Glu35, Lys37, Arg40, Glu58, Trp60, Arg81, Arg83, Lys84, Lys89, Glu91 and Lys119. These residue types are typical of those that populate RNA binding sites in other RNA-associated proteins, where lysine and arginine can have favorable electrostatic interactions with RNA, glutamate can serve as a hydrogen bond acceptor, and tryptophan can interact with RNA via ring stacking. This surface is therefore a likely region of contact with RNA. A second conserved surface contains residues Asn12, Lys13, Glu14, Tyr19, Glu23, Asp49, Phe51, Lys126, Lys128, Lys129, Asp132 and Arg133, and represents a second likely RNA-binding site. The two charge clusters are more apparent in a view from the top of the β-barrel (Figure 4.10). A plot of the electrostatic potential shows that the most conserved regions of the protein surface are, in the main, positively charged (Figure 4.11), which would facilitate interactions with the RNA

backbone. The total charge of SmpB is clearly positive; the calculated isoelectric pH of the protein is 9.92. Interestingly, SmpB contains a surface region centered on the solvent-exposed side of strands β2, β6 and β7, which is mostly hydrophilic but lacking in well conserved residues (Figure 4.9b), and therefore is likely to be solvent exposed when SmpB is bound in complex with the tmRNA.

Based on the structural results and comparison analysis presented above, it is possible to predict how and where SmpB might bind to the tmRNA as well as the function of SmpB in the *trans*-translation process. The tmRNA, with ~360 nucleotides, is substantially larger than SmpB alone. The secondary structure of the RNA (Figure 1.1) contains a tRNA-like domain and an mRNA-like domain comprising a region with an mRNA-like function and four pseudoknots (PK1-PK4), the first of which (PK1) is located between the mRNA- and tRNA-like domains within the tmRNA primary sequence (Karzai *et al.*, 2000; Zwieb *et al.*, 2001). Previously it has been reported that modification of the PK2, PK3 and PK4 pseudoknots does not remove the peptide tagging function of the tmRNA-SmpB complex (Nameki *et al.*, 2000), which suggests that SmpB binds to the mRNA- or tRNA-like regions or PK1. SmpB does not prohibit binding of EF-Tu to the tRNA-like domain or block the interaction of the pseudoknots and mRNA-like region with ribosomal protein S1 (Wower *et al.*, 2000), and actually stimulates the function of alanine tRNA synthetase as it charges the tRNA-like domain with alanine (Karzai *et al.*, 2000; Barends *et al.*, 2001). This information serves to restrict even further where the SmpB-tmRNA interactions may occur. Recently, Barends *et al.* (2001) used a combination of *in vitro* kinetic, gel-shift

and enzymatic protection assays to provide additional evidence that SmpB and EF-Tu simultaneously bind to the amino acid acceptor stem region of Ala-tmRNA. Further insight into the nature of the SmpB-tmRNA interaction is provided by the results of the present work, which show that conserved residues of SmpB are distributed over a large region of the protein surface, and on opposite sides of the protein structure (Figure 4.9), suggesting that SmpB contains more than one interaction surface and makes more than just a single simple contact with the tmRNA. The evidence that is now available suggests that SmpB may serve to bring together and stabilize a particular conformation and relative orientation of the mRNA- and tRNA-like regions of the tmRNA, in a role analogous to that of some of the ribosomal proteins that contact distal regions of the ribosomal RNA and thus induce or stabilize compact tertiary structures which are largely absent in the naked ribosomal RNA. Also, important functional roles of protein-induced RNA conformational rearrangements have been revealed in other RNAs. For example, the *Neurospora crassa* mitochondrial tyrosyl-tRNA synthetase (Cyt-18) was found to bind a group I intron and fold the preexisting RNA secondary structure into the catalytically active three-dimensional structure (Caprara *et al.*, 1996a, 1996b; Myers *et al.*, 2002).

In an alternative role, SmpB, via its substantial and widely spaced RNA-binding surfaces, could serve as a mediator of the interaction between the tmRNA and the stalled ribosome as part of the mechanism of ribosome rescue by transference of the decoding function from the messenger RNA to the mRNA-like region of the tmRNA; perhaps a surface of SmpB can interact with the 16S

ribosomal RNA at or near the sites usually occupied by tRNA to mimic the codon-anticodon recognition structure, noting that the first amino acid incorporated in the tag is not encoded by either the "broken" mRNA or the coding sequence of tmRNA. In support of this scenario, it has been reported that SmpB is required for the association of tmRNA with 70S ribosomes (Karzai *et al.*, 1999), and previous studies suggest that the first elongation cycle with tmRNA may have mechanistic differences with the normal elongation cycles (De la Cruz and Vioque, 2001).

Recent studies provide the third possibility for the function of SmpB. *In vitro* binding studies suggest that tmRNA interacts with the tRNA that decodes the resume codon prior to entering the ribosome (Gillet and Felden, 2001a). Recent published data demonstrate that SmpB binds to canonical tRNA with similar affinity as to tmRNA (Wower *et al.*, 2002). Therefore, SmpB may serve to mediate the interaction between tmRNA and tRNA, with one RNA-binding site interacting with tmRNA and the other with tRNA.

In terms of the function of the unstructured C-terminal tail, lack of structure in this portion does not necessarily imply that it is functionally unimportant; evidence to the contrary is provided by the presence of well conserved, positively charged residues Arg134 and Lys140, which have the potential to specifically interact with RNA. In this sense, SmpB is reminiscent of several of the proteins within the large and small ribosomal subunits, such as L15, L21e and L37e (Ban *et al.*, 2000), and S4 (Sayers *et al.*, 2000), which contain substantial charged tails at the N- and/or C-terminal of an otherwise globular

domain. In the context of the structure of the complete ribosomal subunits, these charged protein tails were found to penetrate to the interior of the ribosome and make specific contacts with the ribosomal RNA (Ban *et al.*, 2000; Wimberly *et al.*, 2000). By analogy, we suggest that the C-terminal tail of SmpB may become structured and have specific interactions within the context of the complete and functional tmRNA-SmpB ribonucleoprotein complex, perhaps penetrating into the interior of the tmRNA. *A. aeolicus* SmpB that is missing the C-terminal tail residues 134-156 was found to have a significantly reduced affinity for the tmRNA (Wower *et al.*, 2002).

Clearly, however, a detailed and definitive answer as to exactly how SmpB binds to the tmRNA and enables the functions of ribosome rescue and peptide tagging must await an NMR or crystallographic analysis of SmpB protein-RNA complexes. NMR perturbation results of the protein/RNA complexes are shown and discussed in chapter 5. Further crystallographic studies to shed more light on the details of the SmpB-tmRNA complex and its interactions are in progress.

End of Chapter 4

Chapter 5

Studies of the interaction of SmpB with RNA by NMR perturbation

5.1 INTRODUCTION

5.1.1 RNA and RNA-protein interaction

RNA molecules play a central role in all the main functions of living cells: storage of genetic information, propagation of the genetic material, and enzymatic activity. Since the finding that RNA can function as an enzyme (for reviews see Cech, 1989; Cedergren, 1990; Lamond and Gibson, 1990), RNA has been suspected to be life's most ancient molecule. Very surprisingly and interestingly, a ribozyme that ligates RNA to protein has been recently created using a combination of rational design and *in vitro* selection (Baskerville and Bartel, 2002).

Many RNAs fold into complex three-dimensional shapes with tertiary structures that have been found to be necessary for their biological activities (Cech, 2000; Nissen *et al.*, 2000). However, large, flexible and structurally heterogeneous biological RNAs may be difficult to purify to homogeneity. This can prevent structural characterization of RNAs by X-ray crystallography due to the difficulties of obtaining well-ordered crystals. As an alternative to

crystallography, NMR has been successfully used to determine the structure and dynamics of many RNA oligonucleotides (Kolk *et al.,* 1998; Schmitz *et al.,* Wimberly *et al.,* 1993; Wimberly *et al.,* 1999; Yoo *et al.,* 2001). Unfortunately, NMR techniques cannot be applied to most large RNA molecules (>50 nucleotides) because of its methodologically determined size limit.

Instead of functioning alone, most RNAs are associated with RNA-binding proteins that control RNA metabolism and/or functionalities. Understanding how RNA-binding proteins and RNA interact with each other is therefore central to understanding a wide range of biological processes.

The past several years have witnessed significant advances in our understanding of protein-RNA recognition and interaction, but in comparison to the extensively studied DNA-protein interactions, the number of atomic resolution structures of RNA and RNA-protein complexes is still very small. A question one would quickly think of is, "Is it possible to gain valuable information for RNA-protein recognition and interaction from these available DNA-protein interaction data, since there are significant chemical similarities between these two types of nucleic acids?" Before answering this question, let us first take a look at the similarities and differences between the RNA-protein interactions and the DNA-protein interactions. For DNA-protein complexes the many existing structures define an important paradigm in intermolecular recognition. In many cases, DNA-binding proteins target DNA molecules via inserting an α-helix into the major groove of a double-stranded DNA. However, this paradigm cannot be applied to protein-RNA recognition, since the major groove of double-stranded RNA is too

narrow to allow the insertion of a protein α-helix or β-strand (Figure 5.1). Instead, all known sequence-specific RNA-binding proteins generally recognize single-stranded regions, bulges, and hairpins or internal loops (Figure 5.2). On the other hand, all DNA double-helical structures are very similar in a sense, but the structures of RNA often vary significantly. Sequence-specific RNA recognition by RNA-binding proteins is usually through unique shapes and charge distributions of different RNAs, which, for DNA-binding proteins, is often achieved via a very precise reading of the identity of individual nucleotides within the DNA double helix. Therefore, RNA-protein interaction mechanisms cannot be readily derived from the existing DNA-protein recognition data, although RNA is chemically similar to DNA (Varani, 1997).

What could one do to study the RNA-protein interactions when the structures of the RNA and the RNA-complex are not available? Thanks to the recently developed biophysical method called NMR perturbation or NMR chemical shift mapping (section 1.3), it is now possible to study some aspects of the RNA-protein recognition in solution without knowing the structure of RNA.

The failure to obtain crystals for the intact 347-nucleotide *A. aeolicus* tmRNA and the availability of the hypothetical binding sites of SmpB on tmRNA led to the NMR perturbation studies of SmpB with several small variants of the tmRNA. The preliminary results are presented in this chapter.

5.1.2 Protein-RNA interaction studies by NMR perturbation

RNA-binding proteins typically have a modular structure and contain RNA-binding domains of 70-150 amino acids that mediate RNA recognition (Mattaj, 1993; Varani, 1997). This provides the opportunity for studying the RNA-protein interaction by NMR perturbation, which requires the known solution structure or the relatively complete resonance assignments of the atoms of only one component in a complex. At the present time, NMR can readily solve the structure of a protein consisting of up to 200 amino acids, provided that the protein is properly folded and exists as a monomer in solution. With its structure known, the interaction between a RNA-binding protein and its RNA partner can be studied by the NMR perturbation method. Many successful NMR perturbation studies on protein-nucleic acids complexes have been reported, but typically the nucleic acids used are relatively short, such as a 10-mer RNA (Lee *et al.,* 1997), a 15-base pair DNA (Foster *et al.,* 1998), a 9-base ss-/ds-DNA (Buchko *et al.,* 1999), a short DNA and a 24-mer RNA (Katahira *et al.,* 2001). These might be attributed to two reasons: the low solubility of the complexes caused by the flexibility of large nucleic acids and/or the change of protein electrostatic properties upon RNA binding, or the formation of nonspecific aggregates at high concentration (Smith, 1998). Nevertheless, the success of the NMR perturbation in a study of the 76-amino acid protein L11 complexed with a 58-nucleotide part of the 23S rRNA (Hinck *et al.,* 1997) encouraged the performance of the work described here.

5.2 RNA PREPARATION

5.2.1 Construction of tRNA-like variants of *A. aeolicus* tmRNA

Three short versions (Figure 5.3) of the tRNA-like domain of *A. aeolicus* tmRNA were designed based on the secondary structure of the tmRNA: construct No. 1, designated C1t, has the similar size as a canonical tRNA; No. 2 (C2t) includes one extra stem and No. 3 (C3t) includes two more stems. A tetra-loop (UUCG) was used in each construct to connect the two strands as well as to stabilize the RNA secondary structure because of its unusual structural stability reported before (Molinaro and Tinoco, 1995). The variants were constructed using the "primer-as-template" (or "template-free") PCR approach. The primers used are listed in Table 5.1. Three endonuclease cleavage sites and a T7 promoter sequence were incorporated in each construct. The three endonuclease recognition sites are: a *Bam*H I site on the 5' end, a *Hind* III site on the 3' end and a *Bst*N I site just ahead of the *Hind* III site. The first two sites were used for inserting the RNA coding sequence into the pUC18 vector, and the last one was used to linearize the plasmid for subsequent *in vitro* transcription. The T7 promoter sequence was added in the upstream of the RNA coding sequence so as to generate target RNA using the *in vitro* transcription reaction catalyzed by the T7 RNA polymerase.

For constructs C1t and C2t, only two primers were required in a single-step PCR reaction. For construct C3t, which is too large to be amplified by a single-step PCR reaction, three primers were used to perform the two-step PCR reaction. In the first step, the 5' end primer and the central primer were used. In

the second step, the purified PCR product from the first step and the 3' end primer were used. All primers were chemically synthesized by Integrated DNA Technologies (IDT) using automated solid phase synthesis techniques, and PAGE-purified. The dried DNA oligos were dissolved in an appropriate volume of ddH_2O to bring the final concentration to 100 pmol/μl. In each PCR reaction, 5 μl of 10 × reaction buffer, 1 μl of dNTPs-Mix (12.5 mM each), 1 μl of each primer, 0.5 μl of Taq DNA polymerase (5 units/μl, BioLabs) and 31.5 μl of ddH_2O were mixed in a 0.6-ml Eppendorf tube. "Touchdown" PCR reaction (Don *et al.,* 1991) was employed to bypass more complicated optimization processes for determining optimal annealing temperatures and to enrich the correct product over any incorrect and nonspecific-annealing products caused by the high GC content as well as the long G and C strings present in the primer sequences. The reactions were carried out on the MiniCycler™ using 92°C (45 sec) for denaturation, 72°C (45 sec) for annealing in the first cycle, 72°C (15 sec) for extension. The annealing temperature was decreased by 2°C every second cycle until the "touchdown" annealing temperature, 56°C, was reached, which was then used for 20 more cycles of polymerase chain reaction. The final step extension was prolonged to 5 minutes to fill in any uncompleted polymerization.

The PCR products described above were purified on a 1% agrose gel, digested by both *Bam*H I and *Hind* III enzymes (BioLabs), and then ligated into double-digested pUC18 vector to generate the recombinant plasmid. The plamid was transformed into the *E. coli* cloning strain DH5α. After the cloned sequence was verified by DNA sequencing, the plasmids were isolated in a large-scale (~2

mg) and then linearized by *Bst*N I enzyme (BioLabs) to generate the linear templates for subsequent *in vitro* transcription.

5.2.2 *In vitro* transcription and RNA purification

The RNA was produced using T7-based *in vitro* runoff transcription (optimized) and purified by preparative scale 5% (w/v) polyacrylamide gel electrophoresis and electroelution using Bio-Rad Green Membrane (Model 422 Electro-Eluter). The eluted RNA was ethyl alcohol precipitated and re-dissolved in the NMR buffer containing 10 mM sodium phosphate (pH 6.0), 80 mM NaCl, 20 mM NaN_3 and 90% H_2O/10% D_2O.

5.3 RESULTS AND DISCUSSION

5.3.1 Interaction between SmpB and the tRNA-like domain of tmRNA

The 1D NMR spectra of all three RNA samples (~0.1 mM each) were collected with water signal suppressed using the jump-return water suppression method. Samples of the SmpB-RNA complexes for NMR spectroscopy were prepared by gradually adding ^{15}N-labeled full-length SmpB protein (in the same NMR buffer as RNA) to each RNA sample. However, a difficult situation arose at this point: white precipitate came out of the solution immediately after the protein was added. To continue the experiments, concentrated salt (5 M NaCl) was added little by little until all precipitate disappeared. The final salt concentration was calculated to be about 0.5 M. The SmpB protein was added in two steps, and three

equivalents of total SmpB with respect to RNA were applied. A 1D RNA spectrum as well as A 2D HSMQC spectrum of the ^{15}N-labeled protein were recorded at each step. To compare the spectra of the free and complexed states, a 1D spectrum of the free RNA and a 2D HSMQC spectrum of the free-state protein were recollected in the presence of the same concentration of NaCl so as to exclude chemical shift changes induced by the increase in salt concentration.

Figure 5.4 shows the 1D NMR spectra of the imino proton resonances of C1t, C2t and C3t, respectively. Based on the dispersion and width of the resonances, it is concluded that C1t and C2t are likely to be properly folded, but C3t seems to aggregate.

The results of the titration of C1t by SmpB protein are shown in figure 5.5. Several significant changes in terms of the chemical shifts of the imino protons in the RNA were detected, which implies that SmpB may interact with the RNA. However, the ^{1}H-^{15}N HSMQC spectra of both the 1:1 and 1:3 mixtures did not show any detectable cross peaks. Apparently, the missing of cross peaks was not due to the low concentration of the protein since the ^{15}N-enriched SmpB by itself showed almost all identified peaks at an even lower concentration (data not shown). It can be speculated that aggregation occurs in the mixture, which slows down the tumbling rate of the protein or the complex, and thus decreases the T2 relaxation time of the nuclei and makes the cross-peaks undetectable.

The same phenomena have been observed for the construct C2t (data not shown). NMR perturbation experiments were not performed on the construct C3t because of the likely aggregation detected from its 1D spectrum (Figure 5.4).

5.3.2 Interaction between SmpB and the 12 base-pair RNA duplex

A 12-base pair RNA duplex (shown in Figure 5.10) has been designed by Wower *et al.* (2002). They proposed that it constitutes the primary binding site of SmpB on tmRNA based on their binding affinity and cross-linking studies. The two separate 12-base RNA oligos were provided by Dr. Wower (chemically synthesized by Dharmacon Research Inc.). The duplex was prepared by mixing the two strands together while monitoring the 1D NMR spectrum of imino proton resonances (10-15 ppm on a 500 MHz NMR spectrometer) and comparing it to the spectrum of each single strand. The final titration product has ~1:1 ratio of both stands in 10 mM phosphate buffer (pH 6.0) containing 100 mM NaCl and 90% H_2O/10% D_2O.

^{15}N-enriched full-length *A. aeolicus* SmpB (in the same buffer as the RNA) was added to above RNA solution gradually. However, white precipitate came out of the solution after just a little bit protein was applied. The salt concentration was increased to ~0.6 M to solubilize the precipitate. The SmpB protein was added in two steps. In the first step, 0.25 equivalent (with respect to RNA) of SmpB was added and a 1D NMR spectrum was collected for RNA. In the second step, another 0.25 equivalent of SmpB was added to the RNA/SmpB solution from the first step and a 1D NMR spectrum (for RNA) as well as a ^{1}H-^{15}N HSMQC spectrum (for protein) was recorded. Comparison of the 1D spectra reveals a few substantial changes for some imino proton resonances of the RNA (Figure 5.6), which suggests that the protein interacts with the RNA. However, examination of the HSMQC spectra of free and RNA-bound states of SmpB did

not identify significant chemical shift perturbations on the backbone nitrogen atoms and amide protons that could be attributed to the bound RNA (Figure 5.7). The two notable changes occurred on residues His88 and Arg90, which may be caused by small changes in the pH of the solution (noting that histidine has a pKa near 6.5). Addition of Mg^{2+}, which may be required for RNA folding, to the RNA/protein mixture did not lead to any detectable chemical shift changes (Figure 5.8). Figure 5.9 shows the overlay of the HSMQC spectra of *A. aeolicus* SmpB at low and high salt concentrations. A majority of the corresponding chemical shifts match pretty well, suggesting that the structure did not significantly change with increased salt. For those resonances showing notable chemical shift differences, none of them are highly conserved (Figure 4.5) and are thus believed to be not essential for the structure and function of the protein.

Apparently, there is a contradiction between the interactions detected in the 1D NMR spectrum of the RNA and the unperturbed chemical shifts in the 2D HSMQC spectrum of the protein upon RNA binding. One explanation is that there is interaction between the protein and RNA but the perturbation is not detectable in the HSMQC spectra. RNA-protein interactions are likely to be predominantly mediated through the side chains of the protein residues, therefore, lack of chemical shift changes for the backbone cannot necessarily be interpreted as lack of interaction between the two components. 3D heteronuclear spectra of the complex, which provide chemical shift information of the side chains, may be worth analyzing in order to re-assess the results obtained in the current chemical

shift studies. Another possibility is that the binding is not specific at the high concentration required for NMR analysis.

Comparison of the sequences of the RNA duplex, the acceptor arm/T stem loop and the tRNA-like domain shows that the duplex is actually the "straightened" acceptor arm/T-stem loop of the tRNA-like domain of *E. coli* tmRNA, with the 3'-end single strand part and the T loop missing (Figure 5.10). This duplex sequence presumably adopts a straight double-stranded helix; however, in the 3D structure of tmRNA the acceptor arm/T stem loop forms a "broken" helix (Stagg *et al.*, 2001; Zwieb *et al.*, 2001). This subtle conformation change could account for the abolishment or change of the specific binding between SmpB and the acceptor arm/T-stem loop of tmRNA. In addition, Wower *et al.* (2002) proposed that there are three SmpB-binding sites on the tRNA-like domain of tmRNA. Therefore, the undetected interaction between SmpB and the 12-base pair RNA duplex may be attributed to the limited size of the RNA, which is analogous to one of the three proposed binding sites.

5.3.3 Interaction between SmpB and yeast tRNA

Commercial yeast tRNA mixture was kindly provided by Dr Browning's laboratory. The dry tRNA was dissolved in 10 mM phosphate buffer (pH 6.0) containing 100 mM NaCl and 90%H_2O/10%D_2O (the typical NMR buffer). A pre-testing experiment mixing a small volume of the tRNA solution with SmpB (in the same buffer) showed that heavy precipitates formed in the solution as expected on the basis of previous experience with the tRNA-like domain and the

12-bp duplex. 5 M NaCl was added to both the tRNA and the SmpB solutions to a concentration pre-determined in the pre-testing experiment, which was just enough to prevent forming precipitates. The final concentration of NaCl was approximately 0.6 mM.

NMR perturbation experiments were carried out by adding ~2 mM (35.9 mg/ml) ^{15}N uniformly labeled full-length *A. aeolicus* SmpB to 0.5 ml of ~0.4 mM (12 mg/ml) yeast tRNA solution in three steps. In each step, 50 μl of the protein solution was applied, which is equivalent to 0.5 of the total tRNA. Thus, the ratios of tRNA and SmpB in the first, second and third step were 1:0.5, 1:1 and 1:1.5, respectively. A 1D NMR spectrum of the imino proton resonances of the tRNA was collected (20°C) in each step, and a 2D HSMQC spectrum of ^{15}N SmpB was lastly recorded (30°C) in the third step.

Due to the multiple species of tRNA, the imino proton resonances in the 1D spectra are clustered between ~10-15ppm, which makes it impossible to distinguish the resonance shifts in the RNA induced by SmpB binding (Figure 5.11). Nevertheless, the ^{1}H-^{15}N HSQMC spectrum of the 1:1.5 (tRNA:SmpB) explicitly demonstrates the interactions between the protein and the tRNA (Figure 5.12). In the spectrum, most of the detected cross peaks did not change upon tRNA binding, but a few new cross peaks emerged after the addition of tRNA. These cross peaks are presumably the indication of successful perturbations, because if there is no interaction at all between the protein and the RNA, no new resonances should be seen even if the protein is in excess. Unfortunately, it is hard, if not impossible, to map the perturbations on the protein structure at this

stage, because most cross peaks present in the spectrum of free state SmpB disappeared after applying tRNA, thereby making sequence specific assignments become enormously difficult.

A titration of SmpB with a single species of tRNA will be worth trying in the future to distinguish the resonance perturbations on the 1D spectrum of the RNA by SmpB, and to assign the chemical shift changes on the HSMQC spectrum to specific residues on the protein, because it will exclude the differentiation of the interactions between the protein and different tRNA species. Also, a set of 3D heteronuclear NMR spectra may be worth analyzing for the complex in case that the chemical shift perturbations induced by the mutual interactions cannot be detected on the backbone atoms.

End of Chapter 5

Chapter 6

Preliminary X-ray crystallographic studies of *A. aeolicus* small protein B

6.1 PROTEIN PREPARATION AND CRYSTALLIZATION

Full-length *A. aeolicus* SmpB protein was purified as described earlier (section 2.2). The sample for crystallization was concentrated to ~10 mg/ml by Centricon YM-10 in the buffer of 10 mM sodium phosphate (pH 6.0) containing 100 mM NaCl. Preliminary NMR studies indicated that the protein is properly folded and behaves as a monomer (Figure 1.6). However, extensive screening by sparse-matrix searches in hope of finding appropriate crystallization conditions, including variations of reservoir solutions (Crystal Screen I and II – Hampton Research, Magic 96, Ammonium Sulfate screen, PEG screen, etc.) and temperature (room temperature, 16°C and 4°C), all failed to yield any crystals.

The short versions of the SmpB protein were then prepared after stable short products were obtained by trypsin treatment. Two fragments, corresponding to residues 1-133 and 3-133, were generated and purified as described in chapter 3. The samples for crystallization were concentrated to 8-12 mg/ml in solution containing only 100 mM NaCl. Initial screens to establish crystallization conditions were carried out by sitting-drop vapor-diffusion using Crystal Screens

I and II (Hampton Research); the droplets contained 1-2 µl of protein mixed with the same volume of the reservoir solutions. After incubation at room temperature for 3 to 7 days, several conditions gave crystals with varying shapes and sizes (Figure 6.1). Further optimizations failed to solve the twin and rough-face problems, and the little crystal grains were reluctant to grow bigger. However, the clustered long needles were successfully tamed and crystals suitable for diffraction experiments were obtained. The crystals were grown by the vapor diffusion method in sitting-drops at room temperature. The reservoir solution contained 0.1-0.2 M ammonium sulfate and 28-32%(w/v) polyethylene glycol monomethyl ether ((PEG-MME) 5,000 in 0.1 M HEPES, pH 6-7. The drops contained 1-2 µl of protein and the same volume of the reservoir solution. Visible crystals grew from the solution after 1 to 2 days and continued growing to maximum size after further 5 to 7 days. The crystal habit is a tetragonal prism of variable length to 0.8 mm and cross section 0.1 mm × 0.1 mm (Figure 6.2).

6.2 DATA COLLECTION AND PROCESSING

The crystals were transferred to a solution containing 80% reservoir solution and 20%(*v*/*v*) glycerol, equilibrated for 2 minutes, mounted on cryo-loops and then flash-frozen in a nitrogen gas stream at 100 K (Cryostream, Oxford Cryosystems). Diffraction data were collected on an imaging plate (MAR Research) using X-rays generated by a Rigaku RU200 rotating anode generator (Molecular Structure Corp., The Woodlands, TX), which was operated at 50 kV and 100 mA. The crystal-to-detector distance was 128 mm. The oscillation angle

was 3 degrees per image, and the exposure time was 15 minutes for each single image. The crystals diffracted to a maximum resolution of approximately 2.8 Å (Figure 6.3). The data were processed and integrated by the program *MOSFLM* (Leslie, 1992). A total of 36143 observations were reduced to a unique set of 5942 reflections.

6.3 PRELIMINARY RESULTS AND DISCUSSION

Preliminary analysis indicates that the crystals belong to a tetragonal lattice, with unit cell parameters a = b = 55.0 Å, c = 65.9 Å, $\alpha = \beta = \gamma = 90°$. The space group cannot be unambiguously assigned at this point. However, examination of the diffraction pattern excluded the two body centered space groups I422 and $I4_122$. Other than that, it could belong to any one of the 14 space groups ($P4_n$, n = 0, 1, 2, 3; $P4_n2_m2$, n = 0, 1, 2, 3 and m = 0, 1). Analysis of the Matthews coefficient (V_M) (Matthews, 1968) provided more information to further narrow down the possibilities to the first 4 space groups ($P4_n$, n = 0, 1, 2, 3). The calculated V_M is 3.26 $Å^3$/Da if the space group is $P4_n$ (4 molecules per unit cell or 1 molecule per asymmetric unit); otherwise, V_M is 1.63 $Å^3$/Da with the space group being $P4_n2_m2$ (8 molecules per unit cell or 2 molecules per asymmetric unit). These two V_M values correspond to the calculated solvent contents of 62% and 24%, respectively. Experience tells that crystals with a solvent content of only 24% usually give very high diffraction resolution. Thus, the relative low diffraction resolution of the SmpB crystals suggests that the crystals most likely have a looser packing mode. It is concluded that the crystals

likely belong to the space group $P4_n$ (n = 0, 1, 2 or 3), although the higher symmetry tetragonal space groups remain a less likely possibility.

The X-ray crystal structure of *A. aeolicus* SmpB is expected to be very similar to the NMR structure provided by current work; therefore, the first logical step for solving the crystal structure is to try molecular replacement (MR) method using the current NMR structure as the search model. In MR calculations, atomic coordinates and crystallographic B-factors are both important in providing information required for calculating the scattering contributions by each atom. An individual NMR model does not contain B-factors. However, information that may be similar to the crystallographic B-factors is embodied in an NMR ensemble: well determined parts have smaller variations in atomic positions, and poorly defined regions exhibit larger variations (reviewed by Chen, 2001). The first application of solving crystal structures using NMR structures as the search model was published ten years ago (Baldwin *et al.*, 1991). Two approaches are now usually used for solving crystal structures using NMR structures as the search models. The first method uses a single conformer with artificial B-factors assigned according to the atomic r.m.s. deviations of individual atoms. The second approach involves using a set of best-defined models as a composite model, with all atoms assigned uniform B-factors. This composite model provides weights to the mutual agreement of equivalent atomic positions (reviewed by Chen, 2001).

Based on the principles described above, the process of solving the crystal structure of *A. aeolicus* SmpB has been launched and is now in progress.

Meanwhile, SeMet-substituted *A. aeolicus* SmpB is being prepared with a view to phasing by multiple wavelength anomalous dispersion (MAD). A search for suitable heavy-atom derivatives may provide an alternative route for solving the crystal structure.

End of Chapter 6

Chapter 7

Conclusion and perspectives

Although either small protein B or tmRNA are not essential for cell growth, their wide presence in prokaryotes suggests that the tmRNA-SmpB quality control system plays an important role in living cells. Significantly, the appearance of SmpB and tmRNA in the organism *Mycoplasma genitalium*, which has the smallest number of genes, provides additional evidence suggesting their importance. Primary sequence comparison indicates that SmpB is a unique RNA-binding protein with little similarity to other known proteins. This dissertation presents the first structural analysis of this type of RNA-binding protein and examines the possibilities of how SmpB participates in the *trans*-translation process. However, a detailed mechanism of *trans*-translation process, such as how the tmRNA-SmpB complex is recruited to the stalled ribosomes, how translation proceeds through the ORF of tmRNA, and what the functions of SmpB and other protein factors are in the process, still remains a mystery.

Future work will necessarily involve multiple approaches. Site-directed mutagenesis based on the current work seems to be the next logical step for determining the residues on SmpB that are responsible for the protein-RNA recognition. *In vitro* and *in vivo* biochemical studies will provide evidence for the

interactions between different components within the RNP complex and between the complex and the translational apparatus. The solution of the structures of 70S ribosome complexed with different ligands points out a possible way for elucidating the structures of tmRNA-SmpB-bound-ribosomes at different stages of *trans*-translation process, provided that the tmRNA-SmpB can be trapped onto the ribosome and these giant complexes can give good-quality crystals.

End of Chapter 7
