1. Introduction

Recombination and plasmid maintenance

The horizontal gene pool in bacteria is composed of an array of accessory mobile genetic elements that profoundly influence genome plasticity, organization, and evolution (*1*). Plasmids are extrachromosomal autonomously replicating members of this pool important for bacterial adaptability and persistence because they provide functions that might not be encoded to chromosome (*2*). Although plasmids can be highly beneficial for the host bacteria by allowing them to persist in otherwise hostile ecological niches, pathogenic properties are also endowed by plasmids. The clinical failure of antibiotics in recent years is in part linked to the rapid dispersal of plasmid-borne resistance genes in bacterial populations. Virulence determinants in bacterial pathogens may also be plasmid-located and similarly dispersed. A typical example of the contribution of plasmids to bacterial diversity is represented by the food-borne pathogen *Bacillus cereus*, the anthrax agent *B. anthracis*, and the insect pathogen and commercial pesticide *B. thuringiensis*; these are very closely related species but, from a human perspective, have quite distinct biological properties. These differences reflect, at least in part, differences in these species plasmid contents (*3*).

Plasmids are separate, autonomous genetic elements present in a cell independently of chromosomes. Most plasmids are small, from several to 100 kb, but sometimes they are so large that using the size criteria their distinction from the chromosome is difficult. Certain plasmids can constitute even up to 50% of bacterial DNA (*4, 5*). It is commonly accepted that plasmid genes do not encode information indispensable for the functioning of the host cell. However, plasmids specify numerous features advantageous for the host in specific environments, such as resistance to harmful agents, ability to degrade rare compounds,

This dissertaion follows the style and format of the journal *Biochemistry*.

pathogenicity, toxin production, nitrogen assimilation etc. Plasmids are very widely distributed among prokaryotes and, in general, are inherited with a high degree of stability. In special environmental conditions some plasmid genes confer a selective advantage. But, in many cases, plasmids are retained over generations without any selective pressure. Thus, there have to exist mechanisms which enable the maintenance of the plasmid during cell growth in nonselective conditions. Systems that contribute to this stability are encoded by DNA cassettes and are, in most cases, independent of one another. A particular plasmid can carry different stabilizing cassettes. Even more, cassettes from different plasmids may be combined to give a stable replicon. So the next fascinating thing to know is how the plasmid distribution takes place from one generation to another.

Random and better-than-random plasmid distribution

During the process of cell division plasmid copies are distributed between descendants. If plasmid distribution within a growing cell is random, in the “ideal replicon”, it is the number of plasmid molecules inside the dividing cell that determines the probability (P_0) that one of the daughter will be plasmid free: $P_0 = 2^{(1-n)}$, where n is the number of plasmid copies (*6*). Therefore, a cell having 30 plasmid copies at the time of division would produce in 10^9 cells and only two-plasmid free ones, whereas in the case of 5 copies, the chance of giving a plasmid-free cell is as much as 1 per 16.

As long as the plasmid copy number remains high, the subpopulation of plasmid-free cells is extremely limited. For low copy number plasmids obeying the random distribution law would mean that a significant fraction of daughter cells will be plasmid free. This is in contrast to the observed maintenance of plasmids in the host cells. Therefore considering this flaw in hypothesis, several mechanisms have been proposed, and shown to operate, to explain this

discrepancy. Broadly evolved hypothesis or mechanisms can be divided into two major classes (*7*):

A. **Site-specific recombination systems**: Mechanism ensures with the help of recombinase proteins that plasmid multimers arisen during replication and/or recombination will be resolved and thus every monomer copy will be independently subjected to random distribution.

B. **Plasmid addiction systems and active partitioning process**: To kill or reduce growth of plasmid free descendant cells. This precisely distributes plasmid copies to each daughter cell at division.

While mechanisms stated in Class A lead to the optimization of random plasmid distribution, whereas class B mechanism ensure better than random plasmid inheritance. I tried to confine all important information, since an extensive review of plasmid distribution in cell is not the purpose of this thesis, reader can refer to review articles by F. Hayes, K. Gerdes and others for early and detailed description of plasmid distribution field (*2, 7-10*). Here in brief, I would like to discuss only the components of site-specific recombination and plasmid addiction systems which deeply concerns the work mentioned in this report.

1. (A) Site-specific recombination systems:

Recombination is a potent evolutionary force that shapes and reshapes the genomes of all organisms. More than 38 years ago, Holliday proposed a model of meiotic recombination in which homologous chromatids exchange single DNA strands to form a partially heteroduplex joint molecule with a four-way junction at the point of exchange (*11*) (Figure 1.1). Resolution of the Holliday junction by a symmetrical strand cleavage, coupled with repair of any DNA base pair mismatches in the heteroduplex regions, provided a plausible explanation for

the formation of recombinants and for the patterns of marker segregation in genetic crosses. Although many features of the model have since been found wanting, the idea that recombinant chromosomes arise through the formation and subsequent resolution of Holliday intermediates has withstood the test of time (*12*). The reaction pathway for *Escherichia coli* has already been dissected in detail. With completion of genome sequencing of many other organisms, eukaryotic homologs and analogs of other prokaryotes like *E. coli*, *Streptococcus pyogenes* protein(s) are emerging (*13-16*), which leads to the expectation that the key features of the reaction mechanism will generally, if not universally, applicable.

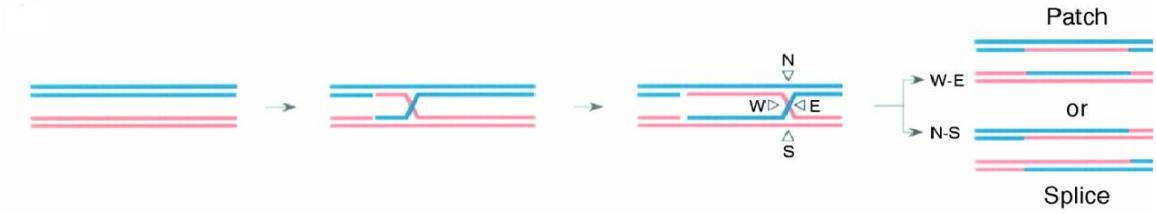

Figure 1.1: The formation and resolution of a Holliday junction in homologous recombination in *E. coli*. Schematic representation of the rearrangement undergone by DNA during homologous recombination. The parent duplexes are blue and pink, and the pairs of RuvC cleavage sites are marked **N** and **S** and **W** and **E**.

Conservative site-specific recombination involves the reciprocal exchange of DNA segments by precise breakage and rejoining processes that involve no loss or synthesis of DNA. In principle, such events can occur intermolecularly (resulting in fusion of the two recombining partners) or intramolecularly (resulting in inversion or excision of one DNA segment relative to the other), although in most biological systems this directionality is strictly controlled. Biological roles of site-specific recombination include chromosomal integration and excision of phage genomes, monomerization of plasmid chromosomes, alternation of gene expression, resolution of transposition intermediates, and fusion of gene cassettes into a functional gene. Site-specific DNA recombination reactions are catalyzed by number of recombination proteins (recombinases) and polynucleotidyl transferases that are specific to each reaction system (*17*).

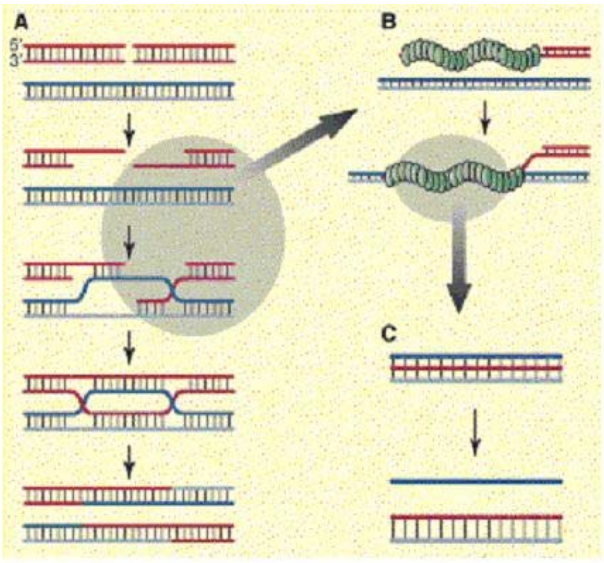

Figure 1.2: Schematic representation of homologous DNA recombination and recombinase function (18)**.** (**A**) A model for one of the functions of homologous recombination, DNA double-strand break repair, illustrated at the level of DNA. Duplex DNA molecules are indicated by the ladders with the rungs representing base pairs. Processing of double-stranded DNA ends result in 3′ single-stranded tails onto which the recombinase proteins form helical filaments. The grey oval represents the part of the process that is shown in more detail in B. (**B**) A helical nucleoprotein filament formed on single-stranded DNA can recognize homologous sequence in intact double-stranded DNA resulting in pairing of the two DNA molecules. (**C**) Within the context of the filament, DNA strand exchange takes place.

In accordance to the work shown in this report, I would like to narrow down to the two major families of site-specific recombination; known as ***tyrosine and serine recombinases*** using different mechanisms for cutting and rejoining the DNA strands at the recombination crossover sites. The tyrosine recombinases complete the cleavage, exchange and rejoining of one pair of DNA strands with the generation of a Holliday junction as a recombination intermediate, before initiating the same set of reaction on the other pair of DNA strands. If the recombination sites are on different DNA molecules, the net result of the process is the integration of the two molecules into a single one (Figure 1.3). However, if the recombination sites are on the same DNA molecule, the recombination reaction can lead to either the deletion or the inversion of the DNA

sequences located between the crossover sites, depending on their relative orientation and the number of interdomainal supercoils trapped on synapsis (Figure 1.2 and 1.3) (*19-21*).

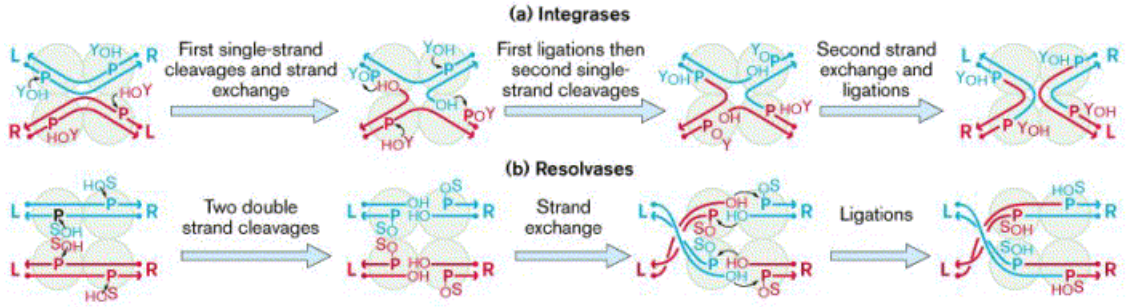

Figure 1.3: Pictorial representation of functioning of Integrases and Resolvases group of recombinases.

The family of recombinases, which catalyze intramolecular strand exchange by a non-replicative double-strand DNA breakage and rejoining mechanism, are known as ***serine recombinase*** family. Recombination involves a 2-bp staggered break across all four DNA strands and the formation of covalent phosphoserine linkages between the DNA strands and recombinase subunits. Two subfamilies of serine recombinases are known. One includes small recombinases (<250 amino acids long) that primarily catalyze intramolecular recombination reactions (*21*). The second subfamily comprises large recombinases (>400 amino acids long) that catalyze both inter- and intramolecular recombination events (*22*). The small Serine recombinase family can be classified into three distinct groups depending on whether they catalyze ***resolution*** (DNA resolvases; γδ, Tn*3* and Sin as prototypes), ***inversion*** (DNA invertases; Hin and Gin as prototypes) or both (***resolvo-invertases***; β as prototype) (Figure 1.3) (*19, 21, 23, 24*). Recombination proceeds via a concerted four-strand cleavage and rejoining mechanism (*19, 21, 25*). These site-specific recombinases form a protein-DNA synaptic complex that includes the two crossover sites, each bound by the recombinase and one or more accessory binding and/or accessory protein(s) (*21, 24, 26, 27*).

1. Resolvases

Resolvases selectively catalyze excision between two directly oriented *res* sites. The *res* site, 114 bp in length, contains three adjacent subsites with dyad axis symmetry, named I, II and III. Each subsite binds a resolvase dimer. Resolvases, which are highly specialized in catalyzing deletions, act on recombination sites (*res*) that are arranged in direct repeat orientation and in the same supercoiled molecule (*21, 24*).

Assembly of a synaptic complex requires binding of three resolvase dimers to each *res* site and DNA supercoiling. No additional factors are needed. Three highly condensed supercoils are trapped in the synaptic complex through interactions of subsites II and III. Recombination takes place at the centre of subsite I. Recently crystal structure of a synaptic $\gamma\delta$ resolvase tetramer covalently linked to two cleaved DNAs has also been revealed (*26*). In the resolution system one (Sin) or two (Tn*3*, $\gamma\delta$, Tn*21*) additional binding sites for two or four additional recombinase units are located close to each crossover site (*21, 24, 28*). In the Sin recombinase the absence of subsite III might be substituted by the essential architectural Hbsu protein (DNA-binding histone-like protein) that binds between subsite I and II (*28*).

2. DNA invertases

DNA invertases (e.g., Hin and Gin), which promote inversions on a supercoiled substrate, act on an inverted recombination site in the presence of a *cis*-acting enhancer bound by two Fis dimers and the HU accessory protein. DNA invertases specifically catalyse inversion between two inversely oriented recombination sites. Invertase dimers bound to these sites can interact to form an inactive synaptic complex, trapping two interdomainal supercoils requires a 65 bp enhancer located on the same DNA molecule, to which FIS binds (*29-31*). Mutant derivatives of DNA resolvases and DNA invertases that are permissive

for inter- and intramolecular events have been isolated. These activated mutant recombinases bypass the requirement for a synaptic structure to initiate recombination and can thus catalyze both resolution and inversion (*32-34*).

3. Resolvo-invertases

Resolvo-invertases are peculiar because they do not show the same selectivity as resolvases and invertases for resolution or inversion (*28, 35*). Resolvo-invertases, whicḥ do not have biases like other, because they can catalyze deletions between two directly oriented recombination sites (*six*) and both deletions and inversions between two inversely oriented *six* sites in the presence of the essential architectural Hbsu protein (*23, 36*). The synaptic complex in resolvo-invertases system is composed by the crossover site bound by a recombinase dimer and one additional binding site (subsites II) for one additional dimer, and an intrisically curved region, which is the target for the architectural Hbsu protein. The resolution reaction requires a supercoiled DNA substrate, but inversion works even with linear DNA (*36, 37*). All these facts dictate the characteristic selectivity of the system.

The 23.8 kDa β recombinase (205 amino acids residues) catalyzes recombination at a 90-bp *six* site both in bacterial and in eukaryotic systems *in vivo* and *in vitro* (*35, 38-40*). β recombinase mediates inversion during replication and flips the orientation of one replication fork with respect to the other, so that both forks travel in the same direction around a circular template avoiding the collision of the two replication machineries after initiation at the two inversely oriented replication origins, and most importantly the resolution of pSM19035 dimers maximized plasmid segregation (*41-43*).

In principle, the absence of bias on substrate utilization by the β protein could be determined either by the protein structure, by the architecture of the DNA substrate, or by both. The synaptosome/invertosome complex at the cross-

over site might be conserved among small serine recombinases (*26, 27, 33*). The β recombinase shows 32% sequence identity with γδ resolvase. Extensive mutational, biochemical and crystallographic analyses of the γδ resolvase have allowed the identification of domains of the protein that participate either in the formation of the active centre, in dimerization, in the interaction of dimers when bound to its target sites, and in specific DNA recognition (*21*).

Like γδ resolvase (*26, 44-46*), limited proteolysis and mutational analysis revealed that the β recombinase has the C-terminal DNA binding domain and an N-terminal catalytic domain, followed by an extended α-arm (the E helix), that promotes the formation of a dimer (*47, 48*). Recently it was shown that a synaptic tetramer of a γδ mutant enzymes, which is capable of recombining linear DNA substrates containing only subsite I, is bound to two subsites I held together by a flat interface consisting of two four-helix bundles (two D and E α-helix pairs) (*26*). Unlike wild type β protein, which forms stable dimers in solution at physiological concentrations, wild type resolvases and DNA invertases exist mainly as monomer in solution (*49, 50*). However, resolvase (*33, 34, 51*) and DNA invertase (*32, 52, 53*) mutant proteins, which have been shown as dimers in solution, have in common with the wild type β protein that they are capable of promoting deletions and inversions and even recombine using linear DNA substrates.

To learn whether the behaviour of β recombinase in solution provide some insight in the recombination reaction different biophysical methods were used. In this work, the thermal and equilibrium denaturation of *Streptococcus pyogenes* β recombinase is reported keeping in mind the evaluated oligomerization behaviour and solution conformation of the protein. We propose models for the thermal and denaturant induced unfolding, which suggest the presence of a folded monomeric intermediate.

1. (B) Plasmid addiction systems:

In simplest words, prokaryotic chromosomes contain toxin -antitoxin loci, termed “addiction modules”, which are composed of two genes organized in an operon that encode a stable toxin and a labile cognate antitoxin, respectively (*54*). In steady state, antitoxins neutralize the effects of toxins by direct protein-protein interaction (*55*). In addition, antitoxins and toxin -antitoxin complexes bind to their promoter DNAs within their own operons and negatively regulate their own transcription (*56*). Hence antitoxins functions as repressor of operon transcription and toxins act as co-repressors. Upon induction by environmental stresses, such as amino acid and carbon source limitation, labile antitoxins are degraded by the ATP-dependent protease (Lon) or the bacterial proteasome system, there leading to rapid growth arrest and cell death by cellular effects of toxins (Figure 1.4) (*57*).

Although there is considerable heterogeneity in the physical and genetic properties of plasmids, large, low-copy-number plasmids associated with antibiotic resistance and pathogenicity tend to include core regions required for replication, segregational stability, and conjugative transfer (*58*). The segregational stability of these and other plasmids in bacterial populations is achieved by the activity of plasmid-specified partitioning proteins that direct plasmid copies to new daughter cells at cell division. Plasmid-directed events resulting in selective killing or growth impairment of cells that have failed to acquire a plasmid copy were identified in the 1980’s (*59, 60*). These mechanisms confer an advantage on plasmid-retaining cells by reducing the competitiveness of their plasmid-free counterparts, thereby ensuring the retention of the plasmid in the population. In common with eukaryotic chromosomes, bacterial plasmids have centromeres that function to segregate plasmid molecules prior to cell division (*61, 62*). However, plasmids also encode maintenance loci whose gene products are activated in plasmid-free cells. These loci function to prevent the

proliferation of plasmid-free progeny (*60, 63*), which increases plasmid maintenance in growing bacterial cultures.

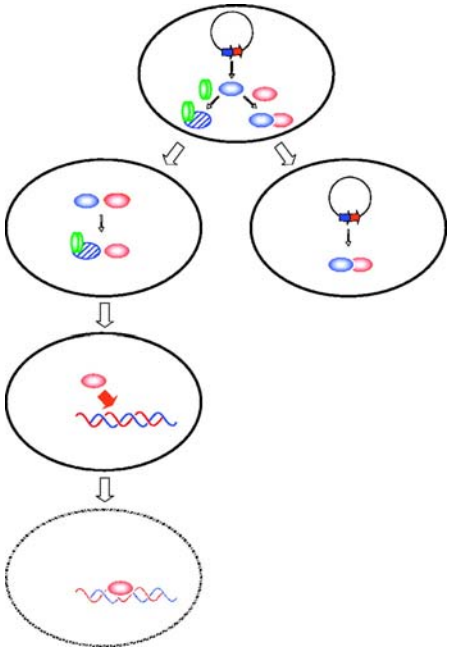

Figure 1.4: Schematic representation of cell death induced by plasmid-located type II TA modules (58). The toxin (red) and antitoxin (blue) proteins form a tight complex that negates the harmful activity of the toxin. The antitoxin is degraded by a protease (green) more rapidly than the toxin, but the latter is continually sequestered by fresh antitoxin. As long as the plasmid is maintained, the cell tolerates the presence of the TA complex (**right**). If a missegregation event or replication defect produces a plasmid-free cell (**left**), the degraded antitoxin cannot be replenished so that the liberated toxin attacks an intracellular target to cause death or growth restriction of the plasmid-free cell. The targeting of DNA by the toxin is illustrative only.

Toxin-encoding loci stabilizes plasmids

Two types of loci that prevent growth of plasmid-free progeny have been characterized: those that are regulated by **antisense RNAs** (*60, 64*) and those that are regulated by protein **antitoxins** (Figure 1.4) (*63*). The antisense RNA-regulated loci encode proteins that kill plasmid-free progeny by damaging the bacterial cell membrane. This unusual phenotype was named post-segregational killing (PSK) (*60*). The complex antisense RNA-regulated RNA folding

mechanism that controls PSK has been reviewed in depth by Greenfield and colleagues (*64*) and to describe it is not in interest of the work presented in this report.

The idea of an “addiction” mechanism leading to very efficient plasmid maintenance comes from Koyama (*65*). In his considerations he pointed out that if cells that lose an established plasmid die, the population would never contain viable cured cells. The protein antitoxin-regulated systems have several names, such as killing -antikilling, post-segregational killing, poison -antidote, toxin -antitoxin, plasmid addiction system or programmed cell death (PCD). All are used to describe the situation when the host cell is selectively killed if it has not received any copy of the plasmid. There has been a recent resurgence of interest in these bacterial TA systems because of new insights that have been acquired into these events, but also because of enhanced appreciation of their widespread distribution both on plasmids of medical importance and on bacterial chromosomes (*7, 54, 66-69*). Now these addiction mechanisms are better known as postsegregational cell killing (PSK) or toxin -antitoxin (TA) systems where attack is within the cells. For brevity and consistency, now on we use nomenclature as suggested by Gerdes and colleagues *i.e.* toxin −antitoxin (TA) loci (*62*). The cell killing due to TA loci is in contrast to the action of colicins or antibiotics that are secreted by bacteria into their environment as inhibitors of neighboring microorganisms. The toxin component produced by TA cassettes is designed to maim bacterial cells, which raises the exciting possibility that these factors might be exploited as novel antibacterial agents in the treatment of infectious diseases. Restriction-modification enzyme pairs can be either plasmid- or chromosomally encoded and are also now viewed as multifunctional TA systems that can promote segregational stability, as well as providing protection against invading DNAs and directing genome rearrangements. The restriction enzyme is analogous to the toxin; the modification methylase is equivalent to the antitoxin (*70*).

TA systems and Plasmid maintenance

TA cassettes have a characteristic organization in which the gene for the antitoxin component precedes the toxin gene (Figure 1.5); the two loci often overlap, reflecting a common autoregulatory mechanism exerted by both components. Although most TA modules conform to this arrangement, there are examples of TA cassettes in which the gene order is reversed, where the antitoxin alone exerts the regulatory effect or where the product of a third gene is implicated (*54, 66, 71*). The mechanisms that control plasmid maintenance by PSK and TA loci are similar: the antagonistic regulators that neutralize the toxins, whether they are antisense RNAs (for PSK loci) or proteins (for TA loci), are metabolically unstable. Rapid depletion of these unstable regulators occurs in newborn plasmid-free cells. As the same cells have inherited stable toxin molecules (or toxin-encoding mRNAs) from the mother cell, the toxin will no longer be removed by the antitoxin, therefore leading to killing (PSK loci) or stasis (TA loci) of plasmid-free cells. In this way, PSK and TA loci prevent the proliferation of plasmid-free cells in growing bacterial cultures.

A recent belief and literature indicates that when cells are subjected to extreme amino acid starvation, above described mechanisms gets into action. Mechanisms rapidly adjust the rate of protein and DNA synthesis. Prokaryotic genomes contain TA loci that might fulfill this function. However, recent database mining has shown that TA loci are ubiquitous in free-living prokaryotic cells (*72*). Obviously, chromosomal TA loci do not function to stabilize plasmids. Recent evidences indicate that TA loci function to modulate exposure to nutritional stress.

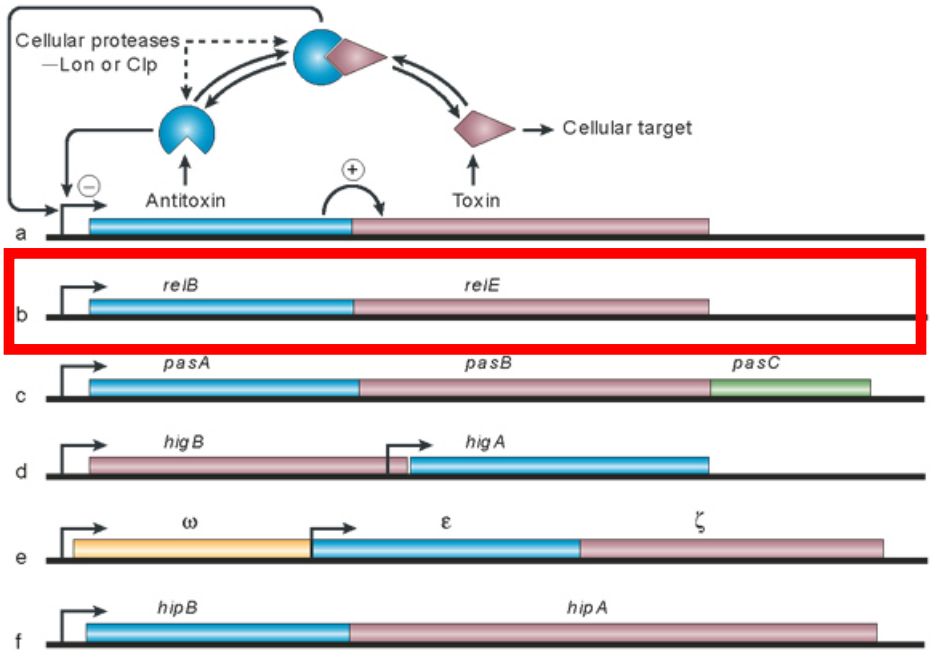

Figure 1.5: Genetic organization and components of TA loci (62). Toxin genes and components are shown in pink, antitoxin genes and components in blue. (a) General genetic setup and components encoded by a typical TA locus. Stippled arrows indicate that cellular protease degrade the antitoxin, either free in solution or in complex with toxin. Arrows pointing rightwards indicate a promoter upstream of the TA operon. The arrows pointing at the rightwards-pointing arrow indicate that the antitoxin and the TA complex bind to the promoter region and repress transcription. The arced arrow indicates translational coupling between the antitoxin and toxin genes. (b-f) Genetic organization of a typical *relBE* locus and other TA systems.

Eight families of TA loci in bacteria

TA loci have been grouped into seven two-component gene families plus one three-component system (ω-ε-ξ) (Table 1.1). Members of all eight families are found on plasmids and chromosomes (*54, 62, 72*). Here we tried to summarize all TA systems in brief.

1. The *ccd* locus of the F plasmid

The *ccd* (coupled cell division) locus is adjacent to the origin of replication of the F plasmid and increases the stability of F and other unrelated replicons

(*63, 73, 74*). Initial studies indicated that plasmid stabilization by *ccd* was caused by coupling of plasmid replication and cell division (*63, 75*). However, it has now been shown that *ccd* increases plasmid stability by inhibiting the growth of plasmid-free daughter cells (*59, 76, 77*). Chromosomal *ccd* loci are rare (Table 1.1) and are confined to a small group of Gram-negative (G-negative) bacteria (*72*).

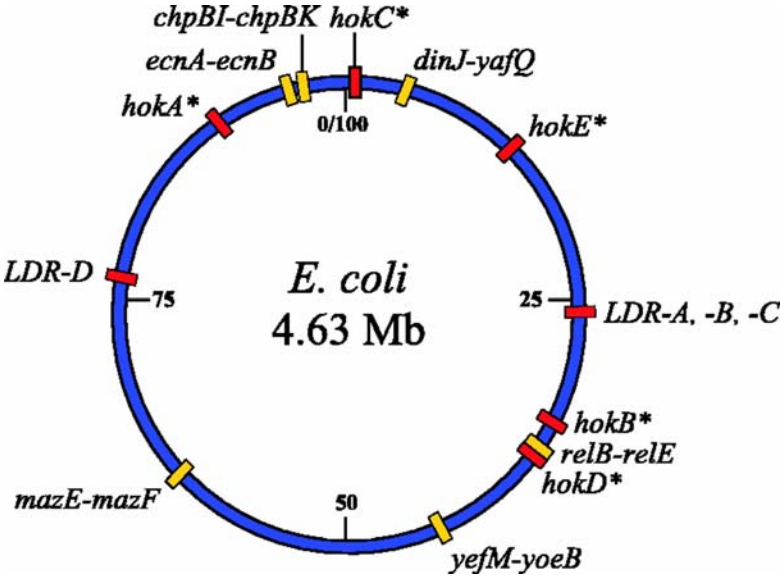

Figure 1.6: Location of known toxin -antitoxin modules on the *E. coli* genome. Asterisks denote genes that are inactive or relics. [type I (red) and type II (yellow)]

2. The *relBE* loci of *Escherichia coli*

The chromosomal *relBE* locus encodes the RelE toxin and the RelB antitoxin, and they form a non-toxic complex (*55, 56*). Overexpression of RelE inhibits cell growth, reduces the number of colony-forming units and inhibits translation (*56, 57, 78*). Initially, Gerdes and colleagues hypothesized that overproduction of RelE killed cells, owing to the dramatic drop in viable cell counts (*56*). However, ectopic overexpression of RelE has now been shown to induce a bacteriostatic condition from which the cells can be resuscitated (*78*). Some plasmids also harbour *relBE* homologues, which increase their maintenance (*79, 80*). Inquisitively, the chromosomal *relBE* locus stabilizes

plasmids as efficiently as the plasmid-borne *relBE* loci, raising the possibility that plasmid stabilization is a secondary phenotype that results from the intrinsic properties of the components encoded by *relBE* loci (*56*). The positions of the *relBE* homologous loci and other TA loci on the *E. coli* K-12 chromosome are shown in Figure 1.6.

3. The *higBA* locus of the Rts1 plasmid

The *higBA* (host inhibition of growth) locus of plasmid Rts1 encodes HigB toxin and HigA antitoxin (*81*). The *higBA* locus stabilizes plasmid Rts1 by inhibiting the growth of plasmid-free cells. Compared with other TA loci, *higBA* is unusual because the toxin-encoding gene is located upstream of the antitoxin-encoding gene. Purified HigA and HigB form a stable complex in which the two proteins are present in an equimolar ratio (*82*). Homologues of *higBA* have been found on the chromosomes of many Gram-negative and Gram-positive bacteria. Interestingly, phylogenetic analyses showed that HigB toxins have weak (but significant) similarity with the RelE group (*72*), whereas HigA and RelB antitoxins belong to different families of DNA-binding proteins.

4. The *parDE* locus of the RK2 plasmid

The *parDE* locus of RK2, a broad host-range plasmid, is a typical TA locus that is located adjacent to the site-specific resolution system of the plasmid (*83, 84*). The *parE* gene encodes a toxin and *parD* encodes an antitoxin. *parDE* increases plasmid maintenance by inhibiting the growth of plasmid-free cells (*85, 86*). Plasmid stabilization by *parDE* is efficient but species-dependent (*73, 87, 88*). Homologues of *parDE* loci are present in Gram-negative and Gram-positive bacteria (Table 1.1). Interestingly, phylogenetic analyses showed that HigB toxins have weak (but significant) similarity with the RelE group (*72*).

5. The *pem (parD)* and *mazEF (chp)* loci

The identical *parD* and *pem* (plasmid emergency maintenance) loci of plasmids R1 and R100, respectively, stabilize their replicons by delaying the growth of plasmid-free cells (*73, 89-91*). The *parD* and *pem* loci encode toxins Kid (killing determinant)/PemK and antitoxins Kis (killing suppression)/PemI, and are organized similar to other TA loci. Two chromosomal homologues of *parD*/*pem* were found in *E. coli* and named *chpA* and *chpB* (chromosomal homologues of *pem*) (*92, 93*). The *chpA* locus is 80 nucleotides downstream of the *relA* gene and was named *mazEF* ('ma-ze' means 'what is it?' in Hebrew) (*94*). These loci are abundant in bacteria but rare in, or absent from Archaea. Kid/PemK toxins do not kill plasmid-free segregants, but instead delay the growth of these cells (*73*). This was the first experimental indication that the MazF family of toxins does not kill the cells.

6. The *phd/doc* locus of the P1 plasmid

Prophage P1 is inherited as a genetically stable, extrachromosomal plasmid. Part of this stability is due to the phd/doc locus of P1. The *doc* (death on curing) gene encodes a toxin, Doc, and *phd* (prevent host death) encodes an antitoxin, PhD, that neutralizes Doc (*95*). As with other TA loci, plasmid stabilization is caused by proteolytic degradation of the antitoxin (*96*). Purified Phd and Doc form a Phd_2Doc trimeric complex, which indicates that Phd inhibits Doc through direct protein−protein contact (*97*). Chromosomal homologues of the *phd*/*doc* family are found in bacteria and a few Archaea. Curiously, Phd is similar to RelB-3 of *E. coli* (YefM) (*68, 98, 99*), but Doc has no similarity to RelE-3 (YoeB) or any other known RelE homologue. The cellular target of Doc has not been identified, but indirect evidence indicates that Doc inhibits translation (*100*).

7. The *vapBC* loci

The *vapBC* (virulence associated protein) locus was identified on a *Salmonella dublin* virulence plasmid (*101*). Inactivation of *vapB* (at that time named *vagC*) inhibited growth on selected media and resulted in loss of plasmid-associated virulence. Stability of the virulence plasmid was impaired by mutations in *vapB*. Several archaeal species, including *Archaeoglobus fulgidus* and *Sulfolobus tokodaii*, have more than 20 *vapBC* loci. The cellular target(s) of VapC toxins is not yet known.

8. The ω-ε-ξ locus of plasmid *pSM19035*

The ω-ε-ξ locus of pSM19035 — a low-copy-number, broad-host-range plasmid from *Streptococcus pyogenes* — encodes three components (Figure 1.5). The ω-repressor autoregulates transcription of the ω-ε-ξ operon (*102*), ε encodes an antitoxin and ξ encodes a toxin (*42, 103*). ξ toxin is neutralized by ε through direct protein−protein contact (*69, 103*) and, as in the case of canonical TA loci, ε is degraded by proteases in vivo (*103*). Moreover, inhibition of translation and transcription induced a strong, ξ-dependent decrease in viable cell counts. These results indicate that the ω-ε-ξ locus encodes a system that kills or prevents the growth of plasmid-free cells (*69*). ε-ξ genes are also present in several Gram-positive bacteria.

Cellular targets of toxins

To understand the function of TA loci, it is crucial to understand the cellular targets of the toxins. Toxin targets are also of practical interest because they could be potential new drug targets in pathogenic bacteria, and might be useful for creating novel genetic tools. Targets known for TA systems are summarized in Table 1.1.

Table 1.1: Assorted TA systems and targets of their respective toxins are summarized below.

TA family (locus)	Toxin	Target of toxin	Antitoxin	Protease	Number of loci	Phyletic distribution
Ccd	CcdB	Replication through DNA gyrase	CcdA	Lon	5	G-negative bacteria
relBE	**RelE**	**Translation through mRNA cleavage.**	**RelB**	**Lon**	**156**	**G-negative and G-positive bacteria, Archaea**
parDE	ParE	Replication through DNA gyrase	ParD	Unknown	59	G-negative and G-positive bacteria
higBA	HigB	Unknown	HigA	Unknown	74	G-negative and G-positive bacteria
mazEF	MazF/PemK	Translation through mRNA cleavage.	MazE/PemI	ClpXP/Lon	67	G-negative and G-positive bacteria, Archaea
Phd/doc	Doc	Translation	Phd	ClpAP	25	G-negative and G-positive bacteria, Archaea
vapBC/vag	VapC	Unknown	VapB	Unknown	285	G-negative and G-positive bacteria, Archaea
ω-ε-ξ	ξ	Unknown	ε	Unknown	16?	G-positive bacteria

The aim of this thesis deals with RelBE TA system from bacteria (*E. coli*) and archaea (*Methanococcus jannaschi*) (*104*). RelE toxin from both bacteria and archaea is known for cleaving mRNA at the ribosomal A-site. Although RelE proteins from Bacteria and Archaea are clearly homologous, their sequence similarities are modest (RelE from *E. coli* and *M. jannaschii* homologue share 18% identical and 40% similar amino acids only) (*104*). Interestingly, in both organisms *E. coli* and *Methanococcus jannaschi*, ectopic expression of RelE inhibits cell growth, reduces the number of colony-forming units and severely

inhibits translation (*56, 57, 78, 104*). Importantly, the *relBE* locus reduces the global level of translation during amino acid starvation (*57*). Therefore, RelE is a global inhibitor of translation that is activated by nutritional stress. RelE cleaves mRNA positioned at the ribosomal A-site *in vitro* and *in vivo* (Figure 1.7.). Cleavage occurs between the second and third bases of A-site codons. RelE-mediated cleavage is catalytic *in vitro*, and cleavage efficiency is dependent on the actual codon present at the ribosomal A-site. UAG and UAA are cleaved 800- and 100-fold more efficiently than the UGA stop codon (expressed as relative K_{cat}/K_m values). Addition of release factor 1 (RF1) reduces RelE-mediated mRNA cleavage *in vitro* (*105*). As RF1 binds firmly to the ribosomal A-site (*106*), this observation indicates that RelE must have access to the A-site to induce mRNA cleavage (Figure 1.7.).

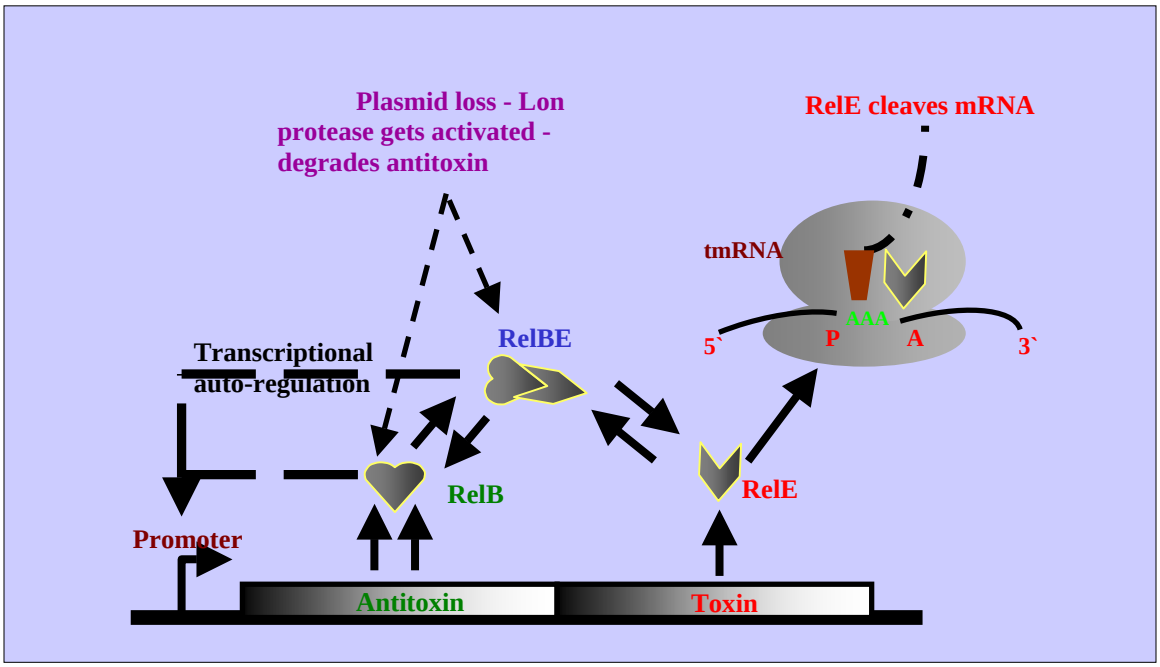

Figure 1.7: Genetic organization and components of the *E. coli relBE* operon. The *relB* promoter is autoregulated by RelB and the RelBE complex. The large and small ribosomal subunits are shown as grey spheres. The A- and P- sites with residing codons are indicated within 30s subunit. RelE-mediated cleavage between the second and third bases of the A-codon is also indicated.

In vivo, RelE cleaves translated RNAs only in their coding regions, consistent with the observation that, in vitro, RelE cleavage depends on the presence of ribosomes (*104, 105*). Interestingly, RelE homologues encoded by the archeaon *Methanococcus jannaschii* cleave test mRNAs *in vitro* with a pattern similar to that of *E. coli* RelE (*104*). RelE-3 of *E. coli* (YoeB) also cleaves mRNA in a pattern similar to that of RelE, and might therefore be considered a RelE homologue (*107*).

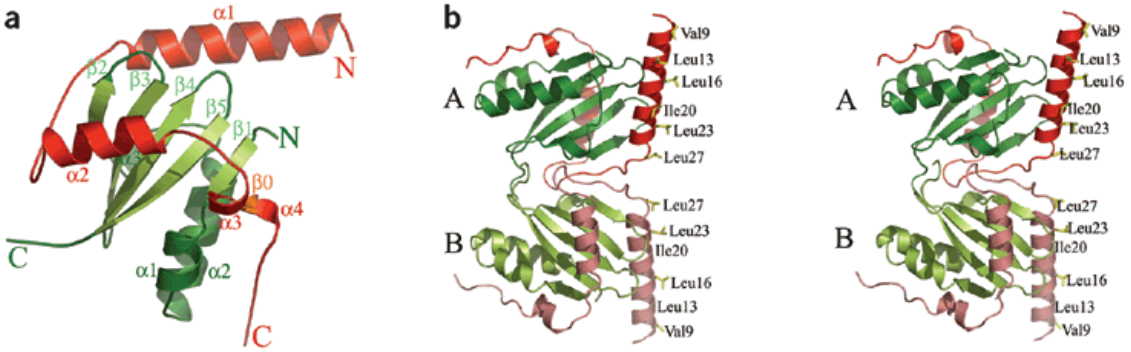

Figure 1.8: Structures of archaeal *P. horikoshii* RelBE complex (108). (a) Ribbon representation of the RelBE complex from archaeal *P. horikoshii*. RelB and RelE are represented as red and green, respectively. N and C termini of both molecules are indicated. (b) Stereo view of the heterotetrameric structure of RelBE complex from *P. horikoshii*. Two molecules (A and B) of the RelBE complex observed in the asymmetric unit are indicated. Side chains of hydrophobic residues at α1 in RelB are indicated.

Recently, the first crystal structure of archaeal RelBE complex (first from RelBE family members) from hyperthermophilic archaeon *Pyrococcus horikoshii* RelBE toxin -antitoxin has been published (Figure 1.8) (*108*). Bacterial *relBE* systems are conserved in archaea such as in *M. jannaschii*, *Archaeoglobus fulgius*, and *P. horikoshii* OT3. As mentioned above, archaeal RelE from *M. jannaschii* cleaves mRNAs on translating ribosomes in a manner similar to the *E. coli* RelE toxin.

Notably, the structure of RelBE is distinct from that of the previously determined MazEF (Figure 1.9) (*109*), interesting regulatory system, genes of

which are present on the *E. coli* chromosome. By crystal structure data for archaeal RelBE, RelE folds into an α/β structure, a single globular domain with an unusual fold, whereas archaeal RelB lacks any distinct hydrophobic core and extensively wraps around the molecular surface of RelE (Figure 1.8) (*108*). Most of the polypeptide chain of RelB has contacts with surface residues of RelE and interactions are predominantly electrostatic. This suggests that the antitoxin has a defined structure only when bound to the toxin. In addition, the lack of tertiary structure could make the unbound RelB easy prey for the ATP-dependent Lon protease, thus explaining the short half-life of the antitoxin (*57*). Similarly, the antitoxin MazE also has a long unstructured extension that wraps around a dimer of the MazF toxin (Figure 1.9) (*109*). Studies revealed a binding mode on RelE which is extensively wrapped by RelB (*105, 108*). RelE binds ribosomes and RelB inhibits the ribosome-binding activity of RelE. It is thus unlikely that RelB simply masks the ribosome-binding site of RelE, thereby preventing RelE from access to the ribosome A-site. Rather suggestions are that the extensive wrapping around RelE enlarges the size of the molecule, thereby precluding it from penetration into the ribosome A-site. As mentioned for TA systems, RelB and the RelBE complex autoregulate the transcription of *relBE* via direct binding to a palindromic nucleotide sequence on the *relBE* promoter region. Because many gene regulatory proteins recognize DNA as homodimers, dimerization of RelBE complex looks obvious. Crystallographic data and gel filtration analysis suggest a heterotetrameric $(RelBE)_2$ complex (*108*). However, the presence of several exposed hydrophobic residues located in the N-terminal helix of RelB suggest that these residues may constitute the interaction site for the two RelBE dimers in solution and that the structure of the biological heterotetramer may be different from the observed in crystal (Figure 1.9) (*108*). In comparison, the MazEF complex forms a heterohexamer ($MazF_2$-$MazE_2$-$MazF_2$), half of which is shown in Figure 1.9., in which MazF monomers are structurally related to the unbound CcdB and kid toxins, despite low sequence homology (16-25%) (*109*).

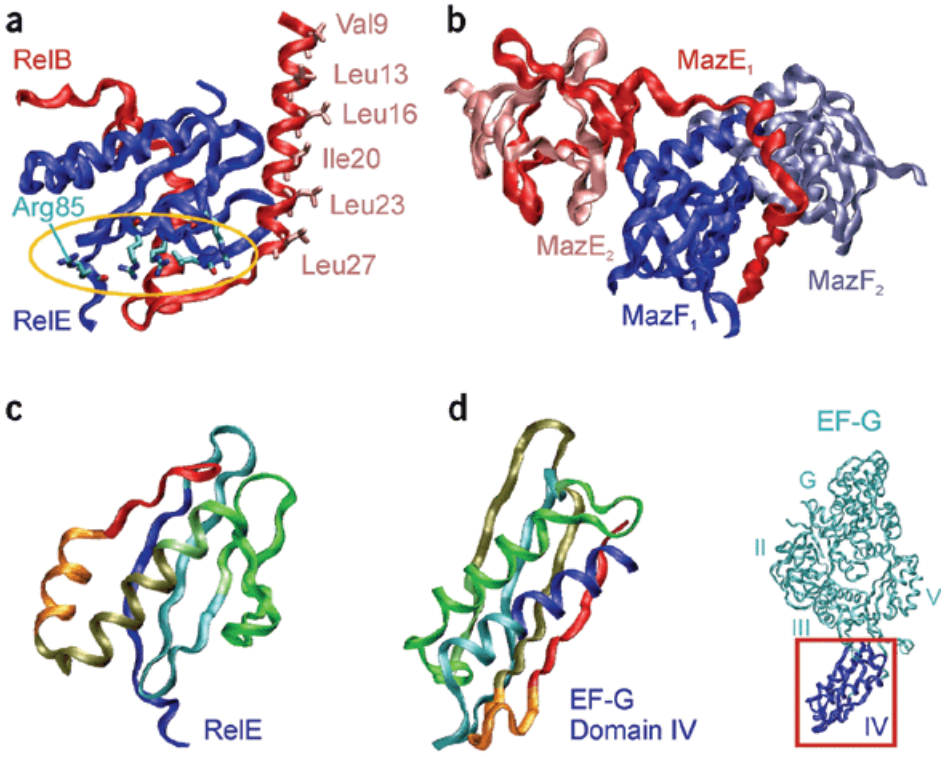

Figure 1.9: Comparison of RelBE from *pyrococcus horikoshii* with MazEF and EF-G (110). (**a**) Crystal structure of the RelE (blue)-RelB (red) dimer (*108*). The charged residues defining the putative active site of RelE (circled) as well as the hydrophobic residues in RelB (labeled) potentially involved in RelBE dimerization are indicated. (**b**) Crystal structure of $MazE_1$-$MazF_1$-$MazF_2$ (red, blue, violet) as one half of the heterohexamer. The region of $MazE_2$ (pink) that forms globular domain of the MazE dimer is also included. (**c,d**) Structures of RelE (c) and domain IV of elongation factor G (EF-G; **d**), illustrate their similarity. The region of EF-G in **d** is boxed in red and is shown in the same orientation as the isolated domain IV.

Since crystal structure of archaeal RelBE is known now, the interplay between RelB, RelE and translational apparatus can be represented in more defined schematic way considering hypothesis and the stoichiometry of components (Figure 1.10).

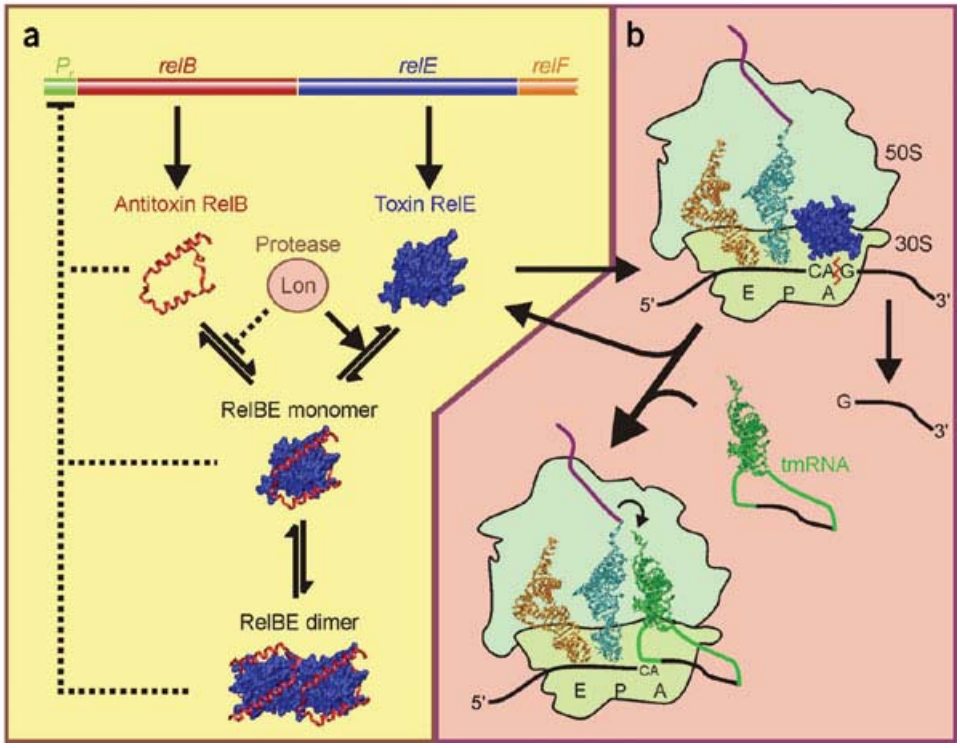

Figure 1.10: The interplay between RelB, RelE and the translational apparatus (110). (a) The antitoxin RelB (red) and toxin RelE (blue) are encoded consecutively in the *relBEF* operon. RelB interacts with RelE to form inactive RelBE monomer and dimer. Free RelB as well as the bound forms bind to the promoter (P_r) of the *relBEF* operon to feedback-inhibit transcription. The action of the Lon protease to degrade both free and bound forms of RelB results in the presence of free RelE, which can inhibit protein synthesis. (**b**) RelE binds to stalled ribosomes and cleaves at the second position of the A-site codon (in this case, CAG) of the mRNA. Blocked ribosomes with truncated mRNAs in the A-site are the substrate for the tmRNA rescue system. The tmRNA molecule binds to the A-site, initially without codon-anticodon interaction, and the tRNA-like part of the molecule (green) accepts the nascent chain (purple) from the P-site tRNA (cyan). Translation then proceeds of the mRNA-like part (black) of the tmRNA, thus tagging the nascent chain for degradation.

Although an extensive study has been done on this regulatory systems some elementary questions are still looking for answers; in terms how single-celled organisms could benefit from suicide. How a small protein can regulate the capacity of the entire protein-synthesizing apparatus by cleavage of ribosome-associated mRNAs, a highly surprising and dramatic mechanism. How a supposedly unstable RelB protein level is regulated by Lon protease. More

interestingly, there are the reports that antitoxin yefM and phd are natively unfolded protein when not in the complex. Is RelB antitoxin also follows similar trend? The major conclusion that can be derived from the first crystallographic data published for one member out of RelBE family is that RelB which does not have its hydrophobic core possibly could be a natively unfolded protein in case not bounded with RelE. If this is the case, then the RelB protein and RelBE TA system are very interesting system for biophysical studies, and definitely studies can provide more insights in the solution behavior of TA systems which can be tried to correlate with the mechanism of system. To gain more insights, we started studying protein folding pathways and the stability of individual components in the system. RelBE complex proteins from *E. coli* and *M. jannaschii* were studied with a belief that energetics could provide a better understanding of system.

Protein folding

Why do we study protein folding? Understanding how proteins fold could be the key to understanding life. Proteins are involved in just about every aspect of the maintenance of cells and are the targets of many drugs. In biophysics, protein field has been center of focus for more than 30 years. The understanding of how a protein folds became more urgent than ever. In addition, there are about 20 known protein-misfolding diseases including, Alzheimer's, Huntington's, scurvy, scrapie etc. With the completion of the DNA sequence of many genomes, the primary structures of many unknown proteins have appeared. The proteomics field has arisen to understand their function if their native structures can be predicted from the amino acid sequence. Although there aren't any existing methods that can reliably predict the 3-D structure of a protein from its sequence, considerable progresses have been made in the understanding of the mechanisms of protein folding. Such advances were achieved by means of experimental and theoretical studies of proteins and simple generic lattice and off-lattice models (*111*). Therefore, physical understandings of how a protein folds or misfolds will help us to develop drugs that recognize target proteins or fix proteins that have misfolded. For example, antibodies, which are proteins indeed, have for many years been seen as useful therapeutics for a number of human diseases ranging from rheumatoid arthritis to leukemia because they are designed to target particular cells and attract other parts of the immune system to the site. There are a dozen antibodies that are approved as therapeutics by the U.S. Food and Drug Administration (http://www.fda.gov/), and many more under development. Although the common person may not know what proteins are, interestingly they affect every aspect of their life.

The origin of the question

Ever since Anfinsen's laboratory demonstrated that denatured bovine pancreatic ribonuclease was able to refold, unassisted by catalysts or cofactors,

to its fully active native state (*112*), the protein folding field grew and involved scientists from many fields. Both experimentalists and theoreticians were trying to answer the question of "how does the amino acid sequence of a protein specify its three-dimensional structure?" (*113*). Anfinsen's explanation for the protein folding problem was that the native protein in its normal physiological milieu was the one in which the Gibbs free energy of the whole system is lowest (*114-116*). Anfinsen approached the protein folding problem from the thermodynamic perspective. However, in the late 1960's, Cyrus Levinthal questioned this approach; by subsequently performing a simple calculation to determine how long it might take for a protein to fold into its native structure. In a standard illustration of *Levinthal's paradox*, if we limit each connection between amino acid residues to three possible states, then a polypeptide chain of 101 amino acids could exist in $3^{100}= 5 * 10^{47}$ configurations. Even if protein were capable of sampling 10^{13} configurations per second, it would take $1.6 * 10^{27}$ years to try them all. Therefore, Levinthal concluded that proteins must fold by specific pathways (*117, 118*). This introduced the hypothesis of "kinetic control" and lead to a search for folding pathways, there are several theories on the role and nature of protein folding pathways, which can be separated into two groups: the hierarchical model which states that performed secondary structure fold into tertiary structures (*119, 120*), and the hydrophobic collapse model in which parts of the protein are brought (or nucleate) to form the beginnings of tertiary structure and then secondary structures form (*121, 122*).

The hierarchical model has intellectual appeal, because it reduces the folding problem to simple understanding of the individual secondary structures and how they assemble. A good metaphor of this folding model is "prefabricated construction", where all the individual parts (α helices, β strands) are fabricated first and then assembled into the final tertiary structure. This model gained support when it was first demonstrated that relatively small peptides could form stable secondary structures. Previously most peptides removed from proteins were not stable enough to form secondary structural elements without tertiary

contacts. Often, when secondary structural elements are isolated they tend to adopt helical structures, turns and less frequently β structures like β hairpins (*119*). These structures might represent the starting point for folding. Unfortunately, the largest opponent is the most peptides removed from proteins do not form structure.

Support for the hydrophobic collapse model came from early studies showing that the energetics of removal of water from hydrophobic group to be considerable (*123*). However, the idea that a conformational search is facilitated within a nonspecific hydrophobic globule presents a problem. The excess of interactions hinders reorganization of the protein. A redefined model suggests that the secondary structural elements formed during the collapse, thus limiting the need for large structural rearrangements.

In the past few years a “new view” of protein folding has emerged (*117*). The “classical view” of protein folding suggested that proteins fold along a pathway with specific populated intermediates. The classical view looked at protein folding occurring along a reaction coordinate where the rate-limiting step was determined by the free energy of a single transition state. The new view holds that the “transition state” is actually an ensemble of many structures and there are multiple folding routes. Whether any or all of these models is correct is uncertain.

One of the first things I learned at university as biochemistry student is that protein structure = function. This addresses the idea that for a protein to work, it must fold into a structure. To understand this paradigm, we must have a clear comprehension of how a protein folds and what forces stabilize the folded structure. While much is known about primary structure and the final native state of folded proteins, the folding pathway itself is still not well understood. Keeping these things as objectives in mind, my dissertation consists of two distinctly different projects. The first was to perform biophysical and an equilibrium folding

studies for the toxin -antitoxin system RelBE complex and its isolated individual components RelB Antitoxin and RelE Toxin from *Escherichia coli* and thermophilic bacteria *Methanococcus jannaschii* whereas; second project investigates folding pathways and other biophysical properties of DNA-binding protein β recombinase from *Streptococcus pyogenes*.

BACKGROUND

Measuring the conformational stability of a protein requires the denaturation of protein from folded (native) to unfolded state (denatured) (Figure 1.11); which further leads to the determination of the equilibrium constant (*Keq*) and ultimately the free energy of this reaction:

$$\text{Native(N)} \xleftrightarrow{Keq} \text{Unfolded(U)} \qquad \text{Equation 1.1}$$

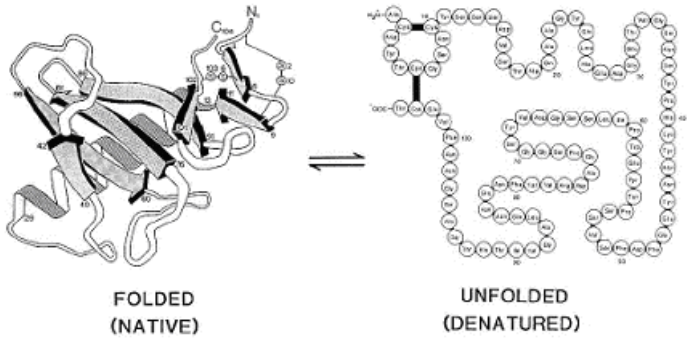

Figure 1.11: Pictorial representation of folded and unfolded protein.

We will explore many methods here and in materials and methods section that have been devised to measure conformational stability including solvent denaturations, thermal denaturations, circular dichroism, fluorescence spectroscopy, analytical ultracentrifugation and size exclusion chromatography.

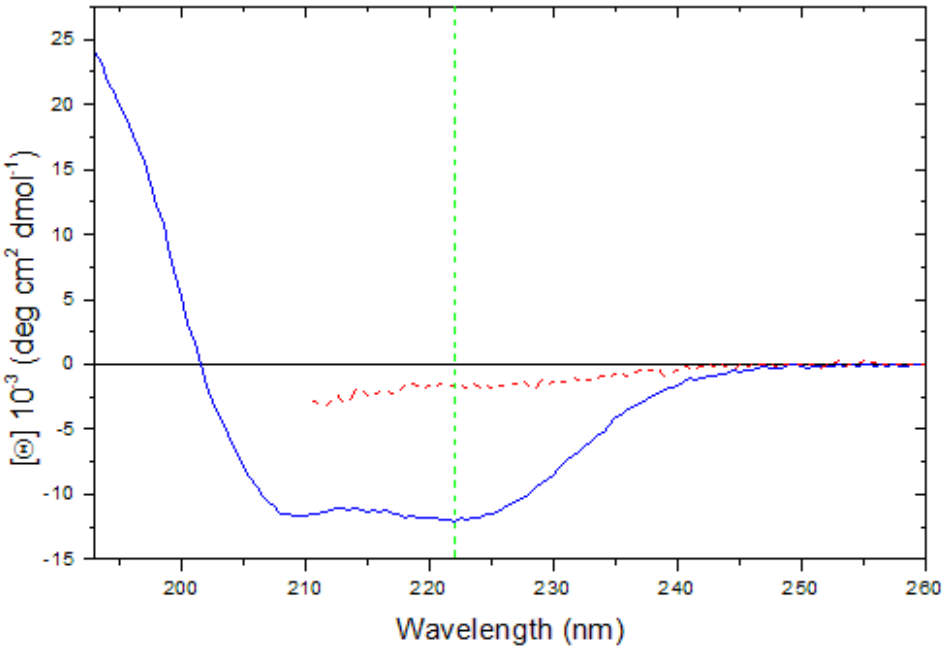

Figure 1.12: Representative circular dichroism spectra of a folded and unfolded protein. The solid line represents the far-UV CD spectrum of a folded protein. The dashed line represents the unfolded spectra of the same protein. The vertical dashed line represents the wavelength of greatest separation *i.e.* 222 nm in the presented case.

Solvent denaturations

For a solvent denaturation, we monitor a spectroscopic signal that represents the protein as a function of denaturant. Urea and guanidine hydrochloride (Gdn-HCl) are the most commonly used denaturants for unfolding studies. The spectral probes most often used to monitor unfolding are tryptophan fluorescence (*124, 125*), and far-UV circular dichroism (CD) (*126*). Tryptophan fluorescence reports on the specific environment of the tryptophan (Trp) residue making it an excellent probe for local environmental fluctuations during unfolding, whereas far-UV CD is a reporter of secondary structure. In order to define the native and unfolded states, the spectra of the native and unfolded protein must be determined. The native state is the spectrum of the protein in the absence of

denaturant, whereas the unfolded state is the spectrum in the highest concentration of denaturant (10 M urea or 8 M Gdn-HCl). A single wavelength that shows a significant difference between native and unfolded states is chosen to monitor the unfolding reaction (Figure 1.12).

Tryptophan fluorescence

Tryptophan fluorescence of proteins has been used for many years to gain insight into a protein's structure and dynamics. In 2001, Vivian and Callis stated that some 300 papers per year abstracted in *Biological Abstracts* reported work that exploits or studies tryptophan fluorescence in proteins (*125*). Figure 1.13 shows typical fluorescence spectra for protein containing one tryptophan in the presence of 0 M and 8 M urea. Among the properties studied are changes in fluorescence intensity, wavelength of maximal intensity (λ_{max}), band shape, anisotropy, lifetimes (τ), and energy transfer. The studies are applied to protein folding, substrate binding, external quencher accessibility, etc. (*125*). However, understanding tryptophan fluorescence is a complicated subject. Many papers in the past several years have demonstrated that the fluorescence of tryptophan residues is strongly dependent on their environment (*127-131*).

Tryptophan is the most important of the intrinsic fluorescence probes in proteins; it has a larger molar absorption coefficient (Table 1.2), it serves as an energy transfer acceptor for the other aromatic amino acids, it can be selectively excited at long wavelengths (e.g., $>$295 nm), and its fluorescence intensity (I_F) and the intensity wavelength maximum (λ_{max}) are sensitive to the microenvironment of the indole side chain (*132-135*). Table 1.3 gives the λ_{max} values of tryptophan, NATA, and a tripeptide, N-acetyl-AWA-amide, in various solvents.

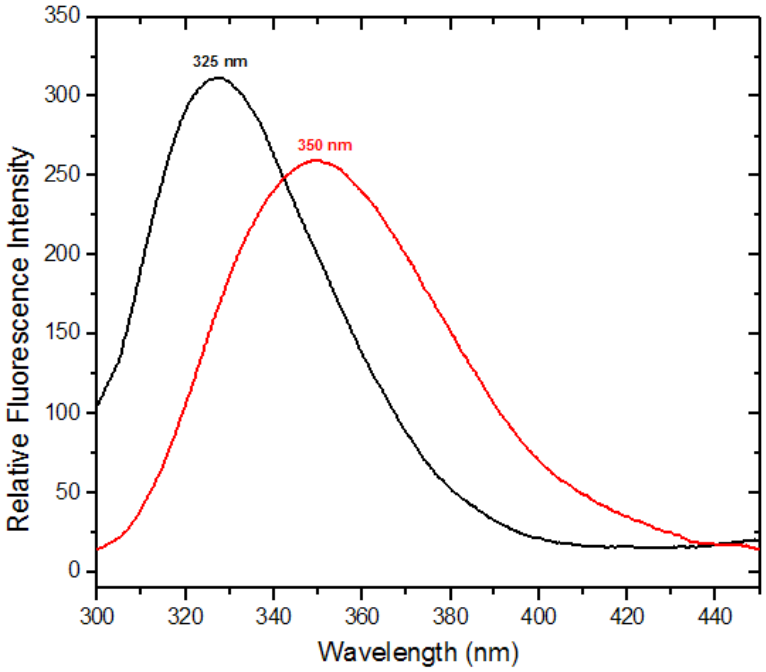

Figure 1.13: Representative urea denaturation fluorescence spectra of protein. The spectra in black and red corresponds to 0 M and 8 M urea in physiological buffer conditions, respectively. Excitation wavelength was 295 nm specifically for excitation of tryptophan.

Table 1.2: Aromatic amino acids, their residue volumes, mean percent buried in proteins, hydrophobicity, absorbance and fluorescence properties.

Amino Acid	Volume (A^3)[a]	Mean Percent Buried[b]	Hydrophobicity (Kcal mol^{-1})[c]	Absorbance[d]		Fluorescence[e]	
				λ_{max}[f]	ε_{max}[g]	λ_{max}[f]	φ_F
Tryptophan	232	87	3.1	280	5600	353	0.13
Tyrosine	197	77	1.3	275	1400	304	0.14
Phenylalanine	194	88	2.4	258	200	282	0.02

[a](*136*)
[b]The mean fraction buried in a set of 61 proteins (*137*)
[c](*138*)
[d](*139*)
[e](*135*)
[f]nm
[g]M^{-1} cm^{-1}

Table 1.3: λ_{max} values for tryptophan and two tryptophan models in various solvents.

Compound	λ_{max} (nm) in various solvents[a]					
	Hexane	Dioxane	Ethanol	Acetonitrile	Water	9 M Urea
NATA[b]	320	328	337	335	351	348
AWA[b]	322	337	337	335	351	349
NATA[c]		329	340	334	352	
Trp[d]		329	338		350[e]	353

[a]The dielectric constants for the solvents are: hexane, 1.9; dioxane, 2.2; ethanol, 24; acetonitrile, 38; Water, 78; 9 M Urea, 99.
[b]NATA is N-acetyl-Trp-amide; AWA is N-acetyl-Ala-Trp-Ala-amide.
[c](*140*)
[d](*141*)
[e]In 0.1 M Tris, pH 7.0.

To help in the understanding of the fluorescence properties of tryptophans, attempts have been made to classify tryptophans. One of the first to suggest discrete classes of tryptophans was Konev in 1967, who suggested two main classes, involving the parameters λ_{max} and quantum yield (q). One of the classes included tryptophans inside the protein in a non-polar, hydrophobic environment with a shorter wavelength of the fluorescent maximum (λ_{max} of ~330nm) and a rather low quantum yield (0.04 to 0.07). The second class consisted of exposed tryptophan residues in a polar aqueous environment with a long wavelength fluorescence maximum (λ_{max} ~350 nm) and a quantum yield equal or higher that that of free aqueous tryptophan (~0.13 to 0.17). This idea was based on the observation that protein spectra shift toward 350 nm upon denaturation by urea, and toward 330 nm upon addition of anionic detergents (*142*). In 2001, Callis stated that the most common use of tryptophan fluorescence λ_{max} information is to assign a tryptophan as buried in a non-polar environment if λ_{max} is ~325 nm or as expressed in a polar environment if λ_{max} is in longer wavelength like ~350 nm (*125*). However, classifying tryptophans into one of these two classes is surely too restrictive. For example, Konev's hypothesis could not explain proteins with high quantum yield and an intermediate λ_{max}. For

example, Ribonuclease T1 has a quantum yield of 0.31 and a λ_{max} of 322 nm (*133*). Furthermore, Callis demonstrated that mere exposure to hydrophilic environment may or may not create a large red shift (*125*).

In 1973 and later refined in 2001, Burstein and co-workers revised and extended Konev's hypothesis of discrete classes additional spectral parameters and approaches (*142-145*). A detailed characterization of the environment of the tryptophan in each class was made in terms of hydrogen bonding, solvent accessibility, packing density, relative polarity, temperature factor, and dynamic accessibility. Burstein assigns tryptophans to one of five discrete classes(*142, 145*). This division into classes should allow researchers to better compare tryptophans in various proteins. As Engelborghs states, "This work represents an important achievement in the characterization of the environment of each tryptophan and the linkage to the spectral properties" (*146*).

Far-UV circular dichroism

Figure 1.14 shows a typical equilibrium urea denaturation using CD at 222 nm to monitor the unfolding reaction. The curve is divided into three regions: the pre-transition region, the transition region and the post-transition region. The pre- and post-transition regions show how the denaturant affects the folded and unfolded protein and the transition region shows how denaturant effects the change in the concentration of the native state with respect to the unfolded state. As with any thermodynamic measurement, it is essential that equilibrium is reached and the reaction is reversible. In the similar way, emission maxima obtained from fluorescence spectra of protein at different denaturant concentration can also be plotted and segregated in the pre-, post-transition and transition regions.

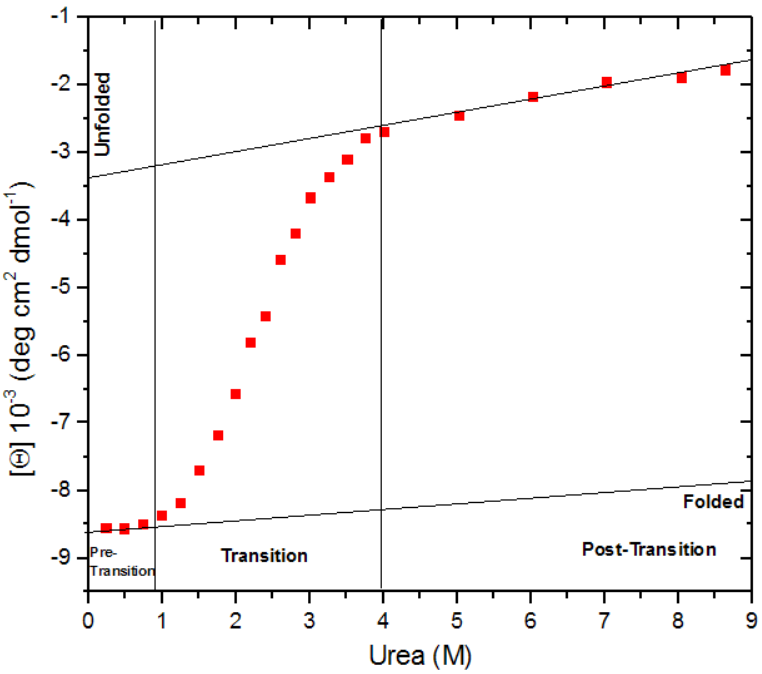

Figure 1.14: Representative urea denaturation curve. The unfolding reaction was monitored by far-UV CD at 222 nm. The three major regions of a denaturant curve are segregated by vertical lines and labeled pre-transition, transition, and post-transition. These regions were chosen arbitrarily and have no mathematical significance. The horizontal lines represent the signals of folded and unfolded protein that will be used to convert the data into fraction unfolded.

Many small proteins have been shown to unfold in a two-state mechanism (*147-151*). We will assume a two-state unfolding mechanism for the discussion here; however, three-state mechanisms will be discussed later as a part of results and appendix. Assuming a two-state mechanism means that for the points on Figure 1.14 only folded and unfolded protein molecules are present ($f_F+f_U=1$), where f_F and f_U are the fraction of protein present in the folded and unfolded state. Thus the observed value of y at any point can be defined as $y=y_Ff_F+y_Uf_U$, where y is a physical parameter to follow unfolding; y_F and y_U represent the spectroscopic values characteristic of the folded and unfolded states, respectively, under the conditions where y is being measured. Combining these equations gives:

$f_U=(y_F-y)/(y_F-y_U)$ Equation 1.2

Using the data from Figure 1.14 and equation 1.2, the data were expressed as fraction of unfolded protein in Figure 1.15. The equilibrium constant K_{eq}, and the free energy ΔG can be calculated using equation (1.3 and 1.4).

$K_{eq}=f_U/f_N=f_U/(1-fU)=(y_F-y)/(y_F-y_U)$ Equation 1.3

$\Delta G=-RT\ln[K_{eq}] = -RT\ln[(y_F-y)/(y_F-y_U)]$ Equation 1.4

where R is the gas constant (1.987 cal/mol K) and T is the absolute temperature in Kelvin. Values of y_F and y_U are obtained by extrapolating the pre- and post-transition baselines to 0 M denaturant (Figure 1.14).

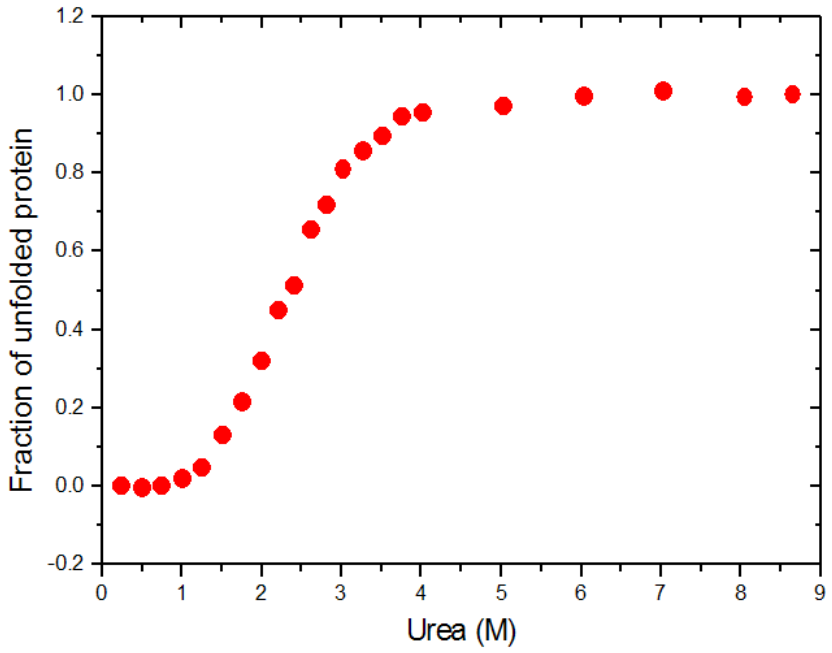

Figure 1.15: Fraction of unfolded protein as a function of urea denaturant. This graph was constructed from the data shown in Figure 1.14 using equation 1.2. The points represent the fraction of unfolded protein present at each urea concentration.

While being able to calculate ΔG in transition region (1 to 4 M urea) is helpful, what we really want is to know the stability in water. Pace and Greene (*152*) noticed that ΔG varies linearly with the molar concentration of denaturant within the transition region (Figure 1.15). Using what has become known as the linear extrapolation method (LEM), they extrapolated the stability to 0 M denaturant (Figure 1.16) (*152*). We utilize equation 1.5 to determine ΔG:

$$\Delta G=\Delta G_{water}-m\cdot[D] \qquad \text{Equation 1.5}$$

where *m* is a measure of the linear dependence of ΔG on denaturant concentration [D].

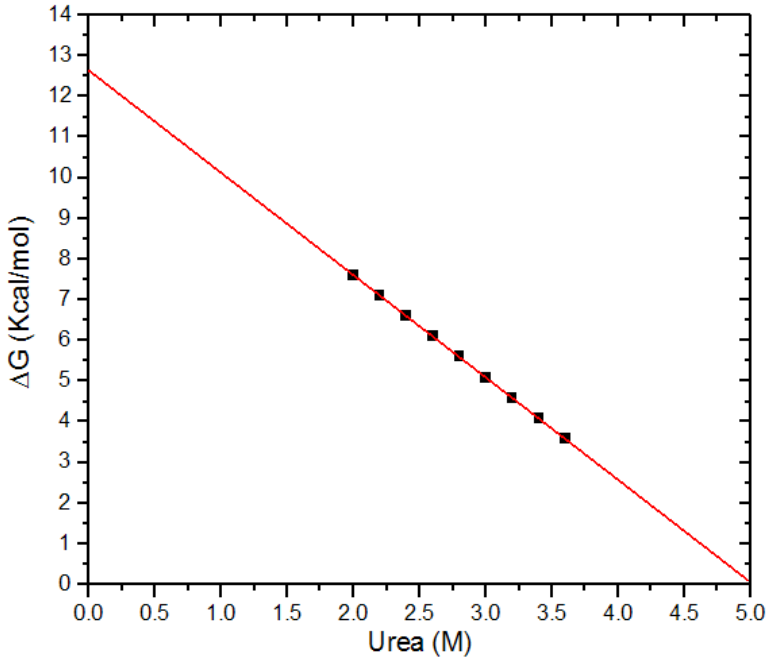

Figure 1.16: The linear extrapolation method. This Figure shows the change in the Gibbs free energy as a function of urea. This graph illustrates the linear dependence of ΔG in the transition region. The solid line represents the linear fit by equation 1.5. The stability of this protein is about 12.6 kcal/mol with a C_{mid} of 2.5 M and an *m* value of 3 kcal/mol M.

Typically the linear extrapolation method (LEM) is implemented through a more precise analysis as described by Santoro and Bolen (*153*). They describe a method by which the transition region is characterized by two parameters *m*, the dependence of ΔG on denaturant and [C½ or C_{mid}] the midpoint of the transition where ΔG=0. The pre- and post-transition regions are characterized by two parameters θ_N and θ_U, the intercepts of the unfolded and native baselines and a_N and a_U which measures the denaturant dependence of the pre- and post-transition baselines. The following equation represents the curve shown in Figure 1.15:

$$[\theta_{obs}] = \frac{([\theta_N] + a_N[D]) + ([\theta_U] + a_U[D]) \times \exp[m \times ([D] - C_{mid})/RT]}{1 + \exp[m \times ([D] - C_{mid})/RT]} \qquad \text{Equation 1.6}$$

The best fit of the six parameters is performed using a least-squares fit program.

Thermal denaturations

Thermal denaturations are the most commonly monitored by CD or by the use of differential scanning calorimetry (DSC). We will first discuss the spectroscopic method of thermal denaturation (van't Hoff analysis). A useful measurement of ΔG requires extrapolating measurements from a narrow temperature range, where unfolding occurs, to a reference temperature such as 20 °C. To obtain the enthalpy change (ΔH), the van't Hoff equation is used:

$$d(\ln K_{eq})/d(1/T) = -\Delta H/R \qquad \text{Equation 1.7}$$

The Van't Hoff plots (ln K_{obs} vs. 1/T) for protein unfolding transitions are usually non-linear, provided that the transition covers a wide temperature range. This indicates that ΔH varies with temperature, which is expected when the heat capacities of the native and unfolded protein differ.

$d(\Delta H)/d(T) = C_p(U) - C_p(N) = \Delta C_p$ Equation 1.8

In equation 1.8, $C_p(U)$ and $C_p(N)$ are the heat capacities of the unfolded and native conformations, and ΔC_p is the change in the heat capacity upon unfolding. With this in mind, ΔC_p and ΔH are both required to calculate ΔG as a function of temperature. Since ΔH is needed at only one temperature, the best temperature to use is T_m, the midpoint of the thermal unfolding curve where $\Delta G(T_m) = 0 = \Delta H_m - T_m \cdot \Delta S_m$. Now with these parameters and the modified Gibbs-Helmholtz equation 1.9 we can calculate $\Delta G(T)$.

$$\Delta G(T) = \Delta H_m(1-T/T_m) - \Delta C_p[(T_m-T)+\ln(T/T_m)] \quad \text{Equation 1.9}$$

We need T_m, and ΔC_p to calculate $\Delta G(T)$. The simplest method to determine ΔH is a plot of ΔG *versus* Temperature. From this plot we get T_m and ΔH_m where T_m is the temperature at which ΔG is 0 and ΔH_m is the enthalpy at the T_m. Now all we need is to determine ΔC_p.

While there are many methods to determine ΔC_p, a useful technique to determine ΔC_p was described by Pace and Laurents (*154*). In this method, ΔG calculated from the transition region performed at different denaturation temperatures are combined with ΔG values from the transition region of a thermal denaturation unfolding curve. A least squares fit to equation 1.9 yields ΔH, ΔC_p and ΔT_m.

Differential scanning calorimetry

The differential scanning calorimeter measures the heat absorption of a sample as a function of temperature (*155, 156*). Pair of cells is placed in a thermostated chamber. The sample cell is filled with a protein solution and the

reference cell is fitted with an identical volume of solvent. The two cells are heated with a constant power input to their heaters during a scan. Any temperature difference between the two cells is monitored with a feedback system so as to increase (or decrease) the input power to the sample cell. Since the masses and volumes of the two cells are matched, the power added or removed by the cell feedback system is a direct measure of the sample and reference solutions. The cell feedback power represents the raw data, expressed in units of cal/min. By knowing the scan rate and the sample concentration, these units are converted to cal/mol-Kelvin.

In practice, the sample and reference cells can be slightly mismatched. The usual practice is to record a reference baseline for the experimental scan; this is subtracted from the experimental data to yield C_p vs. T. The peak maximum occurs near T_m. The heat capacity (C_p) is the temperature derivative of the enthalpy function:

$C_p = (dH/dT)_p$ Equation 1.10

The enthalpy is obtained from a DSC experiment by integration of the heat capacity curve between two temperatures (initial and final):

$\Delta H_{cal} = \int C_p dT$ Equation 1.11

Multiple unfolding transitions

When the unfolding reaction shows more than one transition, unfolding is more complex than a two-state reaction. This behavior is frequently observed for multi-domain proteins. Observing a single transition unfolding does not prove a two-state mechanism; it merely suggests that there are at least two-states. Insight into the folding mechanism can be gained by utilizing different techniques and probes to follow the unfolding transition. Non-coincidence of plots of fraction

of unfolded as a function of temperature or denaturant concentration determined by different spectral probes indicates that an intermediate is present and hence a two-state mechanism cannot be used in analysis of the data. However, coincidence of the unfolding data is only a support and again does not prove a two-state mechanism.

The best support for two-state thermal unfolding is to show that ΔH_{vH} determined by the van't Hoff relationship is identical to that determined by calorimetry ΔH_{cal}. When $\Delta H_{cal} > \Delta H_{vH}$ it is clear evidence that significant concentrations of intermediates are present at equilibrium. If intermediates occur with similar T_m values, the separate equilibria will result in a broadening of the transition curve; a lower slope on the van't Hoff plot and an underestimate of ΔH for the process, *i.e* $\Delta H_{cal} > \Delta H_{vH}$. In some cases, two separate peaks are seen in the C_p vs. T plot. If the transitions correspond to independent folding of two protein domains, they can often be studied separately.