MOLECULAR GENETICS

A Tail of Protein Folding

ROBERT J. VILLAFAÑE, PhD, KRISHNA BAKSI, PhD*

ABSTRACT. This review describes the use of a simple genetic system that has provided important insight into the process of folding and, of its flipside, that of protein aggregation. These studies make use of the tail protein of the bacterial virus P22 which infects *Salmonella typhimurium*. This folding system serves as a model for a number protein structural elements and may also provide important insights into disease-related protein folding defects at a time when an increasing number of diseases are being shown to be due to protein folding alterations. *Key words: Tailspike protein, Protein folding, P22 bacterial virus, Intermediate.*

The study of protein structure, function and folding is critical to basic and clinical sciences alike. Proteins are used in a variety of commonplace situations from cleaning laundry and contact lenses to being used as drugs. In all cases, the ability of a protein to fold is an essential process. The inability to fold is lethal to the structure and function of a protein (1-7). A protein without its normal characteristic 3D structure can not function. The problem thus becomes how to predict (determine) the 3D structure of a protein from its amino acid sequence.

Another phrasing of the same protein folding question is: given the amino acid sequence of a polypeptide chain (one dimensional structure of a protein molecule) can its three dimensional structure be predicted? This basic question has been expanded to include a large number of closely related studies such as: Does the folding path of particular proteins have folding intermediates structures? What interactions occur during the protein folding pathway?

The problem of how a protein folds is immediately evident when one considers that there are over nine thousand proteins whose 3D structures have been precisely determined by either x-ray diffraction or nuclear magnetic resonance methods. Most of the structures have been determined by x-ray methods and in these structures, the exact location of every atom is precisely known (with the exception of hydrogens in diffraction studies). Despite this fact, it is still not known how that three-dimensional structure is achieved.

The study of how the 3D structure of a protein is achieved from its primary sequence of amino acids, is arguably the most important unsolved problem in Biology. But how this problem is displayed in a biological setting can be quite diverse. One example is how a protein travels from one cell compartment to another, such as from the cytoplasm of a eucaryotic cell to the inside of a mitochondrion, is often dictated by its degree of unfolding. Fully folded proteins can not travel between compartments and often cannot leave the cell. Many important cellular functions such as DNA replication contain specific folding proteins. Many of these studies are more easily done in procaryotic cells where mutants in these processes are more easily obtained. More recently it has been shown that many diseases have important roots in various aspects of protein folding involving: defects in the process of protein folding of a particular protein, defects in secretion, or other protein folding related process. Diseases which are in this category are thought to include certain encephalopathies, Alzheimer's disease, osteogenesis imperfecta, Marfan's syndrome, and cystic fibrosis (8, 9). The overall structure of most proteins is tettering on the verge of structural instability and proteins can be made to

From the Departments of Microbiology & Immunology, and *Anatomy & Cell Biology, Universidad Central del Caribe School of Medicine, Bayamón, Puerto Rico.

Address for correspondence: Dr. Robert Villafañe, Department of Microbiology and Immunology, Universidad Central del Caribe School of Medicine, Call Box, 60-327, Bayamón, P R 00960-6032, email: drbob@uccaribe.edu, telephone: (787) 798-3001 ext 229, FAX: 787-740-4300

aggregate under variety of conditions and an understanding of how a protein folds may lead to a manner in which to divert a protein from going into the aggregated state.

Many milestones have been reached in the study of protein structure and protein folding. It is now often possible, for example, to find conditions to reversibly denature many proteins so as to study protein folding under completely defined *in vitro* conditions (1, 2). Proteins can often be purified to homogeneity in one or a very few steps (using such techniques as His tag, fusion proteins). Once a protein has been purified, structural information can rapidly be obtained from methods such as NMR (nuclear magnetic resonance), x-ray diffraction, fluorescence and Raman spectroscopy, and also from cryoelectron microscopy and DNA sequence analysis. Information from 3D protein structures has been very useful in finding structural and folding motifs. Advances in computational and computer modeling methods allow the construction of tentative 3D structures based on homology with known structures . Often protein folding systems are used that are experimentally tractable and have similarities to other more complex proteins for which no assay or no pure protein can be obtained for study.

Because of technological limitations the most fruitful protein folding studies have occurred within the last thirty years. For over two decades, two facts have been taken as foundations: 1) that statistically speaking the protein molecule can not arrive at its three dimensional structure by a random search of all interactions of its amino acids. It would take too long. This has been dubbed the Levinthal Paradox (3). 2) many proteins contain all the information or instructions to fold within their own amino acid sequence. However, in recent times it has been established that some proteins require the assistance of other proteins to fold (chaperones).

A basic outline of how proteins fold is beginning to emerge (10, 11). Current evidence suggests that proteins can be divided into two classes. There are those proteins that fold by a two state model where they go from the unfolded state directly to the folded state without intermediates. The second class of proteins fold via intermediary protein structures that tend to resemble more the native protein than any other structure. However, both classes of proteins tend to start by folding locally, that is by local interactions that tend to form local structures included among these would be the dominant α-helices and β-sheets structures. It is at this point where the controversy remains. There are many controversies, among these, some are about the major forces that are involved in the dominant protein structures and about the driving forces whether thermodynamically driven or kinetically driven. A continuous controversy is whether these proteins fold through a pathway or not.

An invisible problem comes of age. For many years, protein studies dealt with protein structure and function but not its folding. However, a complex mixture of developments resulted in the focus being placed on how a protein folds. A major impetus was the ability to clone genes and to overexpress their encoded proteins to very high quanities. Many proteins became insoluble at these high quantities (concentrations) and were found to clump or aggregate. Such insoluble aggregates often formed densely packed material in cells called "inclusion bodies". Thus the commercial sector, i. e. biotechnology companies, has been an important driving force for the study of protein folding. Researchers including those at these companies found that overexpression of proteins can lead to insoluble aggregates of proteins which are not functional. Studies have lead to the conclusion that many of those insoluble protein masses resulted from protein misfolding.

A system for protein folding studies. A powerful genetic system has been developed for the study of protein folding. It is being studied in several countries. The system used by several international laboratories is that of the tail protein (or tailspike protein, TSP) of the *Salmonella typhimurium* bacterial virus, P22. No other protein folding system possesses as many beneficial attributes such as: 1) folding mutants can be selected directly (*tsf* mutants, temperature sensitive for folding, and their suppressors , 12-17); 2) a protein folding pathway has been determined *in vivo* which revealed a monomeric and a trimeric folding intermediate, preceding the formation of a thermostable trimeric TSP (18-22); 3) a protein folding pathway has been determined *in vitro* (23-24); 4) it is one of an extremely small number of protein folding systems in which there is a direct correlation between folding intermediates identified *in vivo* and those identified *in vitro* (25-27); one is the P22 TSP folding system while another system is that of the human chorionic gonadotropin β subunit (25); however, this latter system does not have the genetics; 5) the mutant *tsf* TSP interferes with the folding at the stage of the monomer (28-29); 6) the 3D structure of a truncated TSP (missing N-terminal 108 amino acids; out of a total 666 amino acids) has been determined (30); 7) monoclonal antibodies have been obtained which can recognize a monomeric folding intermediate, a trimeric intermediate and the native TSP (31-32); 8) it has become a model for aggregation of proteins (1-7, 27) ; 9) it has potential structural similarities to a number of proteins. 10) The folding and assembly of the P22 TSP has also served as a model for LamB (maltoporin) assembly into the outer membrane (33). No other system has this combination of features.

The system at hand. In general the lifecycle of *Salmonella typhimurium* phages follows that of the other phages (34-35). The initial step in phage infection is adsorption. Adsorption to the bacterial cell surface (36-38) consists of at least three stages: 1) Initial binding to the receptor (without which infection could not occur) which is often reversible; 2) Enzyme-substrate interaction between the TSP trimer (endorhamnosidase) and the O-antigen region of the lipopolysaccharide (LPS); 3) Interactions with the host cell outer membrane leading to DNA ejection from the capsid, possibly at the Bayer junctions or other specific sites (39). Adsorption of the phage to the bacterial cell is a known function of the tail or TSP of bacterial viruses (40-42). The P22 phage requires only one adsorption protein (34-36). Following adsorption, the phage DNA generally get transcribed and a set of more phage-specific genes are then activated. Replication ensures multiple phage progeny. Viral structural proteins accumulate late in infection and assemble into the virion particles. Lysis of the bacterial cell occurs within forty minutes of infection.

The receptor for many *Salmonella* phages is the LPS of the appropriate host strain (36-37, 43-45). There is an absolute requirement for a specific interaction between the P22 TSP and LPS at the cell surface during the infection cycle. This requirement has been clearly demonstrated for the tailspikes of the following bacterial viruses: P22, ε^{15} and ε^{34} and is assumed to apply for all converting (lysogenic) *Salmonella* phages . In general, for *Salmonella* host cells, the O-antigen part of the LPS can be up to 40 O-antigen units long. Each unit consists of three or four saccharide units such as mannose, rhamnose, galactose.

Phage adsorption is of some interest because the same protein which serves as the model for protein folding is also responsible for adsorption. Both adsorption kinetics and fluorescence studies have indicated that there is an interval of time between initial binding and injection of DNA into host cells (46-48). This time interval may represent the time required for the phage to hydrolyze its way down close to the bacterial cell surface to reach its surface receptor. Recent studies have indicated that this step may be much more complicated since the enzymatic turnover is only 2 per minute which is clearly too slow to affect a normal infection if one TSP molecule were to act processively (49). However, the substrates were only 2 to 3 O-antigen units long and were labeled at the reducing end. Under physiological conditions, LPS can have up to 40 O-antigen units. In addition, these studies do not explain the fast infection kinetics with such a slow turnover rate which is uncharacteristic of other glycosyl hydrolases and do not explain the fluorescence data of Bayer (46). The kinetics of TSP binding of these chemically-defined 2 to 3 units substrates, indicated that binding was essentially irreversible.

Previous studies have yielded some information about the specificity of P22 TSP hydrolysis on the O-antigen repeating tetrasaccharides (50-54). The P22 TSP cleaves between the rhamnose and galactose residues. Very recent studies have determined the 3D structure of the P22 TSP complexed with two units of O-antigen bound to it (55-57). It is not yet known what are the amino acids that are most important in the binding to the LPS (R. Villafañe, in progress).

The P22 TSP is the most extensively characterized *Salmonella* phage TSP as well as being one of the most well characterized phage LBP, lipopolysaccharide binding protein (LBP; 57-61). This procaryotic LBP has been purified to homogeneity and in its native form is a trimer (18-20, 61). It has unusual properties such as it is resistant to protease, SDS (if unheated) and to heat with a T_m of about 88°C (21-22, 58). Its gene has been sequenced and its structure has been determined (55-57, 61-62).

The 3D structure of the P22 TSP. Biochemical studies had shown that the P22 TSP was a functional trimer of identical chains and genetic studies had shown that it was the product of gene *9* of the bacterial virus (bacteriophage, or simply phage) P22. The structure of the P22 TSP has been solved recently and it has been shown that other proteins have similar structures such as the plant virulence factors PelC and PelE, alkaline proteases and LpxA, lipid A biosynthetic precursor. These proteins define a new structural class of proteins, the β-helices (63-67). Many *Salmonella* phages use their TSPs to interact with the host LPS to initiate an infection. There are indications that some of these phage TSPs may also be in this structural class (43 and R.Villafañe, unpublished results). One member of the β-helix class, Lpx A, shares a hexapeptide structural motif (which may mean it has structural homology) with a number of other membrane proteins (64). Thus this structural class may eventually consist of a large number of proteins. In this protein class, a major part of the structure of each particular protein consists of β-pleated sheets which traces out a helical pattern in 3 dimensions. A normal helical secondary structure requires about 3.5 amino acids per helical turn but in these structures the number of amino acids needed to make a helical turn is much larger. Since proteins with similar structures fold along similar paths *(*69-70), information on the folding of our model protein, the phage P22 TSP, may lead to insights on the folding of its class members. Knowledge of structure and function of the P22 TSP itself is clinically relevant because at least one member of this class is a plant virulence factor while another is involved in the pertussin toxicity and it serves as a model for other clinical relevant proteins.

Originally, the 3D structure of an N-terminally truncated (missing the first 108 amino acids) P22 TSP was solved at a 2 Å resolution (30; Figure 1 and 2). All figures were derived from the Rasmol Molecular Visualization Program. Figure 2 and all subsequent figures show the "subunit" of the P22 TSP for simplicity and ease of visualization. Recently, the structure of the P22 TSP complexed with LPS (it is actually two units of the most distal part of the LPS, the O-antigen) has been solved and this structure has been refined to 1.56 Å (55, 56, Figures 3, 4). The LPS seems to be cradled by two sets of loops. This latter report also included the structure of the head-binding domain of the TSP. The presence of the N-terminal head-binding of the P22 TSP had previously interfered with the ability of the whole protein to crystallize (26, 30, 57). The three most important amino acid residues in catalysis have been identified by analogy with other glycosyl hydrolases and by site-directed mutagenesis to be Asp392, Asp395, Glu359 (49; and Figure 5). As expected these catalytic amino acids are located within the area that is occupied by the LPS. These catalytic residues have normal LPS binding. The turnover number is just two per minute. Reviews on this aspect of the P22 TSP are available (26, 49, 57).

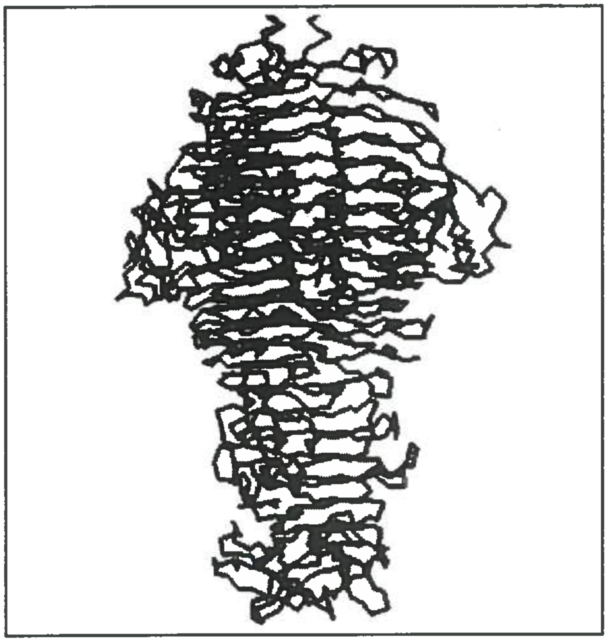

Figure 1. N-truncated- trimeric P22 TSP. This figure displays the trimeric P22 TSP in a Rasmol backbone display mode. In this mode the alpha carbon atom are connected and amino acid side chains are not seen.

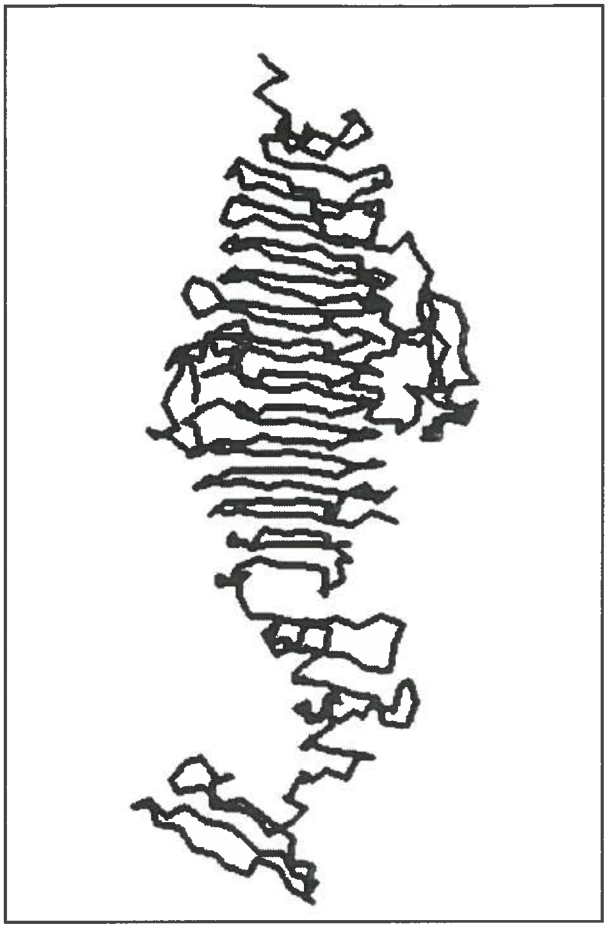

Figure 2. N-truncated "-Monomeric" P22 TSP. This figure is the same as Figure 1 but it just displays the single subunit of the P22 TSP. The N-terminal region starts at the top of the figure while the C-terminus is located at the bottom. To be noted in this figure is the regular periodic structure containing loops. The C-terminus containing the less regular structure is also the means by which the subunits in the trimer are held together through a wrapping around of this C-terminus (interdigitation).

The structure determined is 133 Å long and 35-80 Å wide. P22 TSP consists of a 666 amino acid long polypeptide chain. The β-helix contains the amino acids 143-540 (Figure 6). From this structure protrudes a loop, designated the dorsal fin which contains amino acids 197-259 (Figures 4 and 7). This dorsal fin is very important in folding studies. The C-terminal caudal fin consists of three segments: i aa541-555; ii aa556-619; iii aa620-666. The C-terminal region is where the three identical chains intertwine around each other (interdigitate, 30, 55-57). Although there are eight cysteines per chain, there is no evidence in the native x-ray crystal structure of

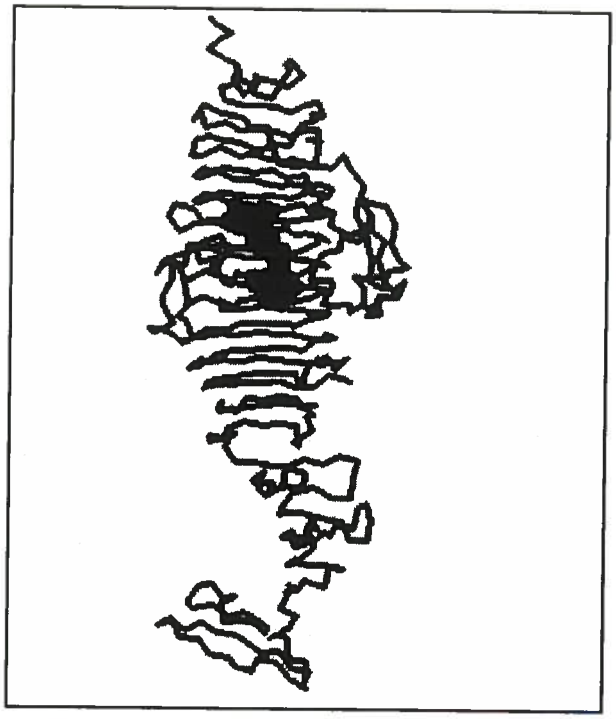

Figure 3. N-truncated- "Monomeric" P22 TSP. (highlighting the LPS, 2 O-antigen units long). This figure is the same as Figure 2 but it contains the LPS (2 O-antigen units) in a spacefill display. The LPS is located in the central periodic region of the P22 subunit. Two large loops are located on each side of the LPS.

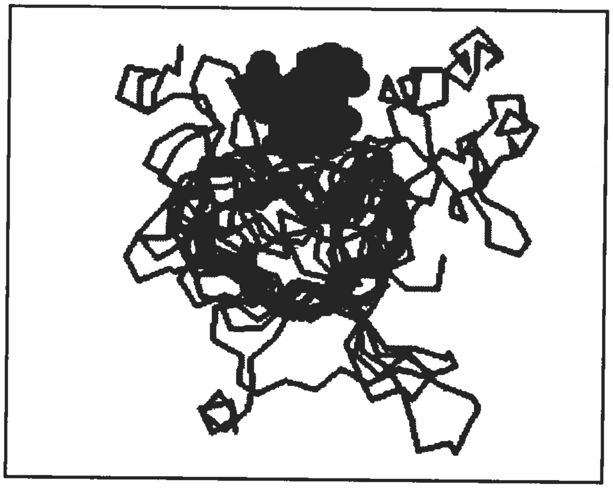

Figure 4. N-truncated- "Monomeric" P22 TSP. (highlighting the LPS with a bottom view of the subunit). This is the view that is obtained when Figure 3 is turned 90° by taking the C-terminal end and lifting it up towards the reader. Note should be taken that the regular periodic mode of the molecule can be seen to be due to a regular periodic circular shape of the polypeptide chain. The LPS (spacefill object) can be readily observed to be located between two loop structures.

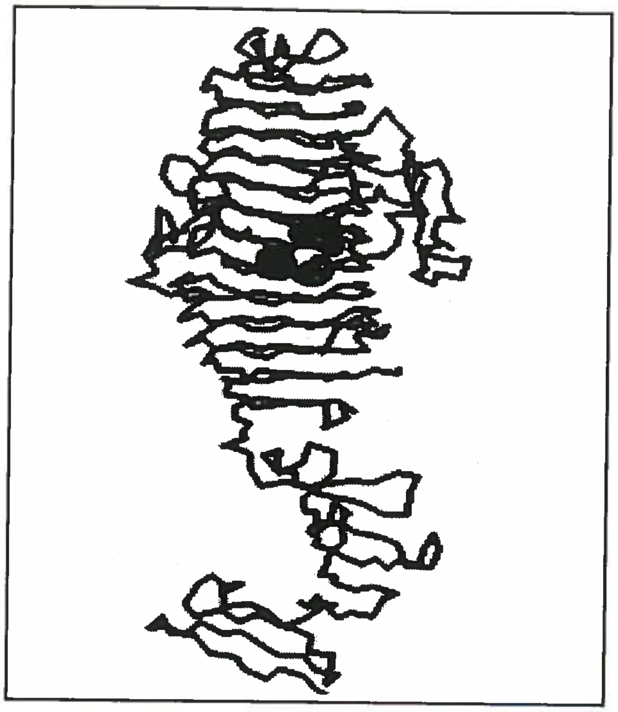

Figure 5. N-truncated- "Monomeric" P22 TSP. (highlighting the three catalytic residues). The threc catalytic amino acid residues, Glu359, Asp392, Asp395 are seen to be located in the same area that contained the LPS. The amino acids are represented by "dots" or van der Waals surfaces of these amino acids.

disulfide bond formation.

A totally genetic folding system. The structural sturdiness of the P22 TSP made it a superb candidate for studying protein folding genetically. It was apparent that once the chain had folded into the native trimeric TSP, it was extraordinarily stable (denatured after 10 min at 80°C, resistant to proteases, and to SDS, if unheated; 19). Since the native state was stable, perhaps the pathway to the stable folded state was more labile. The laboratory of Jon King using conditional lethal mutations (in this case temperature sensitive mutations) was able to show that this was indeed the case (15). Several conditions were favorable and allowed the study of the folding of this P22 TSP *in vivo*. In addition to its stability, the fully folded native P22 TSP could easily be assayed under unusual conditions: since it was stable to SDS, it would migrate to a different position (uncharacteristic of its molecular weight) on a SDS-PAGE than the fully denatured, heated chain. Taking advantage of this peculiar migration of the native P22 TSP trimer and by infecting *Salmonella typhimurium* cells with P22 and using pulse-chase techniques, the kinetics of *in vivo* folding could be determined (18-20). Also important in these early studies was the fact that biological activity of the P22 TSP could be easily assayed as the ability for this virus to form plaques on a bacterial lawn on a petri dish or for its ability to hydrolyze the O-antigen part of the LPS into smaller saccharide units (its endorhamnosidase activity). Intermediates in the folding pathway of this protein did not display these distinctive properties.

Over 100 *tsf* mutants in the tailspike gene have been isolated, defining over thirty sites (12, 15, 17, 71). Once matured at the permissive temperature the tailspike from *tsf* mutants is as stable as the wild type protein. At the restrictive temperature the TSP forms an aggregated species which is not degraded. The protrimer and trimer are not formed in the mutants. These mutants destabilize a thermolabile early intermediate. Early studies had shown that if a cell, which is infected with a *tsf* phage mutant, continues to be incubated at a high restricted temperature, it will not produce any plaques because the TSP is improperly folded. The result is inclusion body formation in the infected cell. However, if the high temperature is lowered within a short period of time, the *tsf* TSP will continue on to a productive pathway and form stable TSP proteins and virus will lyse cells. This reversibility of the *tsf* effects indicated that the critical step was an early one which most probably involved a monomeric folding intermediate.

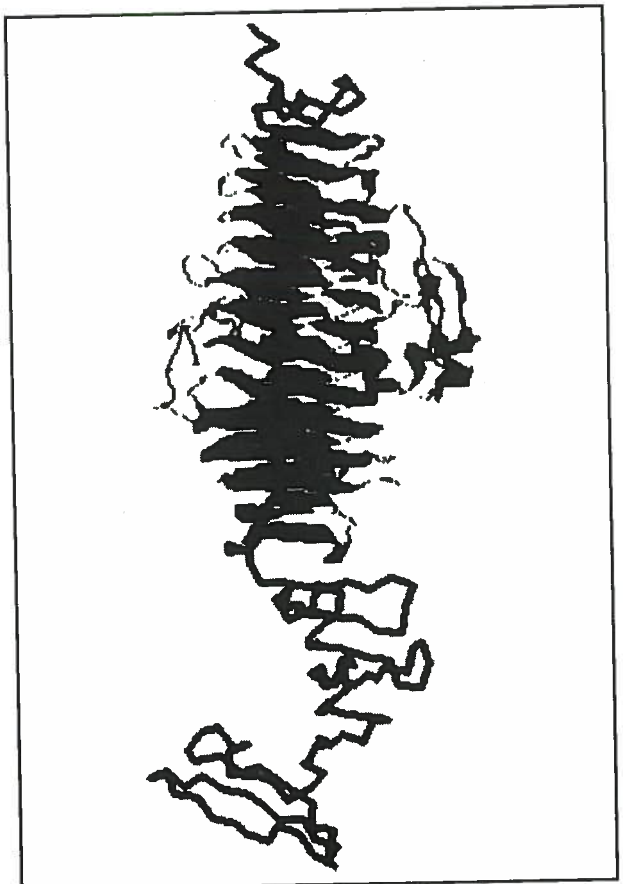

Figure 6. N-truncated-"Monomeric" P22 TSP. (highlighting the beta helix). In this figure the β-helix (amino acids 143-540) is represented by the strand display mode. The β-helix is here shown to contain only the periodic strands of amino acids.

An *in vivo* pathway for the tailspike chain folding and association has been deduced for the TSP (18, 20). *In vivo* the newly synthesized tailspike chain forms a folding intermediate. This converts to a species sufficiently structured for chain-chain recognition. These species associate into the protrimer, in which three chains are associated but not fully folded. In the last step in the pathway: the protrimer folds further to yield the heat stable native spike. Folding occurs both before and after chain association. This subunit association step appears to be the rate-limiting step, probably due to the complicated registration step and a very intricate interweaving of the C-termini from the three polypeptide chains. This protrimer has another interesting characteristic, it contains disulfide bonds which are not present in the native trimeric structure (72). Though the native tailspike is thermostabile, early intermediates in the folding pathway are thermolabile. As a result many tailspikes fail to reach the final conformation at the elevated temperature (19).

Table 1. Folding mutants in the dorsal fin

tsf mutant allele	amino acid change and position
tsfU55	glu<196>lys
tsfU166	thr<199>lys
tsfU5	ser<227>phe
tsfU57	asp<230>val
hyperts	trp<232>xxx
tsfH300	thr<235>ile
tsfU2	ser<238>phe
tsfH304	gly<244>arg
tsfU11	pro<250>ser
tsfU24	ile<258>leu

The dorsal fin loop contains nine sites (out of thirty eight sites) shown to be important for protein folding (Table 1, Figure 7; 17). This suggests that this loop is important for protein folding. Similar loops are present

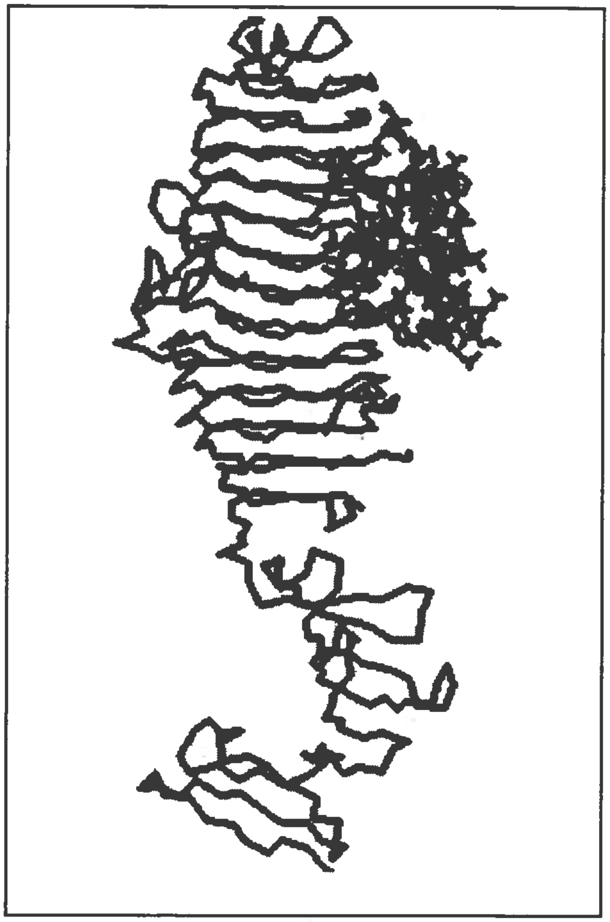

Figure 7. N-truncated- "Monomeric" P22 TSP. (highlighting the dorsal fin). Here the subunit is shown as a backbone structure. The dorsal fin (amino acids 197-259) is shown here in the Stick representation. It is noted as the loop form that is located on the right side of the LPS insertion.

in other β-helix proteins and this result may argue that these loops are also important in structure and folding (63-67).

Before the crystal structure had been obtained, it was argued that the site of the location of the *tsf* mutations must have a repetitive structure since all of the *tsf* mutants had similar characteristics (17). All of the known *tsf* mutations are located within the β-helix region of the protein (amino acid residues 143-540 out of 666 total amino acid residues; Figure 6).

Summary of *tsf* results. The genetic and physiological studies of the P22 TSP showed: 1) the TSP is stable at high temperature (Tm ~88°C) and resistant to proteases and SDS treatments (the latter treatment if unheated, 19); 2) the TSP folding pathway involves two sequential folding intermediates of monomeric and trimeric (Protrimer) size (18-20, 58); technically there is no monomer or subunit since this form of the tailspike is a folding intermediate; 3) only a limited number of sites are involved in its folding path (17); 4) at restriction temperatures the *tsf*TSP forms inclusion bodies (4-7, 22, 28) ; 5) the *tsf* mutations affect the monomer-sized (subunit) folding intermediate (28, 29) ; 6) two suppressor mutants, the global suppressors isolated by the King and Villafañe laboratories, can alleviate many *tsf* defects and are involved in the aggregation path. These suppressor mutants are termed "global suppressors" (13, 14, 16, 28).

The *in vitro* folding pathway. Seckler and coworkers have found conditions in which the trimeric TSP (215KDa) is denatured into unfolded polypeptide chains and its activity and structure is restored after dilution into neutral buffer at 10°C (23, 24). Fluorescence emission, sedimentation analyses, and electrophoretic mobility studies confirmed that the reconstituted protein was identical to the native trimeric TSP. They also showed that reconstitution can be achieved at higher temperatures (30). At these higher temperatures refolding of the *tsf*TSP produced lower tailspike yields similar to what happens *in vivo*.

Characterization of the *in vitro* refolding path by hydrodynamic and electrophoretic migration yielded two folding intermediates of the same size and form as those identified *in vivo* (23). This is the only protein folding system in which it is possible to directly select for folding mutants and in which there is a direct correlation between folding intermediates *in vitro* and *in vivo*.

It had been noted that in the presence of SDS and temperatures between 65-70°C in an SDS PAGE that with time there was complete denaturation of purified P22 TSP with the concomitant appearance of a protein band corresponding to the monomer molecular weight of the P22 TSP. However, before full denaturation there appeared another band that was high molecular weight. Further investigation showed that this band was a thermal intermediate in which the N-terminal 100 or so amino acids were mobile during high temperature incubation and, unlike the P22 TSP, became susceptible to protease (73). Using protease-generated N-truncated P22 TSP, it was also shown that the pattern of formation of the thermal unfolding intermediate and its conversion to the fully unfolded state correlated well with the severity of the original *tsf* defect *in vivo*. With that method, it was shown that the *tsfU2* and the *hyperts* (*Hts*) folding mutants were the most severe *tsf* mutants which corroborated previous *in vivo* studies (Table 1). *Hts* is unique among the *ts* folding mutants because it is temperature sensitive at 30°C. Using a recombinant gene to generate the N-terminally deleted P22 TSP and mutant variants, it was shown that under these thermal unfolding conditions that the *tsf* mutations and suppressors behave similar to the *in vivo* conditions (74).

Mechanism of *tsf* mutations. Using *in vitro* refolding kinetics, Seckler and colleagues have found that the *tsf* mutants affected the "subunit" or "monomer" form of the tailspike and decreased the yield of native tailspike (29). This is the part of the protein folding path which had been genetically defined as the stage at which the *tsf* exerted its affects on the mutant proteins (28). The global suppressors were shown to increase the yield of native TSP at the subunit stage (29).

Interactions between amino acids can be studied genetically. Genetics is an important science in medicine and in all of basic biology because a function can be assigned to a gene. Most often, a mutation defines a gene when its presence directly correlates with a defective (or altered) property or function. These gene assignments are most easily done in a procaryotic cell which does not have the complicating homologous chromosome.

In simple systems, such as the P22 TSP, it is relatively straightforward to identify suppressors (folding interactions) because it is a simple matter to isolate functional mutants at the high restrictive temperature from nonfunctional *tsf* mutants. In genetics a mutation which causes a defective protein identifies the gene and if another mutation is found in the same gene which results in restoration of activity or function, the deduction is that these two mutations have resulted in the corresponding amino acid interacting in a manner to restore activity or function. These two compensatory mutations, located in the same gene, identify genetic interactions. In efforts to identify folding interactions, suppressors to folding mutants have been isolated (13, 14, 16).

Two suppressor mutants, isolated by the King and Villafañe laboratories, can alleviate the defects of many *tsf* mutants. These suppressors appear to correct the aggregation path by increasing the yield of functional TSPs by stabilizing a folding monomeric intermediate. These suppressor mutants have been termed "global suppressors" (13, 14, 16, 28). In the 3D structure the global suppressors are located in the sixth turn of the β-helix where 331 is partly solvent exposed and 334 points into the interior of the β-helix (30). An additional three suppressor pairs have been isolated (16). One of these pairs may identify an ionic interaction important during the folding process but not present as a salt bridge in the native protein structure (16 and R.Villafañe, unpublished data).

The smallest folding units. From the structure of the P22 TSP, it is clear that there is a major structural element located in the central region of the protein , the β-helix (Figure 6). It is the site of the vast majority of the *tsf* mutations (17,71). One recent report has shed light on the importance of this structural feature of the P22 TSP (75). Although previous studies had indicated that the N-terminus was essential for binding of the P22 TSP to the virus head structure (42, 44, 56, 73), other studies had indicated that as far as protein folding was concerned, the N-terminus was not necessary for folding (76). The question now was whether the β-helix was a folding domain and if its structure was similar to the individual chain in the trimer. To address that question, a recombinant gene was prepared (75). This recombinant gene was deleted at the 5' and 3' ends of the gene in order to produce the β-helix gene. The recombinant gene produced a monomeric TSP at low protein concentrations. The monomer TSP was shown to have a spectroscopic signal similar to the native protein. In addition, the monomer β-helix bound LPS and retained enzymatic activity. Because of these characteristics, it could be concluded that the monomer β-helix protein must be similar in structure to one of the subunits in the crystal structure of the P22 TSP (75).

Urea denaturation studies have recently shown that the isolated monomeric β-helix protein exhibits properties of an intermediate in the P22 tailspike folding: the effect of *tsf* mutations is to lower the stability of this protein while the effect of the suppressor mutations is to stabilize the protein (77). This study also showed that the monomer β-helix protein was prone to form aggregates under a variety of conditions. This is a reasonable behavior for a protein conformation that is poised at the point between entry into an aggregation pathway or into a productive protein folding pathway. Since the effect of these mutations (i. e. amino acid substitutions) is expected to occur on the folding intermediate prior to the formation of the protrimer, this study suggests that this monomer β-helix protein exhibits properties of a protein folding intermediate on a folding pathway.

The flip side of folding: folding and aggregation of proteins. An excellent review on this process, containing some relevant and interesting historical views, has recently been published (27). This phenomena is of enormous importance because a change of conditions, such as prolong storage of a therapeutic protein, is often enough to cause protein aggregation. Use of gene carrying vectors for the overexpression of gene products has often resulted in the synthesis and accumulation of the cloned gene at extremely high levels. This overexpression can sometimes even change the nature of the cytoplasm milieu. Such conditions may cause aberrant interactions between highly expressed proteins.

Many overexpressed proteins become trapped in insoluble "inclusion bodies" which can be seen in the electron microscope as crystalline structures occupying large areas of bacteria or yeast cells. Inclusion body formation is so common that it is accepted as a common form of overexpressed product and there are routine procedures to obtain soluble protein from these insoluble masses. It has long been assumed that the formation of aggregates is caused by the interaction between non-specific patches of hydrophobic amino acid residues on the surface of proteins under non-ideal synthetic conditions. The formation of inclusion bodies itself argues strongly for the apparent specific interactions between polypeptide chains.

There are a finite number of folding sites on the protein (17). At higher temperatures, between 35-39°C, the synthesis of the TSP of *tsf* phage mutants results in the formation of inclusion bodies in infected cells. At temperatures of 40°C and above even the wild type P22 TSP forms extensive aggregation.

In the P22 protein folding system, there is a pivotal monomeric folding intermediate which could either proceed into the productive folding pathway or go into the aggregation path which leads to inclusion body formation (4-7, 78). It is at a monomeric stage that the *tsf* defects are expressed. Commitment to the productive or aggregation-specific pathway occurred within seconds of refolding of the urea-denatured P22 TSP (78). Aggregation could be suppressed by initiating refolding in the cold (4°C) and then continuing incubation at 20°C (78). The cold incubation step allowed the accumulation of a monomeric folding intermediate that was past the point of *tsf* sensitivity. These studies suggested that there are at least two monomeric protein folding intermediates. The first intermediate is thermolabile and may be the protein structure that is further destabilized by the *tsf* mutations. This first monomer intermediate is in equilibrium with another aggregation-prone folding intermediate. The second protein monomeric folding is on the productive pathway, past the thermolabile step.

The aggregation of this protein has been shown to proceed by an aggregation pathway (6, 78, 79) and such a pathway may be general (79). A very simple study was done which elegantly showed that in the P22 protein folding system, aggregation is a protein-specific phenomena (79). Using the P22 virus system, it had been shown that both the P22 coat and P22 TSP proteins had folding and aggregation pathways. The electrophoretic gel patterns for the aggregation products of each protein were distinguishable from each other. Aggregates of the P22 coat and TSP proteins were mixed and incubated in the same reaction. No mixed aggregation forms were observed. This study strongly suggests that in some proteins, aggregation is a specific event.

Conclusion. The problem of protein folding can no longer be ignored. The recent award of a Nobel Prize in this area typifies the keen interest and progress that has occurred in recent years (80). The evidence is accumulating that many diseases have their etiologies in the protein folding problem. (8, 9, 80, 81). A number of attributes have made the P22 TSP foldiing system a key for modeling both clinical and basic sciences aspects of protein folding. The P22 TSP has an all β-structure and undergoes aggregation, similar to the β-amyloid disease protein. The *in vitro* and genetic aggregation studies have great commercial and clinical importance. The P22 TSP has a β-helical structure similar to the *Bordetella pertussis* virulence factor (82). Extremely helpful has been the ability to select directly those amino acid sites that are involved in protein folding and the direct correspondence between the *in vivo* and *in vitro* protein folding pathways. It is hoped that this review has illustrated an important protein folding system and that it has also shown the use of procaryotic systems to dissect and understand complicated biological problems.

Resumen

Este artículo describe el uso de un sistema genético simple, que proveee una visión mas profunda del proceso de plegamiento de las proteinas y de su contraparte, la agregación protéica. Estos estudios hacen uso de la proteina de la cola del bacteriófago P22 el cual infecta a *Salmonella typhimurium*. Este sistema de plegamiento sirve como modelo a varios elementos protéicos estructurales y puede aumentar el conocimiento de las enfermedades relacionadas a defectos en el plegamiento de las proteinas, campo que actualmente está en pleno crecimiento.

Acknowledgement

The authors wish to acknowledge discussions with Celia Venza and Clari Salgado as well as use of facilities supported by RCMI grant G12 RR03035 to the Universidad Central del Caribe School of Medicine. The corresponding author also wishes to thank Drs. Jon King at the Massachusetts Institute of Technology and Andrew Wright of Tufts University School of Medicine for long standing discussions and support.

References

1. Jaenicke R. Folding and association versus misfolding and aggregation of proteins. Philos Trans R Soc Lond B-Biol-Sci

1995;348:97-105.

2. Seckler R, Jaenicke R. Protein folding and protein refolding. FASEB J 1993;6: 2545-2552.
3. Levinthal C. Are there pathways for protein folding? J Chim Phys 1968; 65:44-45.
4. Haase-Pettingell C, King J. Formation of aggregate from a thermolabile *in vivo* folding intermediate in P22 tailspike maturation: a model for inclusion body formation. J Biol Chem 1988;263:4977-4983.
5. Mitraki A, King J. Protein folding intermediates and inclusion body formation. Biotechnology 1989;7:690-697.
6. Speed MA, Wang DIC, King J. Multimeric intermediates in the pathway to the aggregated inclusion body state for P22 tailspike polypeptide chains. J Prot Sci 1995;4:900-908.
7. Speed MA, Morshead T, Wang DIC, King J. Conformation of P22 tailspike folding and aggregation intermediates probed by monoclonal antibodies. J Prot Sci 1997;6:99-108.
8. Thomas PJ, Qu B-H, Petersen PL. Defective protein folding as a basis of human disease. Trends Biochem Sci 1995;20:456-459.
9. Horwich AL, Weissman JS. Deadly conformations - Protein misfolding in prion disease. Cell 1995;89:499-510.
10. Baldwin RL, Rose GD. Is protein folding hierarchic? I. Local structure and peptide folding. Trends Biochem Sci 1999;24:26-33.
11. Baldwin RL, Rose GD. Is protein folding hierarchic? II. Folding intermediates and transition states. Trends Biochem Sci 1999;24:77-83.
12. Fane B, King J. Identification of sites influencing the folding and subunit assembly of the P22 tailspike polypeptide chain using nonsense mutations. Genetics 1988;117:157-171.
13. Fane B, King J. Intragenic suppressors of folding defects in the P22 tailspike protein. Genetics 1991;127:263-277.
14. Fane B, Villafañe R, Mitraki A, King J. Identification of global suppressors for temperature-sensitive folding mutations of the P22 tailspike protein. J Biol Chem 1991;261:11640-11648.
15. Smith DH, Berget PB, King J. Temperature-sensitive mutants blocked in the folding or assembly of the bacteriophage P22 tailspike protein. I. Fine-structure mapping. Genetics 1980;96:331-352.
16. Villafañe R, Fleming A, Haase-Pettingell C. Isolation of suppressors of temperature sensitive folding mutations. J Bacteriol 1994;176:137-142.
17. Villafañe R, King J. The nature and distribution of temperature sensitive folding mutations in the tailspike gene of bacteriophage P22. J Mol Biol 1988;204:607-619.
18. Goldenberg DP, King J. Trimeric intermediate in the *in vivo* folding and subunit assembly of the tailspike endorhamnosidase of bacteriophage P22. Proc Natl Acad Sci USA 1982;79:3403-3407.
19. Goldenberg DP, Berget PB, King J. Maturation of the tail spike endorhamnosidase of *Salmonella* phage P22. J Biol Chem 1982;257:7864-7871.
20. Goldenberg DP, Smith DH, King J. Genetic analysis of the folding pathway for the tailspike protein of phage P22. Proc Natl Acad Sci USA 1983;80:7060-7064.
21. Sargent D, Benevides, JM, Yu M-H, King J, Thomas Jr, GJ. Secondary structure and thermostability of the phage P22 tailspike. XX. Analysis by Raman spectroscopy of the wild type protein and a temperature-sensitive folding mutant. J Mol Biol 1988;199:491-502.
22. Sturtevant JM, Yu M-H, Haase-Pettingell C, King J. Thermostability of temperature-sensitive folding mutants of the P22 tailspike protein. J Biol Chem 1989;254:10693-10698.
23. Fuchs A, Seiderer C, Seckler R. *In vitro* folding pathway of the P22 tailspike protein. Biochemistry 1991;30:6598-6604.
24. Seckler R, Fuchs A, King J, Jaenicke R. Reconstitution of the thermostable trimeric phage P22 tailspike protein from denatured chains *in vitro*. J Biol Chem 1989;264: 11750-11753.
25. Huth J R, Mountjoy K, Perini F, Ruddon RW. Intracellular folding pathway of human chorionic gonadotropin b subunit. J Biol Chem 1992;267:8870-8879.
26. Seckler R. Folding and function of repetitive structure in the homotrimeric phage P22 tailspike protein. J Struct Biol 1998;122:216-222.
27. Jaenicke R, Seckler R. Protein misassembly *in vitro*. Adv Prot Chem 1997;50:1-59.
28. Mitraki A, Fane B, Haase-Pettingell C, Sturtevant J, King J. Global suppression of protein folding defects and inclusion body formation. Science 1991;253:54-58.
29. Danner M, Seckler R. Mechanism of phage P22 tailspike protein folding mutations. Protein Sci 1993;2:1869-1881.
30. Steinbacher S, Seckler R, Miller S, Streipe B, Huber R, Reinemer B. Crystal structure of P22 tailspike protein: interdigitated subunits in a thermostable trimer. Science 1994;265:383-386.
31. Friguet B, Djavadi-Ohaniance L, Haase-Pettingell C, King J, Goldberg ME. Properties of monoclonal antibodies selected for probing the conformation of wild type and mutant forms of the P22 tailspike endorhamnosidase. J Biol Chem 1990;265:10347-10351.
32. Friguet B, Djavadi-Ohaniance L, King J, Goldberg ME. *In vitro* and ribosome-bound folding intermediates of P22 tailspike protein directed with monoclonal antibodies. J Biol Chem 1994;269:15945-15949.
33. Misra R, Peterson A, Ferenci T, Silhavy TJ. A genetic approach for analyzing the pathway of LamB assembly into the outer membrane of *Escherichia coli*. J Biol Chem 1991;266:13592-13597.
34. Poteete AR. Bacteriophage P22. In: Calendar R, ed. The bacteriophages. Volume II. New York: Plenum Press; 1988.p.647-682.
35. Poteete AR. P22 bacteriophage. In Webster RG, Granoff A, eds. Encyclopedia of Virology. London: Academic Press; 1994.p.1009-1013.
36. Lindberg AA. Bacteriophage receptors. Ann Rev Microbiol 1973;27:205-241.
37. Lindberg AA. Bacterial surface carbohydrates and bacteriophage adsorption. In: Sutherland I, editor. Surface carbohydrates of the procaryotic cell. New York: Academic Press; 1977. p.289-356.
38. McConnell M, Reznick A, Wright A. Studies on the initial interactions of bacteriophage ε^{15} p.44 with its host cell, *Salmonella anatum*. Virology 1979;94:10-23.
39. Heller KJ. Molecular interactions between bacteriophage and the gram-negative envelope. Arch Microbiol 1992;158:235-248.
40. Casjens S, Hatful G, Hendrix R. Evolution of dsDNA tailed-bacteriophage genomes. Semin Virol 1992;3:383-397.
41. Haggard-Ljungquist E, Halling C, Calendar R. DNA sequences of the tail fiber genes of bacteriophage P2: evidence for horizontal transfer of tail fiber genes among unrelated bacteriophages. J Bacteriol 1992;174:1462-1477.
42. Susskind MM, Botstein D. Molecular Genetics of bacteriophage P22. Microbiol Rev 1978;42:385-413.
43. Greenberg M, Dunlap J, Villafañe R. Identification of the tailspike protein from the *Salmonella newington* phage ε^{34} and partial characterization of its phage-associated properties. J Struct Biol 1995;115:283-289.
44. Israel V, Anderson TF, Levine M. *In vitro* morphogenesis of phage P22 from heads and base-plate parts. Proc Natl Acad Sci USA 1967;57:284-291.
45. Iwashita S, Kanegasaki S. Release of O Antigen polysaccharide from *Salmonella newington* by phage e^{34}. Virology 1975;68:27-34.

46. Bayer ME, Bayer MH. Fast responses of bacterial membranes to virus adsorption: a fluorescence study. Proc Natl Acad Sci USA 1981;78:5618-5622.

47. Israel V. Role of the bacteriophage P22 tail in the early stages of infection. J Virol 1976;18:361-364.

48. Israel V. A model for the adsorption of phage P22 to *Salmonella typhimurium*. J Gen Virol 1978;40:669-673.

49. Baxa U, Steinbacher S, Miller S, Weintraub A, Huber R, Seckler R. Interactions of phage P22 tails with their cellular receptor, *Salmonella* O-antigen polysaccharide. Biophys J 1996:71:2040-2048.

50. Kanegasaki S, Wright A. Studies on the mechanism of phage adsorption: Interaction between phage ε^{15} and its cellular receptor. Virology 1973;52:160-173.

51. Wright A, Kanegasaki S. Molecular aspects of lipopolysaccharide. Physiol Rev 1971;51:748-784.

52. Ericksson U, Lindberg AA. Adsorption of phage P22 to *Salmonella typhimurium*. J Gen Virol 1977;34:207-221.

53. Eriksson U, Svenson SB, Lonngren J, Lindberg, AA. *Salmonella* phage glycanases: Substrate specificity of the phage P22 endorhamnosidase. J Gen Virol 1979;43:503-511.

54. Iwashita S, Kanegasaki S. Smooth specific phage adsorption: endorhamnosidase activity of tail parts of P22. Biochem Biphys Res Chem 1973; 55:403-409.

55. Steinbacher S, Baxa U, Miller S, Weintraub A, Seckler R, Huber R. Crystal structure of phage P22 tailspike protein complexed with *Salmonella* sp. O-antigen receptors. Proc Natl Acad Sci USA 1996;93:10584-10588.

56. Steinbacher S, Miller S, Baxa U, Budisa N, Weintraub A, Seckler R, Huber R. Phage P22 tailspike protein: Crystal structure of the head -binding domain at 2.3 D fully refined structure of the endorhamnosidase at 1.56 D resolution, and the molecular basis of O-antigen recognition and cleavage. J Mol Biol 1997;267:865-880.

57. Steinbacher S, Miller S, Baxa U, Weintraub A, Seckler R. Interaction of *Salmonella* phage P22 with its O-Antigen receptor studied by x-ray crystallography. Biol Chem 1997;378:337-343.

58. King J. Deciphering the rules of protein folding. Chem & Eng News 1989;67:32-54.

59. King J, Fane B, Haase-Pettingell C, Mitraki A, Villafañe R. Genetic analysis of polypeptide chain folding and misfolding *in vivo*. In: Hook JB, Poste G, editors. Protein design and the development of new therapeutics and vaccines. New York: Plenum Press; 1990. p.59-78.

60. Sinnott, ML. Catalytic mechanisms of glycosyl transfers. Chem Rev 1990; 90:1171-1202.

61. Berget PB, Poteete AR. Structure and functions of the bacteriophage P22 tail protein. J Virol 1980;34:234-243.

62. Sauer RT, Krovatin W, Poteete AR, Berget PB. Phage P22 tail protein: gene and amino acid sequence. Biochem 1982;21:5811-5815.

63. Yoder MD, Keen NT, Jurnak F. New domain motif: the structure of pectate lyase C, a secreted plant virulence factor. Science 1993;260:1503-1507.

64. Raetz CRH, Roderick L. A left-handed parallel β–helix in the structure of UDP-N-acetylglucosamine acyltransferase. Science 1995;270:997-1000.

65. Emsley P, Charles IG, Fairweather NF, Isaacs NW. Structure of Bordetella pertussis virulence factor P.69 pertactin. Nature 1996;281:90-92.

66. Yoder MD, Jurnak F. Protein motifs. 3. The parallel beta helix and other coiled folds. FASEB J 1995;9:335-42.

67. Goldenberg DP, Creighton TE. A fishy tail of protein folding. Curr Opin Struct Biol 1994;4:1026-1029.

68. Vuorio R, Harkonen T, Tolvanen M, Vaara M. The novel hexapeptide motif found in the acyltransferases LpxA and LpxD of lipid A biosynthesis is conserved in various bacteria. FEBS Lett 1994;337:289-292.

69. Krebs H, Schmid FX, Jaenicke R. Folding of homologous proteins. The refolding of different ribonucleases is independent of sequence variations, proline content and glycosylation. J Mol Biol 1983;169:619-635.

70. Stackhouse TM, Onuffer JJ, Matthews CR, Ahmed SA, Miles EW. Folding of homologous proteins: conservation of the folding mechanism of the a subunit of tryptophan synthase from Escherichia coli, *Salmonella typhimurium* and five interspecies hybrids. Biochemistry 1988;:824-832.

71. Yu M, King J. Single amino acid substitutions influencing the folding pathway of the phage P22 tail spike endorhamnosidase. Proc Natl Acad Sci USA 1984;81:6584-6588.

72. Robinson AS, King J. Disulfide-bonded intermediate on the folding and assembly pathway of a non-disulfide bonded protein. Nature Struct Biol 1998;4:1997.

73. Chen B-L, King J. Thermal unfolding pathway for the thermostable P22 tailspike endorhamnosidase. Biochemistry 1991;30:6260-6269.

74. Miller S, Schuler B, Seckler R. Phage P22 tailspike protein: removal of head-binding domain unmasks effects of folding mutations on native state-thermal stability. Protein Sci 1998;7:2223-2232.1.

75. Miller S, Schuler B, Seckler R. A reversibly unfolding fragment of P22 tailspike protein with native structure: The isolated β-helix domain. Biochemistry 1998;37:9160-9168.

76. Danner M, Fuchs A, Miller S, Seckler R. Folding and assembly of phage P22 tailspike endorhamnosidase lacking the N-terminal head-binding domain. Eur J Biochem 1993;215:653-661.

77. Schuler B, Seckler R. P22 tailspike mutants revisited: effects on the thermodynamics stability of the isolated b-helix domain. J Mol Biol 1998;281:227-234.

78. Betts SD, King J. Cold rescue of the thermolabile tailspike intermediate at the junction between productive folding and off-pathway aggregation. Protein Sci 1998;7:1516-1523.

79. Speed MA, Wang DIC, King J. Specific aggregation of partially folded polypeptide chains: The molecular basis of inclusion body formation. Nature Biotech 1996;14:1283-1287.1.

80. Prusiner SB. Prion diseases and the BSE crisis. Science 1997; 278:245-251.

81. Cheng SH, Gregory RJ, Marshall J, Paul S, Souza DW, White GA, O'Riordan CR, Smith AE. Defective intracellular transport and processing of CFTR is the molecular basis of most cystic fibrosis. Cell 1990;63:827-834.

82. Emsley P, Charles IG, Fairweather NF, Isaacs NW. Structure of *Bordetella pertussis* virulence factor P.69 pertactin. Nature 1996;381:90-92.