Minireview

Amino acid substitutions influencing intracellular protein folding pathways

Anna Mitraki and Jonathan King

Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received 2 June 1992

Though an increasing variety of chaperonins are emerging as important factors in directing polypeptide chain folding off the ribosome, the primary amino acid sequence remains the major determinant of final conformation. The ability to identify cytoplasmic folding intermediates in the formation of the tailspike endorhamnosidase of phage P22 has made it possible to isolate two classes of mutations influencing folding intermediates – temperature-sensitive folding mutations and global suppressors of *tsf* mutants. These and related amino acid substitutions in eukaryotic proteins are discussed in the context of inclusion body formation and problems in the recovery of correctly folded proteins.

Protein folding in vivo; Global suppressors; Phage P22 tailspike; Inclusion body

1. THE IMPORTANCE OF INTRACELLULAR FOLDING INTERMEDIATES

The nature of the grammar through which amino acid sequences of polypeptide chains direct their three-dimensional fold remains an unsolved problem in molecular biology. A great deal of effort has focused on identifying residues and sequences which contribute to the stability and activity of the native state (for reviews, see [1–4]). However, investigations of the actual pathways through which newly synthesized polypeptide chains reach their native state reveals that considerable sequence information is necessary for folding into the native state, in addition to the information required to stay in that state once achieved.

Within the aqueous environment of the cytoplasm polypeptide chains do not fold through two-state transitions but pass through well-defined, though poorly understood, intermediate states [5–8]. These folding intermediates have to solve a variety of problems: avoiding aggregated states; avoiding proteolysis; reaching particular cellular compartments; transiently binding chaperonins and other auxiliary proteins; and finally, reaching the native state. In fact folding intermediates generally have properties quite distinct from their own native states, in terms of solubility, stability, half-lives and interactions with cellular components, such as chaperonins or signal recognition particles. As a result there is no prior reason to assume that the conformation of folding intermediates will represent simply a partial subset of the native conformation.

One class of sequence information, well defined by cell biologists, are the auxiliary signal sequences involved in bringing the chain to the correct compartment. Another class of auxiliary sequences are actively involved in directing steps in folding and subunit assembly, including the registration peptides of collagen and the prosequences such as those of insulin and and of alpha lytic protease. Nonetheless, the underlying three-dimensional conformation of proteins is determined by the primary amino acid sequence [9]. We have been particularly interested in identifying amino acid sequences *within* the sequence of the structural polypeptide chain which direct the chains into the correct intermediate conformation and through subsequent steps in the pathway. A class of such sequences have been studied for membrane proteins including anchor sequences and signal anchor sequences. [10]. Here we discuss sequences within the structural chains of cytoplasmic proteins or extracellular proteins. These have been identified by the isolation of mutations which affect folding intermediates, without affecting the native state [11].

2. FOLDING INTERMEDIATES, MISFOLDING, AND INCLUSION BODY FORMATION

The diagnostic criterion used for identifying mutations that interfere with productive folding pathways was the accumulation of misfolded polypeptide chains. The generation of misfolded, generally aggregated,

Correspondence address: A. Mitraki, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. Fax: (1) (617) 253 8699.

Published by Elsevier Science Publishers B.V.

chains is also encountered with wild-type amino acid sequences, most notably during the expression of cloned genes in heterologous hosts [12]. This is generally not due to covalent damage or modification, but represents a folding problem, since native protein can frequently be recovered after solubilization of the aggregated material with strong denaturants and subsequently refolded in vitro.

The aggregation of polypeptide chains during in vitro refolding has been carefully studied by Michel Goldberg [13] and Rainer Jaenicke, [14] and co-workers. They found that aggregates were not derived from the native or denatured form of the protein, but from folding intermediates in the pathway. As a result there is kinetic competition between folding intermediates proceeding to the native state and proceeding off pathway to an aggregated state.

In 1989 we reviewed the literature of in vivo and in vitro aggregation, and concluded that in vitro aggregation and inclusion body formation were similar phenomena, originating from the failure of folding intermediates to achieve final folding [15]. In the context of an heterologous intracellular environment, this failure may originate from poor stability of intermediates, absence of essential cofactors or chaperones, etc. The accumulation of partially folded polypeptide species leads to self-association. This self-association seems to be highly selective, since aggregates contain a large percentage of the overexpressed protein, despite the presence of numerous species of partially folded polypeptide chains in the intracellular environment. This selectivity must reflect an aggregation mechanism that proceeds through specific interactions between folding intermediates and, therefore, is in some sense determined by the primary amino acid sequence.

Recent studies of chaperonin function clearly identify that in many cases chaperonins are recognizing folding intermediates that are channeling into the aggregation pathway [16]. The chaperonins transform these to a conformation that is past the off-pathway junction, and are released in a state that proceeds efficiently towards the native state [17,18]. Chaperone proteins are also essential for the folding and assembly pathway of proteins imported into the mitochondrial matrix [19]. Despite the importance of chaperonins in these interactions, the amino acid sequence and environment of folding intermediates remain key players in determing the probability that the chains will successfully fold under given conditions. In fact, reaching the native state is equivalent to avoiding the many non-native but quite stable states polypeptides can fold into.

3. A CYTOPLASMIC FOLDING PATHWAY IN PHAGE-INFECTED BACTERIA

The tailspike of bacteriophage P22 is one of the very few proteins whose cytoplasmic folding intermediates have been identified [7]. A predominantly beta-sheet

Mutation	Residue	Substitution	Local Sequence
tsU9	177	**Gly>Arg**	Phe . Ile . Gly . Asp . Gly . Asn . Leu . Ile . Phe
tsH304	244	**Gly>Arg**	Val . Lys . Phe . Pro . Gly . Ile . Glu . Thr . Leu
tsH302	323	**Gly>Asp**	Asn . Tyr . Val . Ile . Gly . Gly . Arg . Thr . Ser
tsU38	435	**Gly>Glu**	Leu . Leu . Val . Arg . Gly . Ala . Leu . Gly . Val
tsH300	235	**Thr>Ile**	Gly . Tyr . Gln . Pro . Thr . Val . Ser . Asp . Tyr
tsH301	368	**Thr>Ile**	Thr . Trp . Gln . Gly . Thr . Val . Gly . Ser . Thr
tsU18	307	**Thr>Ala**	Asp . Gly . Ile . Ile . Thr . Phe . Glu . Asn . Leu
tsU5	227	**Ser>Phe**	Thr . Leu . Lys . Gln . Ser . Lys . Thr . Asp . Gly
tsN48	333	**Ser>Asn**	Gly . Ser . Val . Ser . Ser . Ala . Gln . Phe . Leu
tsU19	285	**Arg>Lys**	Gly . Phe . Leu . Phe . Arg . Gly . Cys . His . Phe
tsU53	382	**Arg>Ser**	Asn . Leu . Gln . Phe . Arg . Asp . Ser . Val . Val
tsU57	230	**Asp>Val**	Glu . Ser . Lys . Thr . Asp . Gly . Tyr . Glu . Pro
tsmU9	309	**Glu>Val**	Ile . Ile . Thr . Phe . Glu . Asn . Leu . Ser . Gly
tsmU8	344	**Glu>Lys**	Asn . Gly . Gly . Phe . Glu . Arg . Asp . Gly . Gly
tsH303	250	**Pro>Ser**	Glu . Thr . Leu . Leu . Pro . Pro . Asn . Ala . Lys
tsU11	250	**Pro>Leu**	" . " . " . " . " . " . " . "
tsU24	258	**Ile>Leu**	Lys . Gly . gln . Asn . Ile . Thr . Ser . Thr . Leu
tsRAF	270	**Val>Gly**	Glu . Cys . ILe . Gly . Val . Glu . Val . His . Arg
tsU7	311	**Leu>His**	Thr . Phe . Glu . Asn . Leu . Ser . Gly . Asp . Trp
ts9.1	334	**Ala>Val**	Ser . Val . Ser . Ser . Ala . Gln . Phe . Leu . Arg
tsR(am)A	203	**Trp>Gln**	Thr . Thr . Thr . Pro . Trp . Val . Ile . Lys . Pro
am^{ts}H1200	207	**Trp>...**	Val . Ile . Lys . Pro . Trp . Thr . Asp . Asp . Asn
am^{ts}H840	315	**Trp>...**	Leu . Ser . Gly . Asp . Trp . Gly . Lys . Gly . Asn

Table I
Sequences which may kinetically direct turns in the tailspike polypeptide chain.

homotrimer, the tailspike is the cell attachment organelle of phage P22. The native protein is highly thermostabie, with a T_m of 88°C, and is resistant to SDS and proteases [20]. The newly synthesized chains fold and assemble in the host cytoplasm without undergoing any known covalent modifications. During the in vivo folding pathway, the chain passes through single and triple-chain defined folding intermediates that are sensitive to all those factors. Both species can be trapped in the cold and detected by native gel electrophoresis. The SDS-resistance of the native state allows unambiguous identification of the mature polypeptide chains in SDS gels and thus makes it possible to follow the intracellular state of newly synthesized tailspike chains [7,21]. Note that there is no species in the tailspike pathway corresponding to a native monomer – the folding intermediates are in a different conformation from their own native state.

The folding pathway followed by chains refolding from the fully denatured state proceeds similarly, passing through single and triple-chain partially folded intermediates [22,23]. Thus the tailspike provides one of the very few systems where the in vivo pathway off the ribosome can be compared with the in vitro refolding pathway out of denaturant.

At low temperatures (30°C), almost 100% of the newly synthesized folding intermediates chase to the native trimer. As the temperature of folding increases, the early partially folded intermediate partitions between the native pathway and an aggregation pathway leading to inclusion bodies [21]. At 39°C, about 30% of the newly synthesized polypeptide chains reach the mature form, while the remainder partitions to the inclusion body. The temperature dependence of this partitioning suggests that an early intermediate in the pathway is thermolabile and is being partially denatured within the cytoplasmic environment (Fig. 1). The thermal denaturation of folding intermediates, which are generally much less stable than their own native states, is probably one of the major problems to which the inducible heat-shock chaperonins are responding.

If native tailspikes produced at permissive temperatures in vivo are shifted up to restrictive temperatures, the mature tailspikes stay SDS resistant, proving that after the native form is attained, its intracellular stability properties are not altered at high temperatures [24]. Thus the aggregated state is not the product of the intracellular denaturation of the native protein, but represents the polymerization of a folding intermediate that leads to this aggregated intracellular state. If chains that have been synthesized at high temperatures are shifted to low temperature early enough, they can re-enter the productive pathway [21,25]. However, once aggregated the chains cannot be recovered by lowering the temperature. This indicates that aggregation is a conformational trap for folding intermediates that can be kinetically avoided.

4. TEMPERATURE-SENSITIVE FOLDING MUTATIONS IDENTIFYING CRITICAL RESIDUES FOR THE FOLDING PATHWAY OF P22 TAILSPIKE

The clear difference in physical properties between the native state and folding intermediates allowed the identification of a class of mutations, called temperature-sensitive folding (*tsf*), that alter the folding pathway without influencing the properties of the native form [26]. The *tsf* polypeptide chains reach the native state at permissive temperatures in vivo, but fail to reach it at high temperatures [25,27]. The native state of the mutant proteins, once formed at permissive temperature, have similar thermal stability and biological activity properties as the wild-type [24,28,29]. Thus, the residues at the sites of these mutations make very little if any contribution to the stabilization or function of the native state. They behave as if they destabilize the already thermolabile single-chain folding intermediate. They represent a class of informational content inside polypeptide chains which is not crucial for the maintenance of the native form of the protein but makes a significant contribution to the stability and/or kinetic fates of folding intermediates.

Since the native state of the *tsf* mutant polypeptide chains is fully stable at elevated temperatures, the failure of the chains to fold into that state at elevated temperatures must reflect a pathway that is kinetically rather than thermodynamically controlled [24].

The mutations are located at more than 30 sites in the central region of the polypeptide chain and the local sequences resemble those associated with surface beta turns [29,30] (Table I). These sites apparently are kinetically important in directing turns and/or stabilizing a critical folding intermediate in the beta-sheet fold at high temperature. The surface location of these sites explains the tolerance for the mutant substitutions in the native conformation, for example arginines for glycines [29]. Thus recovery of *tsf* mutations selected for surface sites that are kinetically important in chain folding.

5. GLOBAL SUPPRESSORS OF PROTEIN FOLDING DEFECTS AND INCLUSION BODY FORMATION

Starting with mutants kinetically blocked in chain folding, a second set of mutants were selected which alleviated the folding defects of the starting alleles [31]. A group of these seond-site suppressor mutations mapped within the gene coding for the tailspike. Two nearby substitutions in the center of the polypeptide chains were able to suppress diverse *tsf* and absolutely defective amino acid substitutions spanning a 200 residue distance in the primary sequence [32].

The strains containing the suppressor mutations by

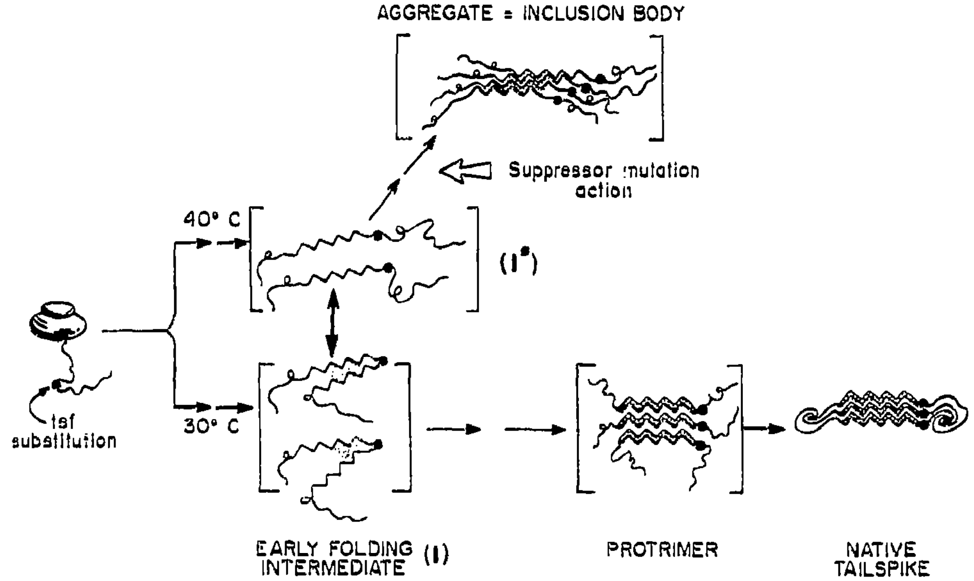
AGGREGATE = INCLUSION BODY
Suppressor mutation action
40° C
(I*)
tsf substitution
30° C
EARLY FOLDING INTERMEDIATE (I)
PROTRIMER
NATIVE TAILSPIKE

Fig. 1. Intracellular folding and association pathway of the P22 tailspike protein. After release from the ribosome, a partially folded single chain intermediate (I) is formed, which can further proceed in the productive pathway and form the protrimer, a species in which the chains are associated but not fully folded. The protrimer folds further into the native spike, with concomitant acquisition of resistance to SDS, proteases and heat. The single chain intermediates are thermolabile even for the wild-type chain, and are partially denatured to form (I*) at restrictive temperature. Tsf mutant locations are important in directing correct turn formation at the early single chain intermediate step. Once this step is completed, they are no longer crucial for the stabilization of the native state. For the mutant chain, the protein can still fold at 30°C, although the stability of the intermediate is lowered, or the equilibrium (between forms I and I*) shifted, but folding competence is retained. However, at restrictive temperature, the equilibrium is shifted to form (I*), this critial interaction cannot be established and the intermediate fails to further proceed in the productive pathway. The accumulation of partially folded beta-sheets leads to illegitimate intermolecular recognition and formation of aggregates.

themselves were not defective in any of their physiological functions. The analysis of the purified suppressor proteins, alone or in combination with a *tsf* mutation, showed that their thermal stability and activity were not altered with respect to wild-type. The chains carrying only the global suppressor mutations mature more efficiently than wild-type at high temperatures, suggesting that they might be 'super-folding' substitutions. In fact examination of the kinetics of intracellular chain folding revealed that the suppressors' substitutions did not change the rate of the intermediate to mature steps in chain folding. Rather they inhibited the loss of folding intermediates to the kinetically trapped inclusion body state [33]. Of additional practical significance is the absence of effects on native function or stability of the suppressor substitutions.

The local sequence surrounding the suppressor mutations is as follows:

Ser-Tyr-Gly-Ser-Val-Ser-Ser-Ala-Gln-Phe-Leu-Arg
331 334

It is difficult to imagine how a single amino acid substitution could alter the aggregation behavior of a 666 residue amino acid chain, if that process were 'non-specific'. Site-specific mutagenesis studies by Myeong-hee Yu and co-workers [34] show that the suppressor phenotype is very specific to the substitution, with only Ala or Gly at site 331 and Val or Ile at 334 functioning as suppressors. The observations support models in which inclusion body formation is a rather specific off-pathway polymerization process involving domains of folding intermediates that would normally productively interact in the intradomain mode (Fig. 1).

Three general models seem possible: (i) the suppressors identify the site on the folding intermediate that initiates the association reaction; (ii) the mutations stabilize the productive conformation of the folding intermediate(s) that the *tsf* mutations are destabilizing; or (iii) the suppressors could create a site for chaperonin recognition or binding that is absent from the wild-type chain.

Regardless of the mechanism this kind of suppressor sequence identifies an additional class of sequence information for the folding of polypeptide chains: sequences that ensure passage of the chain through the productive pathway, by preventing off-pathway traps.

6. MUTATIONAL SUPPRESSION OF INCLUSION BODY FORMATION IN EUKARYOTIC PROTEINS EXPRESSED IN BACTERIA

Inclusion body formation itself can be used as a phenotypic marker for detection of sequences that affect properties of folding intermediates. Mutations that alter inclusion body formation have been described recently for a number of systems. Ronald Wetzel and colleagues have reported that subtle amino acid substitutions can either decrease or increase inclusion body formation for interferon γ and interleukin 1β, without affecting biological activity ([35] and personal communication). The authors have also developed screening methods that allow systematic mutational analysis. The same kind of mutations was found by Rinas et al. [36] on basic fibroblast growth factor. The existence of a global second site suppressor of temperature-sensitive mutants has also been reported for the human receptor-like protein tyrosine phosphatase [37]. This mutation reduced the amount of protein that was intracellularly accumulating in inclusion bodies.

7. A PROREGION KINETICALLY REQUIRED FOR THE FOLDING OF ALPHA-LYTIC PROTEASE

The existence of sequences kinetically required for folding was recently clearly proven in the case of alphalytic protease. This extracellular serine protease is synthesized as a precursor, with a 166 amino acid proregion. The proregion is required for folding in vivo, without being part of the final active form. Expression of the mature portion of the protease in *E. coli* results in little or no activity [38]. Co-expression of the proregion in trans is required in order to reach the native form in vivo, indicating that the proregion could play an essential role in folding [39]. This hypothesis has recently been confirmed with in vitro denaturation–renaturation experiments [40]. Omission of the proregion and unfolding of the mature protein after removal of the denaturant leads to a folding intermediate (I) that is inactive, stable, but remains folding competent. Baker et al. [40] successfully trapped this state under conditions that minimize off-pathway reactions, such as aggregation or proteolysis (i.e. low ionic strength and temperature). At high ionic strength and temperature, this critical intermediate is lost to off-pathway aggregation. The intermediate state remained stable for weeks and there was no detectable interconversion with the native state under similar conditions. When the proregion was added in trans in the refolding mixture, refolding to the active and native state rapidly followed. The authors proposed that the pro-region functions by directly stabilizing the rate-limiting transition state, and by lowering the barrier to allow the I state to further proceed towards the productive pathway. Thus the proregion, although not part of the final native state, might play a similar role with tailspike second site suppressors, namely ensure passage to the productive pathway [41].

8. DISCUSSION

The recognition of a class of sequences that play a unique role in folding pathways suggests the possibility that evolutionary selection operates not only on positions that affect stability and activity of the native state, but also on sequences that contribute to the efficiency of the folding process. The amino acid sequence control of chain folding within cells clearly optimizes response to a variety of variables. Maximizing the yield of physiological folding pathways is a process that differs considerably from maximizing the stability of the native state to environmental stress. It is possible that folding intermediates might have evolved with less than optimal stabilities for a number of proteins, that, however, have stable native states. Therefore, mutations that optimize the stability and/or kinetic fates of intermediates without affecting the stability and activity of the native state may be found for those proteins. Those mutations, aside from being important tools for the study of folding pathways, offer the possibility of improving production of industrially important proteins, without altering their biological function.

Acknowledgements: The Authors were supported by NIH Grant GM 17980 and NSF Grant ECD 8803014, and wish to thank Dr. Florence Davidson for helpful comments.

REFERENCES

[1] Fersht, A. and Leatherbarrow, R.J. (1987) in: Protein Engineering (Oxender D. and Fox, F.C. eds.) pp. 269–278, Liss, New York.
[2] Goldenberg, D.P. (1988) Annu. Rev. Biophys. Biophys. Chem. 17, 481–507.
[3] Shortle, D. (1989) J. Biol. Chem. 264, 5315–5318.
[4] Alber, T. (1990) Annu. Rev. Biochem. 59, 765–798.
[5] Kim, P.S. and Baldwin, R.L. (1982) Annu. Rev. Biochem. 51, 459–489.
[6] Creighton, T.E. (1978) Prog. Biophys. Mol. Biol. 33, 231–298.
[7] Goldenberg, D. and King, J. (1982) Proc. Natl. Acad. Sci. USA 79, 3403–3407.
[8] Gething, M.-J., McCammon, K. and Sambrook, J. (1986) Cell 46, 939–950.
[9] Anfinsen, C.B. (1973) Science 181, 223–230.
[10] von Heijne, G. and Manoil, C. (1991) Protein Eng. 4, 109–112.
[11] King, J., Fane, B., Haas-Pettingell, C., Mitraki, A., Villafane, R. and Yu, M.-H. (1990) in: Protein Folding (Gierasch, L. and King, J. eds.) pp. 225–240, American Association for the Advancement of Science, Washington, DC.
[12] Marston, F.A.O. (1986) Biochem. J. 240, 1–12.
[13] London, J., Skrzynia, C. and Goldberg, M. (1974) Eur. J. Biochem. 47, 409–415.
[14] Zettlmeissl, G., Rudolph, R. and Jaenicke, R. (1979) Biochemistry 18, 5567–5571.
[15] Mitraki, A. and King, J. (1989) Bio/Technology 7, 690–697.
[16] Goloubinoff, P., Christeller, J.T., Gatenby, A. and Lorimer, G.H. (1989) Nature 342, 884–889.
[17] Buchner, J., Schmidt, M., Fuchs, M., Jaenicke, R., Rudolph, R., Schmid, F.X. and Kiefhaber, T. (1991) Biochemistry 30, 1586–1591.
[18] Langer, T., Lu, C., Echols, H., Flanagan, J., Hayer, M.K. and Hartl, F.-U. (1992) Nature 356, 683–689.
[19] Ostermann, J., Horwich, A., Neupert, W. and Hartl, F.-U. (1989) Nature 341, 125–130.
[20] Goldenberg, D.P., Berget, P.B. and King, J. (1982) J. Biol. Chem. 257, 7864–7871.

[21] Haase-Pettingell, C. and King, J. (1988) J. Biol. Chem. 263, 4977–4983.

[22] Seckler, R., Fuchs, A., King, J. and Jaenicke, R. (1989) J. Biol. Chem. 264, 11750–11753.

[23] Fuchs, A., Seiderer, C. and Seckler, R. (1991) Biochemistry 30, 6598–6604.

[24] Sturtevant, J., Yu, M.-H., Haase-Pettingell, C. and King, J. (1989) J. Biol. Chem. 264, 10693–10698.

[25] Smith, D.H. and King, J. (1981) J. Mol. Biol. 145, 653–676.

[26] Smith, D.H., Berget, P.B. and King, J. (1980) Genetics 96, 331–352.

[27] Goldenberg, D. and King, J. (1981) J. Mol. Biol. 145, 633–651.

[28] Yu, M.-H. and King, J. (1984) Proc. Natl. Acad. Sci. USA 81, 6584–6588.

[29] Yu, M.-H. and King, J. (1988) J. Biol. Chem. 263, 1424–1431.

[30] Villafane, R. and King, J. (1988) J. Mol. Biol. 204, 607–619.

[31] Fane, B. and King, J. (1991) Genetics 127, 263–277.

[32] Fane, B., Villafane, R., Mitraki, A. and King, J. (1991) J. Biol. Chem. 266, 11640–11648.

[33] Mitraki, A., Fane, B., Haase-Pettingell, C., Sturtevant, J. and King, J. (1991) Science 253, 54–58.

[34] Lee, S.C., Koh, H. and Yu, M.-H. (1991) J. Biol. Chem. 266, 23191–23196.

[35] Wetzel, R., Perry, L.J. and Vielleux, C. (1991) Bio/Technology 9, 731–737.

[36] Rinas, U. et al. (1992) Bio/Technology 10, 435–440.

[37] Tsai, A.Y.M., Itoh, M., Streuli, M., Thai, T. and Saito, H. (1991) J. Biol. Chem. 266, 10534–10543.

[38] Silen, J.L., Frank, D., Fujishige, A., Bone, R., Agard, D.A. (1989) J. Bacteriol. 171, 1320–1325.

[39] Silen, J.L. and Agard, D.A. (1989) Nature 341, 462–464.

[40] Baker, D., Sohl, J.L. and Agard, D.A. (1992) Nature 356, 263–265.

[41] Creighton, T.E. (1992) Nature 356, 194–195.

NOTE ADDED IN PROOF

Single temperature sensitive mutations that affect the folding pathway but not the conformational stability of the native state have been reported recently for the heterodimeric enzyme luciferase (T. Baldwin, personal communication).