Recombination of protein domains facilitated by co-translational folding in eukaryotes

William J. Netzer* & F. Ulrich Hartl*†

* Cellular Biochemistry & Biophysics Program and Howard Hughes Medical Institute, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, USA

† Max-Planck-Institut für Biochemie, Am Klopferspitz 18a, 82152 Martinsried, Germany

The evolution of complex genomes requires that new combinations of pre-existing protein domains successfully fold into modular polypeptides. During eukaryotic translation model two-domain polypeptides fold efficiently by sequential and co-translational folding of their domains. In contrast, folding of the same proteins in *Escherichia coli* is post-translational, and leads to intramolecular misfolding of concurrently folding domains. Sequential domain folding in eukaryotes may have been critical in the evolution of modular polypeptides, by increasing the probability that random gene-fusion events resulted in immediately foldable protein structures.

The evolution of eukaryotes has been characterized by a dramatic increase in the diversity and structural complexity of proteins. This has been attributed to the organization of eukaryotic genes into exons and introns, which is thought to have increased the probability of favourable gene duplication events as well as the generation of proteins with new functions by joining initially separate DNA sequences ('exon shuffling')^{1–3}. Proteins generated in this way typically contain two or more domains linked sequentially that are encoded either by distinct exons (in the case of more recently evolved structures⁴) or by blocks of exons². Importantly, domains (~100–300 amino acids in length) are not only units of function but also represent autonomous structural units of folding^{5–7}. Independent of the precise genetic mechanism involved, combining ancestral protein domains that were originally separate is apparently the basic principle by which many polypeptides with new functions

evolved. This process has occurred with much higher frequency in eukaryotes than in prokaryotes, as is evident from a simple comparison of prokaryotic and eukaryotic protein structures using the information recently provided by several genome projects (Fig. 1). Although the size distribution of protein domains is uniform throughout the kingdoms of life², eukaryotes contain a much greater number (and proportion) of longer polypeptides combining multiple domains.

In vitro refolding studies indicate that the tendency of polypeptide chains to misfold increases significantly with their chain length, whereas spontaneous folding is usually efficient for small single-domain proteins⁸. Furthermore, expression of eukaryotic multi-domain proteins in bacteria frequently leads to aggregation⁹. One might therefore predict that eukaryotic cells have developed specific mechanisms to ensure the efficient folding of these proteins. For

Figure 1 A comparison of protein length distributions for *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *E. coli* and the archaeobacterium *Methanococcus jannaschi*. Protein number, based on genome analysis, is plotted against protein length (number of amino acids), where length windows correspond to 50 amino acids. The total number of proteins analysed per genome (*n*), the average polypeptide length and the fraction of proteins exceeding a length of 500 amino acids (aa) are given. The last 1% of proteins (>2,000 aa) are not included. Data are derived from the following sources and represent essentially complete genomes, except for *C. elegans*, for which about 95% of the genome is sampled here: *C. elegans*, Batch Entrez NCBI protein database, <http://www3.ncbi.nlm.nih.gov:80/cgi-bin/EntrezBatch/nph-batch/result>; *S. cerevisiae*, FTP://darcy.bu.edu/pub/genomes/; *E. coli* and *M. jannaschi*, TIGR Microbial Database (MDB) of the Institute for Genomic Research, <http://www.tigr.org:80/tdb/mdb/mdb.html>. The following individuals are acknowledged for their help in performing this analysis: M. Adams and O. White (Institute for Genomic Research, Rockville, MD), T. Smith and J. Freeman (The BioMolecular Engineering Center, Boston, MA), and L. Hillier (Washington University, St Louis, MO).

example, their folding might be mediated by the cellular machinery of molecular chaperones, in particular by the cylindrical chaperonins, which can prevent aggregation by sequestering a folding polypeptide within a central cavity¹⁰. However, the chaperonin of the eukaryotic cytosol seems to mediate the folding of only a specific subset of polypeptides^{11,12}. Moreover, the lumen of the endoplasmic reticulum, the site of folding of many composite polypeptides, lacks a chaperonin altogether.

Here we tested the hypothesis that sequential folding of domains during synthesis on ribosomes minimizes the error rate of folding for complex proteins. This mechanism would reduce the problem of folding a large polypeptide to that of folding small polypeptide modules. We have constructed new composite proteins by joining single-domain proteins with a flexible linker. *In vitro* refolding of these fusion proteins is inefficient as a result of intramolecular misfolding. In contrast, during eukaryotic translation, the model proteins reach their native state efficiently by sequential and co-translational folding of domains. Remarkably, folding of the same proteins in a bacterial system is post-translational and leads to aggregation. We propose that co-translational, domain-wise folding is the basis for the efficient folding of a large number of eukaryotic multidomain proteins.

Refolding of two-domain proteins *in vitro*

A fusion protein was constructed consisting of two monomeric, single-domain cytosolic proteins, human H-Ras (M_r 21K) and mouse dihydrofolate reductase (DHFR, M_r 20K). The domain folds of these proteins exist also in prokaryotes. The carboxy terminus of Ras was joined to the amino terminus of DHFR by means of a flexible polypeptide linker containing Ser-Gly-Ile

repeats¹³ (Fig. 2a). Similar linkers join the domains in many eukaryotic proteins, including those that function in signal transduction. Two fusion proteins (with and without a C-terminal six-histidine (His) tag) were expressed in bacteria, mainly as inclusion bodies. The purified proteins were denatured in 6 M guanidinium-HCl, and refolding was conducted by rapid dilution into a refolding buffer. Refolding of Ras-DHFR-His was about 80% efficient for both domains (Fig. 2b). However, Ras-DHFR (minus the His tag) folded only with 20% efficiency. Native fusion protein was efficiently cleaved in the linker region by proteinase K (see below). The protease resistance of the Ras and DHFR domains within the folded fusion protein was indistinguishable from that of the unlinked proteins. In contrast, attempted refolding of Ras-DHFR produced mostly misfolded protein that was readily digested in its non-aggregated state (data not shown).

Refolding of Ras and DHFR was analysed at submicromolar protein concentrations where sedimentable aggregates did not form. Mixing the two single-domain proteins in their unfolded states (in *trans*) had no effect on the yield and rate of refolding (Fig. 2b). The time constants (τ) for folding of the Ras and DHFR proteins were 3 and 1.7 min, respectively. (Refolding rates for DHFR and DHFR-His were indistinguishable). However, in the His-tagged fusion protein both Ras and DHFR domains (in *cis*) refolded concurrently with an 8-fold slower time constant of ~ 24 min, suggesting that there is intramolecular interference between the Ras and DHFR polypeptides.

In contrast to Ras-DHFR-His, refolding of Ras-DHFR occurred with low yield, perhaps because of intermolecular aggregation or intramolecular misfolding. Aggregation is a multimolecular process, but intramolecular misfolding is unimolecular and is therefore

Figure 2 Refolding *in vitro* of Ras-DHFR fusion proteins. **a**, Diagram of the synthetic fusion protein, Ras-DHFR. The C terminus of human H-Ras is linked covalently to the N terminus of mouse DHFR through a 12-residue flexible peptide linker¹³. A second version of the fusion protein (Ras-DHFR-His) contains a six-histidine tag at the C terminus. **b**, Kinetics and yields of refolding upon dilution from 6 M GdmCl into refolding buffer at 25 °C for Ras-DHFR, Ras-DHFR-His, and mixtures of Ras and DHFR. Final protein concentrations were 0.5 μ M for the fusion proteins and for Ras and DHFR. Activities (³H-GDP binding for Ras (filled squares) and enzymatic activity for DHFR (open squares)) are expressed as percentage of native, single-domain Ras and DHFR controls. Refolding kinetics of DHFR and DHFR-His do not differ significantly, so only the kinetics of refolded DHFR is shown. Refolding of mixtures of Ras and DHFR were indistinguishable from folding of each protein in isolation. Data were fitted to first-order exponentials.

Figure 3 Partitioning of *in vitro* refolded Ras-DHFR into native and non-native folds. **a**, Refolding of Ras-DHFR at varying final concentrations measured as the increase in DHFR activity over time (see Fig. 2b). Insert shows concentration independence of refolding yields. **b**, Tryptophan fluorescence spectra excited at 295 nm for refolded Ras-DHFR (intermediate) in comparison to native Ras-DHFR-His and denatured Ras-DHFR in 6 M GdmCl (0.5 μ M final protein concentration). The six-histidine tag did not influence the fluorescence signal.

Figure 4 Efficient folding of Ras-DHFR upon translation in reticulocyte lysate. Formation of ^{35}S -methionine-labelled protease-resistant Ras and DHFR domains on translation in reticulocyte lysate detected by immunoprecipitation. Translation was for 40 min at 25°C in a non-synchronized reaction. Lanes 1–3, translation product was incubated in the absence or presence of proteinase K (PK, 20 and 50 $\mu\text{g ml}^{-1}$); lanes 4–7, immunoprecipitations with anti-Ras and anti-DHFR antibodies. Analysis by 12.5% SDS-PAGE and phosphorimaging. FL, full-length protein.

Figure 5 Co-translational folding of Ras-DHFR in the eukaryotic cytosol. **a**, Synchronized translation of Ras-DHFR in rabbit reticulocyte lysate. Reactions were stopped by dilution into ice-cold buffer containing cycloheximide either with or without proteinase K (+PK and -PK) and analysed by SDS-PAGE and fluorography (top). FL, full-length protein. Note that the protease-resistant Ras domain contains only half as many methionines as the DHFR domain. Data from synchronized translations were quantified by densitometry (average of three independent experiments) and corrected for the differences in methionine contents in Ras and DHFR. **b**, Expression of Ras-DHFR in COS cells. Transfected COS cells were labelled at 37°C with ^{35}S -methionine. Cells were lysed in ice-cold buffer containing digitonin or digitonin and PK. Polypeptides were immunoprecipitated with anti-Ras and anti-DHFR antibodies and analysed by SDS-PAGE and phosphorimager quantification. (Expression levels of endogenous Ras and DHFR are less than 10% of that of the fusion protein.) Emergence of each protein species is shown as the fraction of the amount detected at 120 s of labelling when the relative amounts of the different proteins were constant. The approximate rate of polypeptide synthesis was calculated by measuring the time required after adding ^{35}S -methionine to cells to achieve a constant proportion of labelled, proteolysed Ras domain relative to labelled, full-length product. The time required to aminoacylate tRNA was taken to be 15–30 s.

concentration independent. The fraction of protein that reached the native state remained the same ($\sim 20\%$) at increasing concentrations of fusion protein in the range 25–500 nM, indicating that intermolecular aggregation was not governing the efficiency of refolding (Fig. 3a). Dynamic light scattering determined that the non-native intermediate, once formed, remained monomeric (M_r 43K) up to low micromolar concentrations (data not shown). The intermediate had the following physical properties: a maximum tryptophan fluorescence emission at 340 nm (compared with 331 nm for native Ras-DHFR-His), consistent with a collapsed conformation in which Trp residues are incompletely buried (Fig. 3b); the presence of significant α -helical structure by far-ultraviolet circular dichroism spectroscopy; enhanced binding of the fluorescent probe 8-anilidonaphthalene-1 sulphonate compared to the native protein; and hardly any cooperativity of unfolding (data not shown).

In summary, during refolding of Ras-DHFR, the two domains interfere intramolecularly, resulting in the accumulation of a non-native intermediate(s). The presence of the histidine tag seems to destabilize this intermediate such that slow folding to the native state occurs.

Co-translational folding in eukaryotic cytosol

Remarkably, the Ras-DHFR fusion protein folded with almost full efficiency on translation in a rabbit reticulocyte lysate. Folding was assayed by partial proteinase K digestion followed by immunoprecipitation of Ras and DHFR domains (Fig. 4). Nearly all newly synthesized protein was soluble and was cleaved in the linker into Ras and DHFR fragments, the signature of the native fusion protein.

Next we determined the timing of domain folding during synthesis by synchronizing translation with aurintricarboxylic acid, which inhibits reinitiation of translation¹⁴ and allows only a single round of elongation (Fig. 5a). The time required to synthesize a single molecule of Ras-DHFR was approximately 9 min at 24°C. Formation of protease-resistant Ras domain occurred consistently before the full-length fusion protein, indicating that Ras folded co-translationally. The folded DHFR domain appeared only after synthesis of full-length protein was completed, consistent with the notion that folding of a single protein domain generally requires its complete extrusion from the ribosome¹⁵. Equimolar amounts of Ras and DHFR bands were present at late time points of translation, demonstrating that the fragments resulted from digestion of productively elongating polypeptide chains (note that DHFR contains eight methionines, compared with just four in the protease-stable Ras domain). Thus, in contrast to refolding *in vitro*, during cell-free translation folding of Ras-DHFR fusion protein occurs in a domain-wise manner and is highly efficient. Attempts to detect an interaction of translating Ras-DHFR with the eukaryotic chaperonin TRiC were unsuccessful (data not shown), suggesting that co-translational folding of Ras-DHFR is independent of chaperonin.

We then investigated whether co-translational folding of Ras-DHFR occurred *in vivo* and at 37°C. Folding was assayed in COS cells transiently transfected with the fusion construct (Fig. 5b). It was not feasible to produce a synchronized translation *in vivo*. However, we reasoned that if co-translational folding were to occur, it could be observed after short times of labelling. The approximate rate of polypeptide synthesis was calculated as one Ras-DHFR polypeptide per 120 s, corresponding to an elongation rate of ~ 3 amino acids per second (see Fig. 5b). Radiolabelling of cells was stopped after various times by dilution into ice-cold buffer containing cycloheximide and digitonin with or without proteinase K. Folded Ras domain was seen in immunoprecipitations after just 30 s of labelling, at a time when very little labelled full-length product and folded DHFR domain were detectable (Fig. 5b). Based on the rate of translation, any labelled, full-length Ras-DHFR occurring at that time could be labelled only in the DHFR domain. Therefore, any labelled, protease-resistant Ras appearing at this time must be

derived from protein that folded before elongation was completed. This folded Ras is likely to represent productively translating protein, because the proportion of Ras domain to full-length protein at later time points corresponded to the composition of the fusion protein (data not shown). Thus co-translational, domain-wise folding occurs for Ras-DHFR under physiological conditions *in vivo*. Furthermore, virtually all of the protein expressed in COS cells was folded and soluble (Table 1).

Post-translational folding in bacterial cytosol

Only 10% of total Ras-DHFR-His was soluble and native when expressed in *E. coli* (at 25–37°C), independent of the time of induction. In contrast, no folded Ras-DHFR (minus His tag) was detected under the same conditions of expression (Table 1). Because the C-terminal His tag could only influence a post-translational folding mechanism, it was likely that folding in *E. coli* might occur after the fusion polypeptide is released from the ribosome.

Figure 6 Post-translational folding of Ras-DHFR-His in *E. coli* cytosol. **a**, Expression of Ras-DHFR-His in *E. coli* *in vivo*. Time course of production of full-length protein in total (open squares), soluble fraction (open circles) and insoluble pellet (filled circles), and of soluble, proteinase K-resistant Ras (open diamonds) and DHFR (open triangles) domains. Spheroplasts were pulse-chase labelled at 30°C with ³⁵S-methionine. Amounts of Ras are corrected for methionine content. Analysis by SDS-PAGE and phosphorimager quantification. **b**, Synchronized translation of Ras-DHFR-His and β -lactamase (from the same plasmid) at 25°C in *E. coli* S30 lysate. Reactions were stopped by dilution into ice-cold buffer containing chloramphenicol (top); chloramphenicol and proteinase K (PK) (middle); or chloramphenicol, CDTA and PK (bottom). In the bottom panel, PK was added 5 min after CDTA. FL, full-length Ras-DHFR-His. Analysis was by SDS-PAGE and phosphorimaging. CDTA treatment was correlated with the release of nascent chains from ribosomes, as demonstrated by sedimentation analysis. Note that β -lactamase is synthesized as a precursor containing a hydrophobic signal sequence. Instability of β -lactamase in the presence of PK may be enhanced by binding to chaperones, such as GroEL.

To explore this possibility, viable spheroplasts prepared from *E. coli* cells expressing Ras-DHFR-His were labelled with ³⁵S-methionine at 30°C, followed by unlabelled methionine after 15 s (Fig. 6a). Ras-DHFR-His was the predominant polypeptide species synthesized. Further incorporation of radiolabel into newly synthesized polypeptide was effectively prevented after 2 min of chase (data not shown), but production of labelled full-length protein continued owing to completion of elongating chains (Fig. 6a). At different times spheroplasts were rapidly cooled and lysed with digitonin in the presence or absence of proteinase K. As soon as full-length protein was synthesized about 50% of the protein aggregated and was recovered in the pellet fraction. This material was protease resistant (data not shown). The remainder was soluble and protease sensitive, but only ~25% was stably folded, as indicated by cleavage in the linker region into protease-resistant Ras and DHFR domains. Furthermore, the two domains folded concurrently, and only after full-length protein had been synthe-

Figure 7 Post-translational and co-translational folding of *E. coli* OmpR. **a**, Left, schematic representation of the two-domain structure of OmpR. The linker segment has the sequence RQANELPGAPSQEE (single-letter code starting with residue 122)^{18,19}. Arrows indicate the positions of the major trypsin cleavage sites in the folded protein. Right, SDS-PAGE of OmpR after 20 min translation in S30 followed by trypsin treatment. FL, full-length OmpR; fragment 1-190 results from trypsin cleavage in the folded C-terminal domain. It consists of two peptides encompassing residues 1-190 and 1-193 (ref. 18). The two lower fragments correspond to the N-terminal domain. **b**, Synchronized translation of OmpR in *E. coli* S30 lysate (top) and in reticulocyte lysate (bottom). Analysis was as in Fig. 6b and Fig. 5b, respectively, except that trypsin was used to monitor folding. Amounts of full-length OmpR (open circles), as well as trypsin-resistant fragments 1-121 plus 1-142 (N domain) (filled squares) and of fragment 1-190 (C domain) (open squares), were quantified and are expressed as a percentage of the amounts present at 20 (S30) or 25 min (reticulocyte lysate) of translation. At the times indicated, translation reactions were incubated with or without CDTA (5 min at 25°C in the presence of chloramphenicol) followed by trypsin. Insets show analysis by SDS-PAGE. Note that the speed of polypeptide elongation is very similar in the two systems, although initiation of translation is faster in bacterial S30.

sized (Fig. 6a). During the chase period a second, slower phase of protein aggregation occurred from the soluble pool of Ras–DHFR molecules that did not reach the native state (Fig. 6a). These results suggested that folding of the fusion protein occurred by a post-translational mechanism and was inefficient as a result of intramolecular misfolding.

Translation in *E. coli* S30 lysate was used for further investigation. Compared with expression in intact *E. coli*, Ras–DHFR–His synthesized in S30 folded with a higher efficiency of 30% (Fig. 6b and Table 1). The remainder of the full-length protein was relatively protease resistant, suggesting the formation of misfolded or aggregated species as *in vivo*. (Almost no folded Ras or DHFR domain was produced from Ras–DHFR (Table 1).) A synchronized translation confirmed that folding of Ras–DHFR–His was indeed post-translational (Fig. 6b, top and middle panels), in marked contrast to the observations in reticulocyte lysate. The speed of translation in S30 and reticulocyte lysate, and also the physical conditions of translation (pH, protein and salt concentrations), are very similar. Furthermore, the amount of protein synthesized in S30 is small compared to the amount *in vivo*, and inclusion bodies are absent. Thus, misfolding on translation does not seem to be dependent on the physiological speed of *E. coli* translation, nor on the concentration of the newly synthesized polypeptide or the presence of pre-existing aggregates.

To test whether it is the association with translating *E. coli* ribosomes that prevented polypeptide chains containing a completed Ras domain from folding, the Mg^{2+} -chelator CDTA was added to release ribosome-bound chains¹⁶ before digestion with proteinase K (Fig. 6b, bottom). This demonstrated that, before the appearance of full-length protein, nearly all of the Ras domain synthesized (but still ribosome bound) was foldable. Thus, during ongoing translation in the absence of CDTA, folding of Ras is delayed until full-length protein is synthesized. The proportion of foldable Ras was maximal when ribosome release was induced before synthesis of the C-terminal DHFR. At later time points the amount of foldable Ras decreased concomitantly with the increase observed in protease-resistant, apparently misfolded full-length protein (Fig. 6b, middle). This effect is very similar to the misfolding of fusion protein during refolding *in vitro*. (CDTA treatment reduced the protease resistance of misfolded full-length protein, presumably by destabilizing protein aggregates, but had no effect on the amount of folded Ras and DHFR.)

Chelating Mg^{2+} is expected to stabilize the interaction of unfolded polypeptide chains with ATP-dependent chaperones such as DnaK (Hsp70) and GroEL^{15,17}, rendering bound polypeptides sensitive to digestion by protease. It is therefore unlikely that these components are responsible for the prevention of co-translational folding of the Ras domain. Moreover, increasing the concentrations of chaperones in S30 lysate by 5- to 10-fold by adding purified DnaK, DnaJ, GrpE, GroEL and GroES did not affect the efficiency or timing of folding (data not shown). Thus the failure of the bacterial system in folding the fusion proteins co-translationally is related either to properties of the bacterial ribosome itself, or to components of the translation/folding machinery that interact tightly with nascent polypeptide chains, but dissociate upon chain release from the ribosome.

Post- and co-translational folding of OmpR

To determine whether post-translational folding in *E. coli* lysate is a unique property of Ras–DHFR, we performed similar experiments using the natural bacterial transcription factor OmpR (M_r 27K). OmpR consists of an N-terminal phosphorylation domain and a C-terminal DNA-binding domain joined by a ~12-residue flexible linker^{18,19} (Fig. 7a). Analysis of the folded structure of OmpR has been established using limited trypsin digestion¹⁸. Only a few of the 34 potential trypsin cleavage sites are accessible in the native protein. Cleavage at a major site in the C-terminal domain yields a fragment of residues 1–190 containing the entire N-terminal

Table 1 Folding efficiencies upon synthesis (% of total protein)

	Ras	DHFR	Ras–DHFR	Ras–DHFR–His
<i>E. coli</i>	>90*	>90	<2	10
S30 lysate	>90	>90	<5	30
COS cells	n.d.	n.d.	>90	n.d.
Reticulocyte lysate	>90	>90	>90	>90

Folding efficiencies of Ras, DHFR and Ras–DHFR fusion proteins upon synthesis *in vivo* and *in vitro* are expressed as the fraction of soluble protein having native structure as confirmed by limited proteolysis and DHFR activity measurements (for *E. coli* only). Data are averages of 3–6 independent experiments.

* Ras expression in *E. coli* was also determined in ref. 39.

domain, the linker and part of the C-terminal domain (Fig. 7a). Its appearance signifies that both the C-terminal and the N-terminal domains are folded. The two lower fragments (residues 1–122 and 1–142) are derived from cuts in or close to the linker region (occurring with lower efficiency) and contain the N-terminal domain plus part of the linker¹⁸. Appearance of these fragments alone indicates that only the N-terminal domain has folded. The remaining C-terminal segment (residues ~142–190) is not resolved on the gel.

Folding of OmpR was analysed during translation in S30 and reticulocyte lysates (Fig. 7b). Efficient folding in the bacterial lysate was similar to that of Ras–DHFR–His in that folded domains were produced only after synthesis of full-length protein (Fig. 7b, top). Again, CDTA-induced release from ribosomes showed that folding of the N-terminal domain was retarded as long as the protein was associated with ribosome (Fig. 7b, inset). Nevertheless, the N-terminal domain folded slightly faster than the C-terminal domain. An intrinsically slower rate of folding of the C-terminal domain could be an adaptation of OmpR to allow productive post-translational folding. Alternatively, the N-terminal domain may initiate folding co-translationally, but reaches its native, protease-resistant state only after synthesis of full-length polypeptide. Because the speed of bacterial translation *in vivo* is much more rapid than in S30 lysate, this would strongly favour a completely post-translational folding mechanism of OmpR. Remarkably, domain folding of OmpR was sequential and co-translational in reticulocyte lysate (Fig. 7b, bottom). Addition of CDTA before trypsin treatment did not increase the amount of folded N domain (Fig. 7b, insert), indicating that its folding was not retarded by the eukaryotic ribosome. These results confirm that the eukaryotic and bacterial translation systems differ in their ability to support the co-translational folding of modular proteins.

Discussion

Our observations provide evidence for the biological significance of co-translational polypeptide chain folding. Rapid and efficient folding of modular model proteins in the eukaryotic system depends on sequential folding of their domains during synthesis. Strikingly, the bacterial translation system lacks the capacity for folding these model proteins co-translationally, and produces misfolded species that form inclusion bodies.

There are estimated to be only about 1,000 protein modules with unique three-dimensional structures^{7,20}. Based on our observations, we propose that a shift from post-translational to co-translational folding mechanisms has occurred during evolution. This may have provided the basis for the efficient exploitation of a limited set of available structures to evolve modular polypeptides with new functions through gene fusion events (the ‘folding shift hypothesis’). Alternatively, the restricted capability of prokaryotes to fold these proteins, along with their streamlined genomic organization, may explain the limited complexity of prokaryotic proteins. In the folding shift hypothesis we do not specify whether the ability to support co-translational domain folding has been lost (or reduced) by prokaryotes (perhaps together with split genes²), or whether it has been acquired by eukaryotes. Our distinction

between prokaryotic and eukaryotic folding does not predict that bacteria are unable to generate immediately foldable (and thus selectable) proteins from random gene fusions, but that such events would be rarer in organisms that do not generally use co-translational folding.

Co- versus post-translational folding. Co-translational folding has been observed in eukaryotic systems^{15,21–23}, but has not previously been identified as a mechanistic requirement for successful folding of modular proteins. Independent support for the predominance of post-translational folding in bacteria comes from our finding that most of the polypeptides synthesized in the *E. coli* cytosol can associate with the chaperonin GroEL before folding (unpublished observations), although GroEL does not recognize ribosome-bound chains^{10,24}. The physical basis for the difference between the prokaryotic and eukaryotic translation and folding machineries remains to be defined. Generally, the higher speed of bacterial translation (~15 residues per second²⁵) would favour post-translational folding. However, post-translational folding was observed even when the speed of translation was reduced *in vitro* to that of the eukaryotic system. Prevention of folding of ribosome-bound protein in *E. coli* may be caused by either molecular chaperones or by the bacterial ribosome itself.

Although we do not exclude the possibility that co-translational folding may occur for certain bacterial proteins (for example, exceptionally large polypeptides such as β -galactosidase), it is plausible that the selection (or maintenance) of co-translational folding mechanisms may not have been stringent in bacteria, where rapid polypeptide elongation is geared towards short generation times. Present-day bacterial proteins may be selected to fold rapidly when released from the ribosome. This may explain why the solubility of small eukaryotic proteins in *E. coli* sometimes increases upon their translation as fusions with a bacterial 'carrier' polypeptide (or with glutathione S-transferase), independent of whether the carrier is fused to the N terminus or the C terminus of the eukaryotic protein^{26,27}. On the other hand, the inefficient folding of many modular eukaryotic proteins in *E. coli*⁹ may result from the basic inability of bacteria to fold these proteins co-translationally. Another effect of co-translational folding in eukaryotes may be the expression of enzymes cooperating in specific metabolic pathways as large, multidomain proteins, whereas in prokaryotes these activities are synthesized as separate polypeptides²⁸.

The role of molecular chaperones. The bacterial chaperonin acts in the post-translational folding of a wide variety of cytosolic proteins^{10,29}, but a general chaperonin seems to be missing from the eukaryotic cytosol. Unlike GroEL, the eukaryotic chaperonin TRiC is thought to be involved in the folding of only a restricted subset of cytosolic proteins, including actin and tubulin^{11,12}. It is conceivable that co-translational folding might have eliminated the selective pressure for a more generalized chaperonin activity by reducing the formation of kinetically trapped, aggregation-sensitive intermediates. Actin is composed of domains that consist of discontinuous amino-acid sequences³⁰, and, although multidomain, would not be expected to fold co-translationally. Proteins with similar architecture may have a strong dependence on TRiC for folding.

Many translating chains in eukaryotic cells seem to interact with Hsp70, however^{31,32}. The Hsp70 chaperone system is thought to prevent the aggregation of nascent polypeptide chains synthesized on polyribosomes. This may be critical before complete synthesis of a domain (and its extrusion from the ribosome) that is required for stable folding¹⁰. ATP-dependent binding and release cycles of Hsp70 and co-chaperones may be sufficient to mediate the co-translational folding of modular proteins, whereas shielding within the chaperonin cavity during folding may be needed predominantly by proteins with structurally unstable (or discontinuous) domains that are unable to fold independently. Thus, the capacity of eukaryotic cells to fold complex, modular proteins co-translation-

ally may have simplified the chaperone requirement of *de novo* protein folding. □

Methods

Plasmid construction. Human H-Ras (Val 12) cDNA and mouse DHFR cDNA were cloned in *E. coli* TG1 from pSP64 and pGEM-4 (Promega), respectively. Ras cDNA and DHFR cDNA were combined in a three-way ligation with a synthetic DNA linker encoding Val-Leu-Ser (the C-terminal tripeptide of Ras)-Gly-Gly-(Ser-Gly-Gly)₂-Ser-Gly-Ile-Met¹³. One clone was sequenced and identified as pSP64/pGEM-Ras-DHFR. Ras-DHFR and Ras-DHFR-His were cloned into pET3A (Novagen) using the polymerase chain reaction (PCR) and pSP64/pGEM4-Ras-DHFR as template. Ras-DHFR was also cloned into plasmid pcDNA3 (Invitrogen) for transient transfection of COS cells. Additionally, H-Ras (Val 12) and mouse DHFR were cloned into pET 3A for expression in *E. coli*.

Protein purification. Pet3A-Ras-DHFR-His was expressed in *E. coli* strain BL21 by IPTG induction (3 h at 37 °C). Protein was purified to >95% on Ni-NTA³³ either from the soluble fraction or after dissolving insoluble protein in 6 M guanidinium-HCl (GdmCl). Purified protein was transferred into buffer A (20 mM Tris HCl, pH 7.4, 80 mM KOAc, 1 mM MgOAc, 2 mM DTT) containing 6 M GdmCl in case of protein purified from inclusion bodies. Protein concentrations were based on measurements of absorption at 280 nm (A_{280}). pET-3A-Ras-DHFR was induced as above. Inclusion bodies were washed, dissolved in 6 M GdmCl and purified to >80% by reverse-phase HPLC (C4, VYDAC column) over a 30–60% gradient of acetonitrile/0.1% trifluoroacetic acid. When native Ras-DHFR-His (previously purified on Ni-NTA) was subjected to the same purification scheme, no differences in specific activity of refolded Ras-DHFR-His were observed compared to Ras-DHFR-His purified by Ni-NTA alone. DHFR-His and DHFR (mouse) were expressed in soluble form and purified on Ni-NTA or methotrexate-agarose³³, respectively.

In vitro refolding. Protein dissolved in 6 M GdmCl was rapidly diluted 100-fold into refolding buffer (20 mM Tris-acetate, pH 7.4, 80 mM KOAc, 1 mM MgOAc, 2 mM DTT) containing protein ligands (50 μ M dihydrofolic acid or 2–5 μ M GDP). Refolding of DHFR was measured in refolding buffer containing 50 μ M dihydrofolic acid and 50 μ M NADPH by following the decrease in A_{340} over time¹⁷. Ras renaturation was measured in the presence of 2 μ M ³H-GDP (1.0 mCi ml⁻¹, NEN), and bound ³H-GDP was determined³⁴.

In vitro transcription and translation. *In vitro* transcriptions for translation in rabbit reticulocyte lysate were performed using pSP64/pGEM-4-Ras-DHFR linearized with BglII as template and SP6 polymerase (Promega). Translations were carried out in 50–70% nuclease-treated reticulocyte lysate (Promega) in the presence of 0.8 mCi ml⁻¹ ³⁵S-methionine (>1,000 Ci mmol⁻¹, Amersham) at 25 °C (ref. 15). Coupled *in vitro* transcription/translation in *E. coli* S30 lysate³⁵ was carried out at 25 °C with pET3A Ras-DHFR, pET3A Ras-DHFR-His, and the *E. coli* OmpR gene (inserted in pJES307, a derivative of pT7-7; a gift from A. West)³⁶ as templates in the presence of 0.6 mCi ml⁻¹ ³⁵S-methionine, 0.06–0.1 mg ml⁻¹ plasmid DNA, 2 μ g ml⁻¹ rifampicin, and 1.5 units ml⁻¹ T7 RNA polymerase. Translations were synchronized by adding 100 μ M aurintricarboxylic acid (Sigma) 3–9 min after addition of RNA (reticulocyte lysate) or 2.5–3 min after addition of DNA (S30 lysate)¹⁴.

For limited proteolysis, 3–5 μ l of translation reaction was diluted 50-fold in ice-cold buffer A (minus DTT) containing 5 mM MgOAc, 2 mM unlabelled methionine and 2 mM cycloheximide (reticulocyte lysate) or 10 μ g ml⁻¹ chloramphenicol (*E. coli* S30). Proteinase K (Boehringer) was present at 10 or 20 μ g ml⁻¹ and trypsin (Boehringer) at 10 μ g ml⁻¹, and incubation proceeded on ice for 10 min. Proteinase K was inhibited by addition of 1 mM phenylmethylsulphonylfluoride (PMSF) and trypsin by 2 molar equivalents of soybean trypsin inhibitor. When used, CDTA (10 mM) was also present during proteolysis. In several experiments the original Mg²⁺ concentration of the buffer was restored after CDTA induced dissociation of ribosomes (before addition of protease) to demonstrate that protease activity was not Mg²⁺ dependent. Proteolysed samples were precipitated by trichloroacetic acid (TCA) and analysed by SDS-PAGE.

Expression and labelling of Ras-DHFR *in vivo*. Transfected COS cells³⁷ were suspended at ~10⁶ cells ml⁻¹ in Dulbecco's modified Eagle medium lacking methionine and incubated for 30 min at 37 °C. Cells (800 μ l) were pulse

labelled with 1.7 mCi ml⁻¹ ³⁵S-methionine. Portions of 50 µl were removed at times after addition of label and mixed in 200 µl ice-cold buffer A (minus DTT) containing 2 mM cycloheximide and 0.5% digitonin with or without 5 µg ml⁻¹ proteinase K (see above). Reactions were centrifuged at 15,800g at 4 °C for 10 min (pellets did not contain Ras-DHFR). TCA precipitates of supernatants were dissolved in 2% SDS buffer. Portions were diluted 1:10 in 10 mM Tris acetate, pH 7.4, 0.3 M NaCl, 5 mM EDTA, and 1% Triton X-100, containing either 10 µg ml⁻¹ anti-Ras monoclonal antibody (pan Ras AB-3; Oncogene Science), or 1/10 volume of anti-DHFR antiserum from rabbit, and incubated for 1 h on ice, followed by addition of protein A-Sepharose.

PET3A-Ras-DHFR-His was expressed in *E. coli* BL21 by IPTG induction (2 h at 30 °C). Before spheroplasting³⁸, cells were grown at 30 °C in M63 minimal medium lacking methionine (doubling time ~90 min). Spheroplasts were incubated with 60 µCi ml⁻¹ ³⁵S-methionine at 30 °C followed by addition of 1 mM unlabelled methionine after 15 s. Incorporation of label (TCA-precipitable radioactivity) stopped completely 2 min after adding unlabelled methionine. Portions were diluted 2-fold into ice-cold buffer A (minus DTT)/0.5% digitonin (lysis buffer) with or without proteinase K. After inhibition of proteinase K, reactions were separated into pellets and supernatants (5 min, 15,800g at 4 °C). Pellets were washed in lysis buffer and analysed either by SDS-PAGE or by immunoprecipitation for Ras and DHFR proteins.

Received 24 March; accepted 9 June 1997.

- Gilbert, W. The exon theory of genes. *Cold Spring Harb. Symp. Quant. Biol.* **111**, 901–905 (1987).
- Blake, C. C. F. Exons and the evolution of proteins. *Int. Rev. Cytol.* **93**, 149–185 (1985).
- Doolittle, R. F. & Bork, P. Evolutionary mobile modules in proteins. *Sci. Am.* **269**, 50–56 (1993).
- Stolzfus, A., Spencer, D. F., Zuker, M., Logsdon, J. M. & Doolittle, W. F. Testing the exon theory of genes: The evidence from protein structure. *Science* **265**, 202–207 (1994).
- Richardson, J. S. Describing patterns of protein tertiary structure. *Methods Enzymol.* **115**, 341–380 (1985).
- Doolittle, R. F. The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64**, 287–314 (1995).
- Orengo, C. A., Jones, D. T. & Thornton, J. M. Protein superfamilies and domain superfolds. *Nature* **372**, 631–634 (1994).
- Jaenicke, R. Folding and association of proteins. *Prog. Biophys. Mol. Biol.* **49**, 117–237 (1987).
- Marston, F. A. O. The purification of eukaryotic polypeptides synthesized in *Escherichia coli*. *Biochem. J.* **240**, 1–12 (1986).
- Hartl, F. U. Molecular chaperones in cellular protein folding. *Nature* **381**, 571–580 (1996).
- Kubota, H., Hynes, G. & Willison, K. The chaperonin containing t-complex polypeptide 1 (TCP-1)—Multisubunit machinery assisting in protein folding and assembly in the eukaryotic cytosol. *Eur. J. Biochem.* **230**, 3–16 (1995).
- Lewis, S. A., Tian, G. L., Vainberg, I. E. & Cowan, N. J. Chaperonin-mediated folding of actin and tubulin. *J. Cell Biol.* **132**, 1–4 (1996).
- Weissman, J. S. & Kim, P. The Pro region of BPTI facilitates folding. *Cell* **71**, 841–851 (1992).
- Stewart, M. L., Grollman, A. P. & Huang, M.-T. Aurintricarboxylic acid: Inhibitor of initiation of protein synthesis. *Proc. Natl Acad. Sci. USA* **68**, 97–101 (1971).
- Frydman, J., Nimmesgern, E., Ohtsuka, K. & Hartl, F. U. Folding of nascent polypeptide chains in a high molecular mass assembly with molecular chaperones. *Nature* **370**, 111–117 (1994).

- Gesteland, R. F. Unfolding of *Escherichia coli* ribosomes by removal of magnesium. *J. Mol. Biol.* **18**, 356–371 (1966).
- Martin, J. et al. Chaperonin-mediated protein folding at the surface of groEL through a 'molten globule'-like intermediate. *Nature* **352**, 36–42 (1991).
- Kenney, L. J., Bauer, M. D. & Silhavy, T. J. Phosphorylation-dependent conformational changes in OmpR, an osmoregulatory DNA-binding protein of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **92**, 8866–8870 (1995).
- Martinez-Hackert, E. & Stock, A. M. The DNA-binding domain of OmpR: crystal structure of a winged helix transcription factor. *Structure* **5**, 109–124 (1997).
- Chotia, C. One thousand families for the molecular biologist. *Nature* **357**, 543–544 (1992).
- Bergman, L. W. & Kuehl, W. M. Formation of an intrachain disulfide bond on nascent immunoglobulin. *J. Biol. Chem.* **254**, 8869–8876 (1979).
- Chen, W., Helenius, J., Braakman, I. & Helenius, A. Cotranslational folding and calnexin binding during glycoprotein synthesis. *Proc. Natl Acad. Sci. USA* **92**, 6229–6233 (1995).
- Kolb, V. A., Makeyev, E. V. & Spirin, A. S. Folding of firefly luciferase during translation in a cell-free system. *EMBO J.* **13**, 3631–3637 (1994).
- Gaitanaris, G. A., Vysokanov, A., Hung, S.-Z., Gottesman, M. & Gragerov, A. Successive action of *E. coli* chaperones in vivo. *Mol. Microbiol.* **14**, 861–869 (1994).
- Bremer, H. & Dennis, P. P. In *Escherichia coli and Salmonella: Cellular and Molecular Biology* (ed. Neidhardt, F. C.) 1553–1569 (ASM, Washington DC, 1996).
- LaVallie, E. R. & McCoy, J. M. Gene fusion expression systems in *Escherichia coli*. *Curr. Opin. Biotechnol.* **6**, 501–506 (1995).
- Sharrocks, A. D. A T7 expression vector for producing N- and C-terminal fusion proteins with glutathione S-transferase. *Gene* **138**, 105–108 (1994).
- Hawkins, A. R. & Lamb, H. K. The molecular biology of multidomain proteins. Selected examples. *Eur. J. Biochem.* **232**, 7–18 (1995).
- Horwich, A. L., Low, K. B., Fenton, W. A., Hirschfield, I. N. & Furtak, K. Folding in vivo of bacterial cytoplasmic proteins: role of GroEL. *Cell* **74**, 909–917 (1993).
- Kabsch, W., Mannherz, H. G., Suck, D., Pai, E. F. & Holmes, K. C. Atomic structure of the actin: DNase I complex. *Nature* **347**, 37–44 (1990).
- Beckmann, R. P., Mizzen, L. A. & Welch, W. J. Interaction of Hsp 70 with newly synthesized proteins: implications for protein folding and assembly. *Science* **248**, 850–854 (1990).
- Nelson, R. J., Ziegelhoffer, T., Nicolet, C., Werner-Washburne, M. & Craig, E. A. The translation machinery and 70 kd heat shock protein cooperate in protein synthesis. *Cell* **71**, 97–105 (1993).
- Mayhew, M. et al. Protein folding in the central cavity of the GroEL–GroES chaperonin complex. *Nature* **379**, 420–426 (1996).
- Feuerstein, J., Goody, R. S. & Wittinghofer, A. Properties and characterization of nucleotide-free and metal ion-free p21 'Apoprotein'. *J. Biol. Chem.* **262**, 8455–8458 (1987).
- Pratt, J. M. In *Transcription and Translation: A Practical Approach* (eds Hames, B. D. & Higgins, S. J.) 179–210 (IRL, Oxford, 1984).
- Cameau, D. E., Ikenaka, K., Tsung, K. L. & Inouye, M. Primary characterization of the protein products of the *Escherichia coli* ompB locus: structure and regulation of synthesis of the OmpR and EnvZ proteins. *J. Bacteriol.* **164**, 578–584 (1985).
- Felgner, P. L. et al. Lipofection: a highly efficient, lipid-mediated DNA-transfection procedure. *Proc. Natl Acad. Sci. USA* **84**, 7413–7417 (1987).
- Ausubel, F. M. et al. *Current Protocols in Molecular Biology* pp. 6.11.5 (Wiley, New York, 1993).
- Tucker, J. et al. Expression of p21 proteins in *Escherichia coli* and stereochemistry of the nucleotide-binding site. *EMBO J.* **5**, 1351–1358 (1986).

Acknowledgements. We thank K. Ewalt for assistance with the *in vivo* labelling experiments; M. K. Hayer-Hartl for the fluorescence analysis; R. Deep and E. Breslow for circular dichroism measurements; J. E. Rothman, J. Young, M. Tector and W. Houry for critically reading the manuscript; and various colleagues for providing support in the genomic analysis (listed in Fig. 1).

Correspondence and requests for materials should be addressed to F.U.H. (e-mail: uhartl@biochem.mpg.de).

KNOW YOUR COPY RIGHTS RESPECT OURS

The publication you are reading is protected by copyright law.
Photocopying copyright material without permission is no different from stealing a magazine from a newsagent, only it doesn't seem like theft.

If you take photocopies from books, magazines and periodicals at work your employer should be licensed with CLA.

Make sure you are protected by a photocopying licence.

The Copyright Licensing Agency Limited
90 Tottenham Court Road, London W1P 0LP
Telephone: 0171 436 5931 Fax: 0171 436 3986