doi:10.1016/j.jmb.2010.03.066 J. Mol. Biol. (2010) 400, 579–588

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Physicochemical Determinants of Chaperone Requirements

Gian Gaetano Tartaglia[1], Christopher M. Dobson[1], F. Ulrich Hartl[2] and Michele Vendruscolo[1]*

[1]*Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK*

[2]*Department of Cellular Biochemistry, Max Planck Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany*

Received 1 February 2010; accepted 26 March 2010
Available online
21 April 2010

Edited by B. Honig

We describe a series of stringent relationships between abundance, solubility and chaperone usage of proteins. Based on these relationships, we show that the need of *Escherichia coli* proteins for the chaperonin GroEL can be predicted with 86% accuracy. Furthermore, from the observation that the abundance and solubility of proteins depend on the physicochemical properties of their amino acid sequences, we demonstrate that the requirement for GroEL can also be predicted directly from the sequences with 90% accuracy. These results indicate that the physicochemical properties of the amino acid sequences represent an essential component of the cellular quality control system that ensures the maintenance of protein homeostasis in living systems.

Keywords: molecular chaperones; protein solubility; protein abundance; protein homeostasis

Introduction

The presence of misfolded or incompletely folded proteins, which almost invariably lack all activity, not only represents an energetic drain on the cell but, in the case of the malfunctioning of cellular quality control mechanisms, can also result in the accumulation of aggregates that range from inclusion bodies in bacteria to amyloid fibrils in mammals.[1,2] Such assemblies can cause the impairment of biological processes and affect the viability of the organism.[1–5] Indeed, protein misfolding and aggregation phenomena are known to be associated with more than 30 human diseases, and amyloid fibrils or their oligomeric precursors have been found to be involved in many of the most debilitating, feared and rapidly proliferating pathologies of the modern world, including Alzheimer's disease, Parkinson's disease, Huntington's disease, Creutzfeldt–Jakob disease and type II diabetes.[1,6]

It is becoming increasingly clear that the maintenance of protein solubility and hence the avoidance of misfolding and aggregation are crucial requirements for proteins to perform their functions in the crowded environment of the cell.[7,8] A recent analysis of a set of human proteins has revealed a close relationship between the aggregation rates of the proteins and the expression levels of their corresponding mRNA molecules.[9] Since the aggregation propensities of proteins can be accurately predicted from their amino acid sequences,[10–12] the link between expression level and aggregation propensity offers a series of opportunities to understand the factors defining the behaviour of proteins in a cellular context. In an initial study based on this conclusion, we have shown that it is indeed possible to predict the order of magnitude of mRNA expression levels using the physicochemical properties of the corresponding amino acid sequences.[13]

In this study, we have explored the possibility of identifying the physicochemical principles that underlie the interactions between proteins and molecular chaperones. This study has been prompted by recent reports in which the level of abundance of specific proteins in living systems has been linked to their requirements for chaperones in order to fold successfully[14] and to maintain their solubility.[15] As the solubility and abundance of proteins can be predicted from their amino acid sequences,[12,13] it should be possible to define specific properties of the sequences that determine their dependence on

*Corresponding author. E-mail address: mv245@cam.ac.uk.

Abbreviations used: emPAI, exponentially modified protein abundance index; ENO, enolase; METK, *S*-adenosylmethionine synthetase; GATD, galactitol-1-phosphate 5-dehydrogenase.

chaperones. We show here that this conclusion is correct and we formulate methods to predict the DnaK and GroEL requirements of *Escherichia coli* proteins from their amino acid sequences.

We have considered *E. coli* proteins in order to establish these principles since the proteome of this organism is relatively small and characterized in great detail. The most extensively studied molecular chaperones in *E. coli* are the ATP-dependent DnaK/DnaJ/GrpE and GroEL/GroES systems that are required for the folding of a subset of proteins, including a number of essential enzymes.[16–19] According to current models for chaperone-assisted protein folding, substrate proteins released from DnaK/DnaJ/GrpE either fold into their native conformations or are transferred into the central cavity of GroEL.[20] When the latter binds to its cofactor GroES, a large complex that can prevent aggregation by encapsulating individual proteins inside a molecular cage is created; in this environment, the polypeptide chain has the opportunity to find its native state without the risk of inappropriate interactions with other molecules. A detailed characterization of the proteins that bind to GroEL/GroES *in vivo* has been performed by lysing *E. coli* cells in the presence of glucose and hexokinase in order to convert cellular ATP to ADP,[14] which stabilizes GroEL/GroES as a complex with its substrate enclosed. Peptide analysis by mass spectrometry enabled the identification of the extent of complexation with GroEL for 250 mainly cytosolic proteins. On the basis of its dependence on GroEL/GroES deduced from *in vitro* and *in vivo* refolding assays,[14] each of these proteins was assigned to one of three classes: (i) class I, in which substrates fold largely independently of GroEL/GroES but may use it to optimize their folding yield; (ii) class II, in which substrates are highly chaperone dependent, at least under mildly unfavourable environmental conditions, but can utilize either DnaK/DnaJ or GroEL/GroES for folding; and (iii) class III, in which substrates have an absolute requirement for the GroEL/GroEL system in order to fold correctly.

In this work, we have explored the origin of these observations through an analysis of the solubility, abundance and chaperone requirements of *E. coli* proteins, which have revealed that these properties are closely linked. We show further that it is possible to predict them with high accuracy from the physicochemical properties of their amino acid sequences.

Results

Relationship between protein aggregation propensity and GroEL requirement

We have recently shown that it is possible to estimate the intrinsic propensity of a protein to aggregate;[21] this propensity is defined in the unfolded state when the regions that contribute to intermolecular association are exposed to the solvent. When proteins are structured, however, it becomes necessary to take into account the fact that at least some of the most aggregation-prone regions will be buried within the native structure.[12,22] Since the conformations that are more likely to interact with GroEL are neither completely unfolded nor in their native states,[23–26] we introduce here an approach for calculating aggregation propensities that takes into account the effects of the formation of intramolecular interactions on the aggregation process (see Eq. (4) in Methods). We applied this method to predict the aggregation propensities of proteins in the set of experimentally identified GroEL substrates described above[14] (Fig. 1). We found that class I proteins have on average substantially lower aggregation propensities than class II and class III proteins. The aggregation propensities of class II and III proteins are found to be similar, consistent with the fact that proteins belonging to these two classes are largely chaperone dependent, while the lower aggregation propensities of class I proteins correspond to their less stringent chaperone requirements.

Comparison of the aggregation propensity profiles of class I and class III proteins revealed that the latter exhibit a larger number of aggregation-prone regions; this result is illustrated in Fig. 1a, considering the cases of a class I protein (enolase, ENO) and a class III protein (*S*-adenosylmethionine synthetase, METK). We have also predicted the aggregation propensities of an extended set of 1158 cytosolic proteins whose GroEL requirements using a method described below (see Section 'Prediction of GroEL requirement'). Analysis of the aggregation propensities of these proteins (Fig. 1b, red bars) confirmed the conclusions just described for the proteins whose classification is known experimentally.

In addition to confirming the existence of a close link between the GroEL requirement and the aggregation propensity of proteins,[14] the approach that we have taken in this study enables insight about the type of structures adopted by GroEL substrates to be obtained. We found that the aggregation propensity estimated from fully unstructured conformations,[21] or from nearly native conformations,[12,22] provides less accurate predictions of the GroEL classes (see Prediction of GroEL requirement), consistent with the idea that the conformations that are most prone to interact with GroEL are partially structured.[23–26]

Relationship between protein solubility and GroEL requirement

By exploiting the link between aggregation rates and mRNA expression levels, we have recently introduced the *CamEL* algorithm to predict the solubility of proteins expressed in *E. coli*.[13] Here, we used *CamEL* to analyse the relationship between protein solubility and GroEL requirements. In agreement with the results reported by Kerner *et al.*[14] for a small number of proteins whose soluble fraction was measured *in vivo*, we found that the number of highly soluble proteins decreases from class I to class III, while the number of poorly soluble

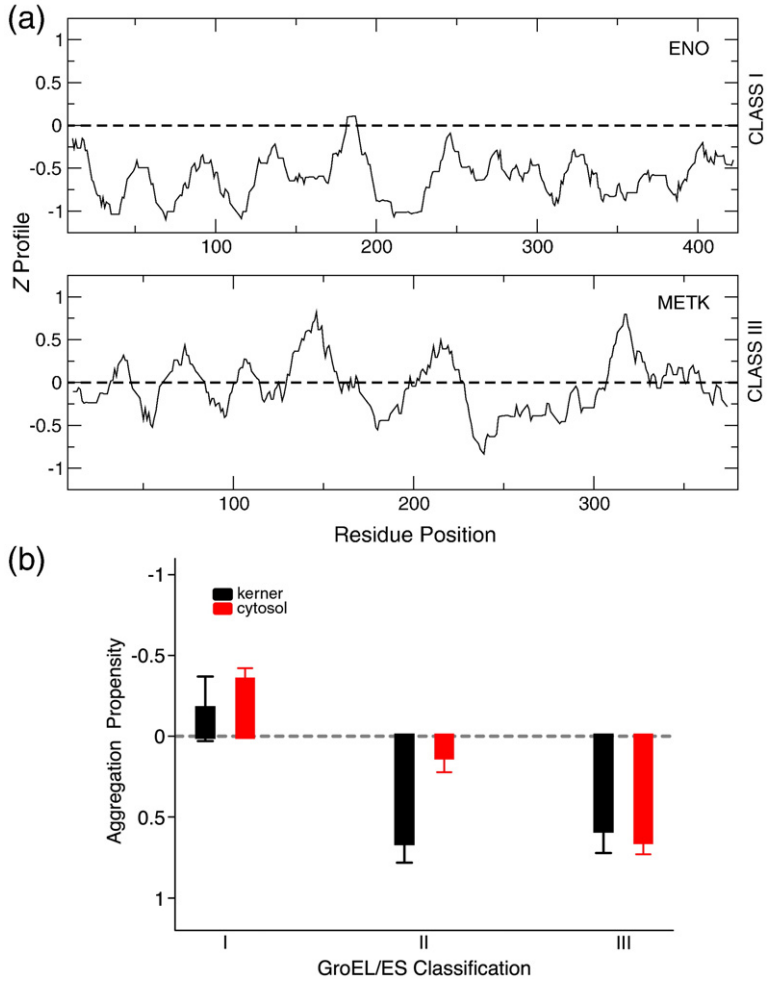

Fig. 1. Relationship between aggregation propensity and GroEL requirement. (a) Comparison of typical aggregation propensity profiles (Z Profile[12]) a class I and a class III proteins; we present the cases of ENO (class I) and METK (class III). The aggregation propensity profile of METK exhibits several peaks above zero, indicating that this protein has a high tendency to aggregate. By contrast, the aggregation propensity profile of ENO presents only one small aggregation peak, indicating a low aggregation propensity. (b) Sequence-based predictions of aggregation propensities (Z scores[12]). In addition to the 250 proteins whose GroEL requirements were originally classified by Kerner *et al.*[14] ('kerner', black bars), we also present the aggregation propensities ('cytosol', red bars) of a set of 1158 cytosolic proteins for which we predicted a GroEL class (690 of class I, 297 of class II and 171 of class III).

proteins increases from class I to class III (Fig. 2a). These trends are confirmed by the properties of the 1158 cytosolic proteins for which we predicted their individual GroEL classification (see Prediction of GroEL requirement) (Fig. 2b).

Relationship between protein abundance and GroEL requirement

The highly crowded environment of a living cell, typically corresponding to about 300 mg/ml of macromolecules, imposes stringent conditions on the properties of the amino acid sequences of proteins that function within the cell.[6] By investigating the nature of these conditions, we have recently found a close relationship between *in vivo* mRNA expression levels and *in vitro* protein aggregation rates[9] for a group of human proteins for which the latter have been measured under near-physiological conditions. We suggested that this relationship is the consequence of balance between the evolutionary pressure acting to decrease the risk of aggregation in the cell, which can give rise to cellular malfunction and disease,[1–5] and the effects of random mutations, which show a marked tendency to destabilize the native folds of proteins and enhance their propensity to aggregate.[28,29]

Since the maintenance of solubility is an essential requirement for homeostasis in living organisms, the presence in the cell of factors capable of responding to the initial stages in the formation of protein aggregates is crucial, especially under conditions that promote aggregation. Indeed, an inverse relationship between GroEL requirement and abundance in the cytosol has been observed,[15] and it has also been reported that the aggregation propensity and the exponentially modified protein abundance index (emPAI) are anticorrelated for cytosolic proteins in *E. coli*.[14]

As mRNA expression levels and protein abundances are known to be correlated in *E. coli*,[30] we investigated the relationship between these quantities and the GroEL requirement of proteins; by simultaneously considering these quantities, we can reduce the intrinsic noise associated with abundance measurements[31,32] and achieve a more accurate description. By using a consensus of mRNA levels[33] and protein abundances,[14] we observed that both mRNA expression levels (Fig. 3a, black bars) and protein abundances (Fig. 3b, black bars) decrease when the GroEL requirements increase. We have also analysed the mRNA expression levels and protein abundances of the 1158 proteins for which we predicted the GroEL classification (see Prediction

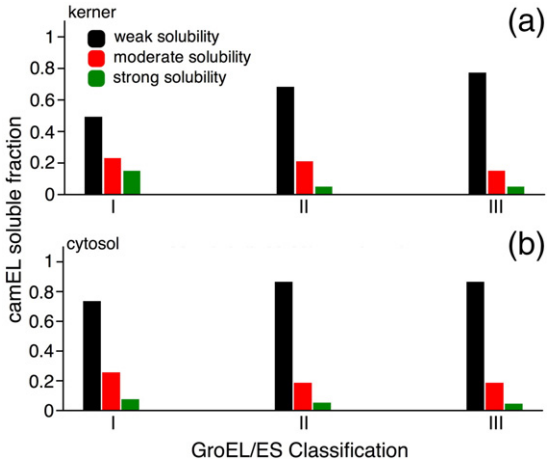

Fig. 2. Relationship between protein solubility and GroEL requirement. (a) We used the *CamEL* algorithm[13] to predict the solubility of the set of 250 *E. coli* proteins whose GroEL requirements were determined experimentally by Kerner *et al.*[14] ('kerner'). We found that the number of proteins of low solubility exhibits a significant increase in going from class I to class III, while the number of proteins with moderate to high solubility shows a significant decrease in going from class I to class III. The three categories of low solubility, moderate solubility and high solubility follow the definition given in a previous paper.[27] (b) Similar trends were found for an extended set of 1158 cytosolic proteins ('cytosol') whose GroEL requirements are predicted in this work.

of GroEL requirement), and we found similar results (Fig. 3a and b, red bars).

Relationship between DnaK and GroEL requirements

The function of the Hsp70 system of *E. coli* (DnaK/DnaJ/GrpE) is to prevent aggregation and to promote the correct folding or refolding of proteins.[20,34,35] It has been suggested that DnaK binding sites generally occur at rather regularly spaced positions in protein sequences and that the binding motif consists of a hydrophobic core of four or five residues enriched in Leu, Ile, Val, Phe and Tyr residues, with two flanking regions enriched in basic residues;[36] on the basis of this analysis, an algorithm for predicting the DnaK binding propensity has been established.[36] In this work, we have considered the effects of the formation of partially structured conformations on the DnaK binding propensities (see Eq. (5) in Methods) by using a procedure similar to that adopted for the aggregation propensities (see Eq. (1) in Methods). By using these DnaK requirement predictions, we have found that class I and class III proteins have low DnaK requirements, while class II proteins have significantly higher DnaK requirements (Fig. 4a and b). These results are consistent with the finding that class II proteins exhibit a strong tendency to use the DnaK system for folding, while class III proteins use DnaK for preventing aggregation but GroEL to fold.[14] Our results (Table S2, "DnaK–GroEL" scale) indicate that these two classes are mainly differentiated by higher hydrophobicity and secondary structure propensity and by lower flexibility and charge of class II proteins with respect to class III proteins. Two examples of DnaK binding propensity profiles are shown in Fig. 4a for ENO (class I) and galactitol-1-phosphate 5-dehydrogenase (GATD, class II). We observed that exposed regions with a strong hydrophobic character have a high tendency to promote protein aggregation.

Prediction of GroEL requirement

We have developed two complementary methods for predicting the GroEL requirement of proteins. The first method, which is presented in this section, is based on the observation that GroEL requirements are closely related to aggregation propensities, DnaK requirements, mRNA expression levels and protein abundances. The second method, which is described in the next section, uses only the knowledge of the amino acid sequences to predict the GroEL requirement of proteins.

The trends that we found for protein aggregation propensities (Fig. 1), mRNA expression levels (Fig. 3a), protein abundances (Fig. 3b) and DnaK requirements (Fig. 4) suggest that these quantities could be combined to predict GroEL requirements. As experimental GroEL requirements are reported in terms of a classification,[14] a support vector

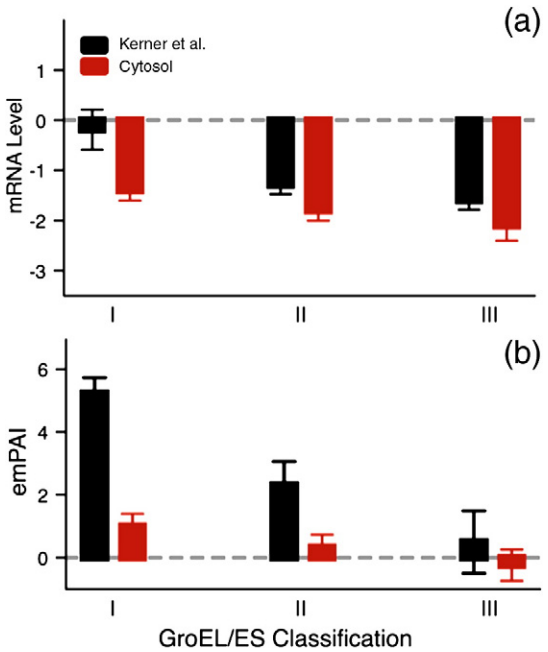

Fig. 3. Relationships between mRNA expression levels, protein abundance (emPAI score) and GroEL requirement. (a) mRNA expression levels refer to the log phase of the cell cycle; values are reported in a log scale, and the mRNA level is calculated as discussed by Selinger *et al.*[33]: Level = 13,000*ln(mRNA Copies) + 39,000. Black bars: Kerner *et al.*[14] ('Kerner *et al.*'), red bars: this work ('cytosol'). (b) Protein abundances are reported on the scale of the emPAI score.[14]

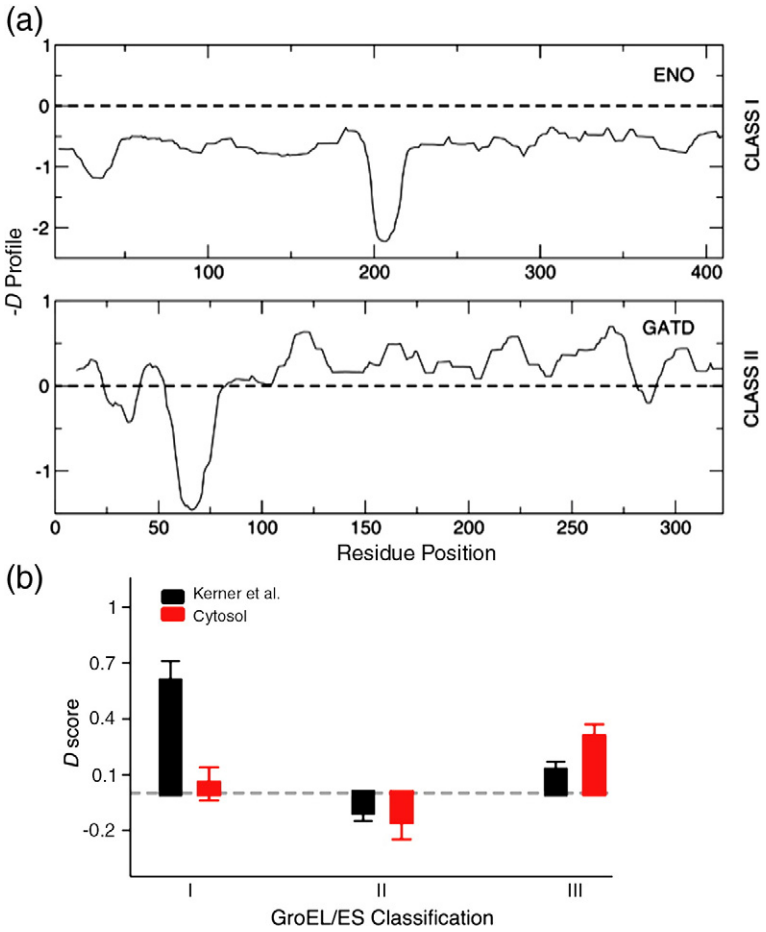

Fig. 4. Relationship between DnaK and GroEL requirements. (a) DnaK requirement profiles (*D* Profile) were calculated using Eq. (5). We provide examples of these profiles for ENO (class I) and GATD (class II). No region with a significant propensity to interact with DnaK was predicted for ENO, in agreement with the fact that this protein does not require DnaK to fold. By contrast, many regions of GATD were found to have a high propensity for interacting with DnaK. The sign of the *D* profile is reversed to facilitate the visualization of the predicted DnaK binding sites in analogy with the Z^{agg} profiles. (b) Average DnaK requirement (*D* score) for the proteins in the three GroEL classes; experimental GroEL classification (black bars, Kerner *et al.*[14], 'Kerner et al.') and predicted GroEL classification (red bars, this work, 'Cytosol') are shown.

machine was used for fitting the data because it provides an efficient algorithm when dealing with discrete data points.[37] We adopted a Gaussian radial-basis function for training the support vector machine and then performed a *leave-one-out* test for the probability of assigning a correct class. This method achieved an 86% accuracy level in assigning proteins to their experimental GroEL class (Fig. 5). It is possible to monitor the error associated with the predictions and thus establish a confidence parameter for testing by modifying the penalty term in the training phase of the method. We found that our classification only slightly changes by increasing the penalty term, indicating that the algorithm is stable and reaches convergence after a few optimization cycles.

We then used this approach to perform a GroEL classification of a set of 1158 cytosolic proteins for which mRNA and protein abundances are known (Table S1). The results indicate that a small fraction of these proteins (14%) is likely to require GroEL for folding under normal growth conditions (class III). We also found that the number of essential proteins is strongly depleted in class III proteins (2%), in agreement with expectations based on the lethality of their misfunction.[14] We observed that proteins characterized by a high β-sheet content are enriched in class III proteins, suggesting a higher propensity to form ordered aggregates (Table S1). We also predicted that the fraction of GroEL independent proteins (class I) is 59%, which is substantially enriched with respect to the percentage of class I proteins in the original classification (16%), which was obtained by using a detection method more sensitive to abundant proteins. Overall, we found that class II and class III comprise a small fraction (297 and 171 proteins, respectively) of them and that class I (690 proteins) contains the majority. In addition to the 85 class III proteins described in the original study by Kerner *et al.*,[14] we identified 86 other ones. There are two possible explanations for these results—the first is that the prediction method should be improved, and the second is that the abundance of the extra 86 class III proteins was too low to enable experimental detection in the study by Kerner *et al.*[14] The latter explanation is supported by the finding that the average abundance of the 86 extra class III proteins is 50-fold lower than that of the 85 original class III proteins.

Sequence-based prediction of GroEL requirements

We introduce in this section a method for predicting the requirement of proteins for GroEL from the physicochemical properties of their amino acid sequences. This method, in addition to providing insights into the physicochemical determinants of the interactions between GroEL and its substrates, is particularly useful in cases for which experimental

information about protein and mRNA concentrations is not available, and therefore the method presented in the previous section is not applicable.

We note first that class III proteins are characterized by an overall higher flexibility, lower hydrophobicity and lower burial than other cytosolic proteins (Fig. 6), suggesting that GroEL substrates tend to populate highly dynamic intermediate states. These results reinforce the expectation that an analysis of the physicochemical properties of the amino acid sequences should enable the prediction of their propensity to interact with GroEL. To perform such sequence-based GroEL requirement predictions, we thus calculated an amino acid propensity scale (Eq. (7) and Table S2) that enables the overall preference of a given sequence to be assigned to one of two classes (see Eq. (10) in Methods). By using this method, we obtain a correct assignment for 97% of class I proteins with respect to class III proteins and a correct assignment for 90% of class I proteins with respect to class II proteins (Table S2). A slightly lower accuracy, 78%, is found for the discrimination between classes II and III, suggesting that the physicochemical properties of the proteins included in these classes tend to be rather similar (Table S2). We found that charge and flexibility give positive contributions that help discriminate class I from class II and class III and that secondary structure is associated with a negative contribution (Table S2). By contrast, charge and β-sheet propensity give positive contributions to discriminate class II from class III, while hydrophobicity and flexibility are associated with negative contributions (Table S2). It is possible to assign correctly the GroEL class with an accuracy of 90% by using a consensus of the three scales.

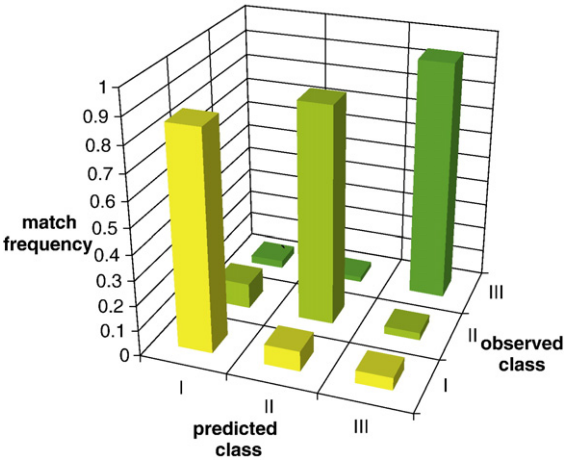

Fig. 5. Prediction of GroEL requirement. The GroEL requirement score is obtained by using a support vector machine approach to combine aggregation propensities, DnaK requirements, mRNA expression levels and protein abundance emPAI scores. We used this method to partition the 250 proteins studied by Kerner *et al.*[14] into the three classes that were originally identified by experiment. We tested the predictive power of the algorithm by using a leave-one-out procedure that indicates an accuracy level ('match frequency') of 86% in assigning the correct class.

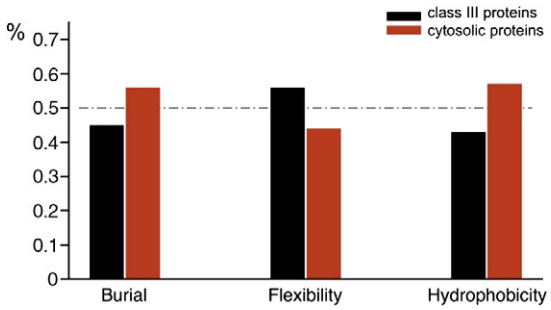

Fig. 6. Characteristic physicochemical properties of class III proteins. Class III proteins exhibit physicochemical properties distinct from those of other cytosolic proteins. Our results indicate that cytosolic proteins have higher degree of hydrophobicity and burial than class III proteins; by contrast, class III proteins have lower flexibility than other cytosolic proteins.

Examples of amino acid GroEL requirement scores (*G* profiles) are shown in Fig. S1. We observed that both the number and the height of the negative peaks increase with the GroEL requirement.

We also considered the efficiency of predictions based on the individual properties included in the sequence-based GroEL requirement predictions (see Eq. (7) in Methods), such as flexibility, burial or hydrophobicity (Fig. 6); the latter is also known to be weakly anticorrelated with chaperone requirements in *Saccharomyces cerevisiae*.[38] Although in these cases we found that predictions were still possible, the accuracy was never higher than 70%, indicating that the particular combination of factors that we employed describes more accurately the behaviour of these proteins.

Discussion and Conclusions

In this study, we have explored the relationships between protein solubility, abundance and chaperone usage. Our results indicate that these relationships impose stringent conditions on the amino acid sequences of proteins (Fig. 7). The existence of a link between protein solubility and protein abundance (arrow A) had already emerged from a study in which the *in vitro* aggregation rates of a set of human proteins were shown to be correlated with the maximal levels of mRNA expression;[9] further evidence for this correlation has also been provided by Ishihama *et al.*[15] Similarly, a link between protein abundance and chaperone requirements (arrow C) has been discussed by Kerner *et al.*[14] In the light of these results, we suggest that a link between protein solubility and chaperone requirements must exist (arrow B), a conclusion that we have investigated in the present study. Our results (Figs. 1 and 2), together with complementary recent studies,[39,40] provide evidence that such a link indeed exists. We have exploited the relationships between these quantities (arrows A, B and C in Fig. 7) to show that GroEL requirements can be predicted with 86%

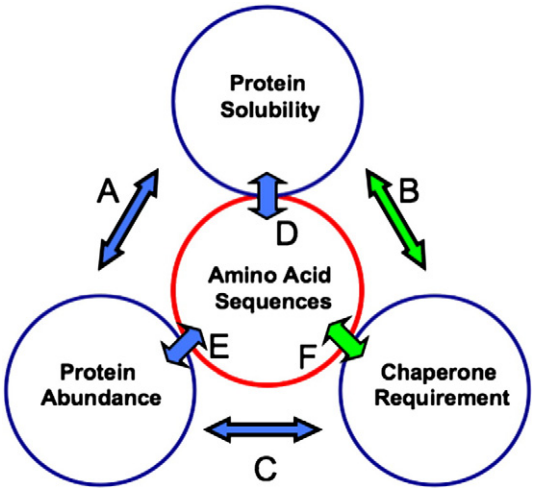

Fig. 7. Schematic relationships between protein solubility, protein abundance and chaperone requirements. The results discussed in this article (green arrows) and those already established in the literature (blue arrows) suggest that all these quantities are closely linked and that they can all be predicted from amino acid sequences. We identify the links by letters: A,[13,15] B[39,40] (see Sections 'Relationship between protein aggregation propensity and GroEL requirement, Relationship between protein solubility and GroEL requirement', and Figs. 1 and 2), C[14] (see Sections 'Relationship between protein abundance and GroEL requirement' and Fig. 3), D[10–12,22] (see Sections 'Relationship between protein aggregation propensity and GroEL requirement'), E[13] and F[38,41,42] (see Sections 'Sequence-based prediction of GroEL requirement', Fig. 5 and Table S1).

accuracy (Fig. 5 and Table S1) from the knowledge of protein solubility and abundance, mRNA expression levels and DnaK requirements.

We have also discussed the relationships between amino acid sequence and protein solubility, abundance and chaperone requirements (arrows D, E and F, respectively, in Fig. 7). Since solubility and abundance can be predicted from the amino acid sequence,[13] we investigated whether GroEL requirements can also be predicted in this manner. Our results indicate that the accuracy in the classification that can be obtained by following this strategy is of 90% (Fig. S1 and Table S2).

The results presented in this study indicate that the behaviour of proteins in the cell is regulated at two levels: The first is a "molecular" level, in which the properties of the amino acid sequences safeguard protein solubility at the concentrations required by the cell for optimal function. The second is the "cellular" level, in which quality control mechanisms are in place to maintain homeostasis, with the chaperone response ensuring that any incipiently misfolded assemblies are prevented from developing further. These two levels of regulation are highly complementary. Since there is essentially no evolutionary pressure on the amino acid sequences to be selected to increase the solubility beyond the concentrations required for their optimal function, cellular quality control mechanisms should be in place to balance the increase in the propensity to aggregate caused by variations in the normal conditions, such as those associated with stress. In essence, therefore, our results indicate that evolution has exploited both the chemical properties of proteins and the properties of their environments in order to optimize their activity in the cell.

Our results therefore suggest that the complex behaviour of proteins in the cell can be predicted with rather high accuracy from their amino acid sequences. With this conclusion, we are not suggesting that the complex processes that regulate the abundance and solubility of proteins in the cell are not important, but rather that amino acid sequences have coevolved with their cellular environments to maintain solubility and ensure homeostasis.

Methods

Sequence-based prediction of protein aggregation propensity

We have introduced a correction to the intrinsic aggregation propensity Z_{0i}^{agg}[21] to take account of the conformational properties of partially structured states of proteins. In order to estimate these properties, we considered the burial b[43] and flexibility f[44] propensities, which are two values that can be predicted from the amino acid sequences. We thus estimated the aggregation propensity of partially structured proteins as:

$$Z_i^{\mathrm{agg}} = w_b T(b_i) + w_f T(f_i) + w_{\mathrm{agg}} T\left(Z_{0i}^{\mathrm{agg}}\right) \qquad (1)$$

In this equation, b_i and f_i are the burial and the flexibility of residue i, respectively, and the function $T(x)$ is the hyperbolic tangent of $\alpha + \beta x$, where α and β are two parameters used to normalize the variable x. The aggregation propensity Z_{0i}^{agg} was defined as:[12,22]

$$Z_{0i}^{\mathrm{agg}} = w_0 + w_{\mathrm{hydr}} I_i^{\mathrm{hydr}} + w_{\mathrm{ss}} I_i^{\mathrm{ss}} + w_{\mathrm{ch}} I_i^{\mathrm{ch}} + w_{\mathrm{pat}} I_i^{\mathrm{pat}} + w_{\mathrm{gk}} I_i^{\mathrm{gk}} \qquad (2)$$

where I_i^{hydr} is the hydrophobicity of residue i, I_i^{ss} is its secondary structure propensity, I_i^{ch} and I_i^{gk} are functions for charged amino acids and I_i^{pat} takes into account the effects of patterns of polar and nonpolar residues; w_A, w_B, w_F, w_H, w_R and w_P are the relative coefficients.

The parameters w_b and w_f were determined by minimizing the function $\sum_i |w_p T(p_i) - w_b T(b_i) - w_f T(f_i)|$ on a data set of proteins used in a previous study,[12] where p_i represents the amplitude of the native fluctuations[45] and the parameter w_P represents the corresponding weight, which was determined by fitting

$$Z^{\mathrm{agg}}(p) = (\tilde{w}_l l + c)^{-1} \sum_i \left[w_p T(p_i) + \tilde{w}_{\mathrm{agg}} T\left(Z_{0i}^{\mathrm{agg}}\right) \right] \qquad (3)$$

to the experimental aggregation rates for the same set of proteins.[12]

We the define the overall aggregation propensity of an amino acid sequence as:

$$Z = (w_l l + c)^{-1} \sum_i Z_i^{agg} \quad (4)$$

where l is the protein length and $(w_l l + c)^{-1}$ is a normalization factor.

Sequence-based prediction of DnaK requirement

In order to carry out sequence-based predictions of the DnaK requirement of proteins, we used a modified version of the method of Rudiger *et al.*,[36] which provides the propensity D_i^0 of residue i to interact with DnaK. Since DnaK binding sites are likely to be inaccessible when a protein is folded, we followed the procedure introduced above to enable the estimation of the aggregation propensity of partially folded proteins and use the burial and flexibility propensities to modify the D_{0i} profiles:

$$D_i = w_b T(b_i) + w_f T(f_i) + w_D T(D_{0i}) \quad (5)$$

where b_i and f_i are the burial and the flexibility propensities, respectively. We then define the overall D score for the DnaK requirement by summing over the entire sequence,

$$D = (w_l l + c)^{-1} \sum_i D_i \quad (6)$$

where l is the protein length and $(w_l l + c)^{-1}$ is a normalization factor.

Sequence-based prediction of GroEL requirement

As a first step in establishing the method, we define a GroEL requirement score for an amino acid i as,

$$G_i^0 = w_A A_i + w_B B_i + w_F F_i + w_H H_i + w_R R_i + w_P P_i \quad (7)$$

The variables A, B, F, H, R and P account for α-helical propensity,[46] β-sheet propensity,[47] flexibility,[44] hydrophobicity,[48] aromatic clustering[49] and polar/nonpolar patterns, respectively;[50] w_A, w_B, w_F, w_H, w_R and w_P are the relative coefficients.

G_i profiles are then calculated from the G_i^0 scores averaging over a sliding window of 15 amino acids that moves from the N-terminus to the C-terminus. In Fig. 7, we report examples of G_i profiles for six proteins studied by Kerner *et al.*[14]

We define an overall GroEL requirement score (G score) by summing over the individual amino acid propensities and scaling with a function of the negative peaks and of the protein length,

$$G = w \sum_i G_i T(G_i) \quad (8)$$

The normalization function w is defined as:

$$\log(w) = \alpha \log(l) + \beta \log \sum_i T(G_i) \quad (9)$$

where l is the protein length.

In order to determine parameters w_A, w_B, w_F, w_H, w_R and w_P in Eq. (7) capable of discriminating between two GroEL requirement classes, we define a relative fitness function,

$$F(X,Y) = \sum_{a \in X} \sum_{b \in Y} \theta[-G(a) + G(b)] \quad (10)$$

where the indices a and b run on the proteins of GroEL classes X and Y, respectively, and the step function $\vartheta(x)$ is defined as 1 for $x \geq 0$ and 0 for $x < 0$. For each pair of classes I/II, II/III and I/III, we used a Monte Carlo approach to estimate the parameters w_A, w_B, w_F, w_H, w_R and w_P that maximize the relative fitness function $F(X,Y)$—i.e., that separate the two classes more efficiently. We used a *leave-one-out* procedure to estimate the predictive power of the method.

A web server is available to use this method†.

Acknowledgements

This work was supported by the Royal Society, the Leverhulme Trust, the Wellcome Trust, MRC and BBSRC.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2010.03.066

References

1. Chiti, F. & Dobson, C. M. (2006). Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.* **75**, 333–366.
2. Haass, C. & Selkoe, D. J. (2007). Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid beta-peptide. *Nat. Rev. Mol. Cell Biol.* **8**, 101–112.
3. Lesne, S., Koh, M. T., Kotilinek, L., Kayed, R., Glabe, C. G., Yang, A. *et al.* (2006). A specific amyloid-beta protein assembly in the brain impairs memory. *Nature*, **440**, 352–357.
4. Mattson, M. P. (2004). Pathways towards and away from Alzheimer's disease. *Nature*, **430**, 631–639.
5. Silveira, J. R., Raymond, G. J., Hughson, A. G., Race, R. E., Sim, V. L., Hayes, S. F. & Caughey, B. (2005). The most infectious prion protein particles. *Nature*, **437**, 257–261.

† http://www-vendruscolo.ch.cam.ac.uk/camGroEL.php

6. Dobson, C. M. (2003). Protein folding and misfolding. *Nature*, **426**, 884–890.
7. Ellis, R. J. (2001). Macromolecular crowding: obvious but underappreciated. *Trends Biochem. Sci.* **26**, 597–604.
8. Powers, E. T., Morimoto, R. I., Dillin, A., Kelly, J. W. & Balch, W. E. (2009). Biological and chemical approaches to diseases of proteostasis deficiency. *Annu. Rev. Biochem.* **78**, 959–991.
9. Tartaglia, G. G., Pechmann, S., Dobson, C. M. & Vendruscolo, M. (2007). Life on the edge: a link between gene expression levels and aggregation rates of human proteins. *Trends Biochem. Sci.* **32**, 204–206.
10. Chiti, F., Stefani, M., Taddei, N., Ramponi, G. & Dobson, C. M. (2003). Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature*, **424**, 805–808.
11. Fernandez-Escamilla, A. M., Rousseau, F., Schymkowitz, J. & Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* **22**, 1302–1306.
12. Tartaglia, G. G., Pawar, A. P., Campioni, S., Dobson, C. M., Chiti, F. & Vendruscolo, M. (2008). Prediction of aggregation-prone regions in structured proteins. *J. Mol. Biol.* **380**, 425–436.
13. Tartaglia, G. G., Pechmann, S., Dobson, C. M. & Vendruscolo, M. (2009). A relationship between mRNA expression levels and protein solubility in *E. coli*. *J. Mol. Biol.* **388**, 381–389.
14. Kerner, M. J., Naylor, D. J., Ishihama, Y., Maier, T., Chang, H. C., Stines, A. P. *et al.* (2005). Proteome-wide analysis of chaperonin-dependent protein folding in *Escherichia coli*. *Cell*, **122**, 209–220.
15. Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., Hartl, F. U., Kerner, M. J. & Frishman, D. (2008). Protein abundance profiling of the *Escherichia coli* cytosol. *BMC Genomics*, **9**.
16. Frydman, J. (2001). Folding of newly translated proteins *in vivo*: the role of molecular chaperones. *Annu. Rev. Biochem.* **70**, 603–647.
17. Horwich, A. L., Fenton, W. A., Chapman, E. & Farr, G. W. (2007). Two families of chaperonin: physiology and mechanism. *Annu. Rev. Cell Dev. Biol.* **23**, 115–145.
18. Hartl, F. U. & Hayer-Hartl, M. (2009). Converging concepts of protein folding *in vitro* and *in vivo*. *Nat. Struct. Mol. Biol.* **16**, 574–581.
19. Kramer, G., Boehringer, D., Ban, N. & Bukau, B. (2009). The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. *Nat. Struct. Mol. Biol.* **16**, 589–597.
20. Hartl, F. U. & Hayer-Hartl, M. (2002). Protein folding—molecular chaperones in the cytosol: from nascent chain to folded protein. *Science*, **295**, 1852–1858.
21. Dubay, K. F., Pawar, A. P., Chiti, F., Zurdo, J., Dobson, C. M. & Vendruscolo, M. (2004). Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. *J. Mol. Biol.* **341**, 1317–1326.
22. Tartaglia, G. G. & Vendruscolo, M. (2008). The Zyggregator method for predicting protein aggregation propensities. *Chem. Soc. Rev.* **37**, 1395–1401.
23. Chatellier, J., Buckle, A. M. & Fersht, A. R. (1999). GroEL recognises sequential and non-sequential linear structural motifs compatible with extended beta-strands and alpha-helices. *J. Mol. Biol.* **292**, 163–172.
24. Horst, R., Bertelsen, E. B., Fiaux, J., Wider, G., Horwich, A. L. & Wuthrich, K. (2005). Direct NMR observation of a substrate protein bound to the chaperonin GroEL. *Proc. Natl Acad. Sci. USA*, **102**, 12748–12753.
25. Stan, G., Brooks, B. R., Lorimer, G. H. & Thirumalai, D. (2006). Residues in substrate proteins that interact witn GroEL in the capture process are buried in the native state. *Proc. Natl Acad. Sci. USA*, **103**, 4433–4438.
26. Sharma, S., Chakraborty, K., Mueller, B. K., Astola, N., Tang, Y. C., Lamb, D. C. *et al.* (2008). Monitoring protein conformation along the pathway of chaperonin-assisted folding. *Cell*, **133**, 142–153.
27. Bussow, K., Quedenau, C., Sievert, V., Tischer, J., Scheich, C., Seitz, H. *et al.* (2004). A catalog of human cDNA expression clones and its application to structural genomics. *Genome Biol.* **5**.
28. DePristo, M. A., Weinreich, D. M. & Hartl, D. L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet.* **6**, 678–687.
29. Kumar, M. D. S., Bava, K. A., Gromiha, M. M., Prabakaran, P., Kitajima, K., Uedaira, H. & Sarai, A. (2006). ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions. *Nucleic Acids Res.* **34**, D204–D206.
30. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E. M. (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* **25**, 117–124.
31. Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y. & Barkai, N. (2006). Noise in protein expression scales with natural protein abundance. *Nat. Genet.* **38**, 636–643.
32. Raj, A. & van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, **135**, 216–226.
33. Selinger, D. W., Cheung, K. J., Mei, R., Johansson, E. M., Richmond, C. S., Blattner, F. R. *et al.* (2000). RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nat. Biotechnol.* **18**, 1262–1268.
34. Glover, J. R. & Lindquist, S. (1998). Hsp104, Hsp70, and Hsp40: a novel chaperone system that rescues previously aggregated proteins. *Cell*, **94**, 73–82.
35. Mogk, A., Tomoyasu, T., Goloubinoff, P., Rudiger, S., Roder, D., Langen, H. & Bukau, B. (1999). Identification of thermolabile *Escherichia coli* proteins: prevention and reversion of aggregation by DnaK and ClpB. *EMBO J.* **18**, 6934–6949.
36. Rudiger, S., Germeroth, L., SchneiderMergener, J. & Bukau, B. (1997). Substrate specificity of the DnaK chaperone determined by screening cellulose-bound peptide libraries. *EMBO J.* **16**, 1501–1507.
37. Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* **20**, 273–297.
38. Gong, Y. C., Kakihara, Y., Krogan, N., Greenblatt, J., Emili, A., Zhang, Z. & Houry, W. A. (2009). An atlas of chaperone–protein interactions in *Saccharomyces cerevisiae*: implications to protein folding pathways in the cell. *Mol. Syst. Biol.* **5**.
39. Niwa, T., Ying, B. W., Saito, K., Jin, W., Takada, S., Ueda, T. & Taguchi, H. (2009). Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins. *Proc. Natl Acad. Sci. USA*, **106**, 4201–4206.
40. Noivirt-Brik, O., Unger, R. & Horovitz, A. (2007). Low folding propensity and high translation efficiency distinguish *in vivo* substrates of GroEL from other *Escherichia coli* proteins. *Bioinformatics*, **23**, 3276–3279.

41. Yam, A. Y., Xia, Y., Lin, H. T. J., Burlingame, A., Gerstein, M. & Frydman, J. (2008). Defining the TRiC/CCT interactome links chaperonin function to stabilization of newly made proteins with complex topologies. *Nat. Struct. Mol. Biol.* **15**, 1255–1262.
42. Rousseau, F., Serrano, L. & Schymkowitz, J. W. H. (2006). How evolutionary pressure against protein aggregation shaped chaperone specificity. *J. Mol. Biol.* **355**, 1037–1047.
43. Janin, J. (1979). Surface and inside volumes in globular proteins. *Nature*, **277**, 491–492.
44. Chou, P. Y. & Fasman, G. D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.* **47**, 45–148.
45. Best, R. B. & Vendruscolo, M. (2006). Structural interpretation of hydrogen exchange protection factors in proteins: characterization of the native state fluctuations of CI2. *Structure*, **14**, 97–106.
46. Pace, C. N. & Scholtz, J. M. (1998). A helix propensity scale based on experimental studies of peptides and proteins. *Biophys. J.* **75**, 422–427.
47. de la Paz, M. L. & Serrano, L. (2004). Sequence determinants of amyloid fibril formation. *Proc. Natl Acad. Sci. USA*, **101**, 87–92.
48. Meek, J. L. (1980). Prediction of peptide retention times in high-pressure liquid-chromatography on the basis of amino-acid composition. *Proc. Natl Acad. Sci. USA*, **77**, 1632–1636.
49. Gazit, E. (2002). A possible role for pi-stacking in the self-assembly of amyloid fibrils. *FASEB J.* **16**, 77–83.
50. Xiong, H. Y., Buckwalter, B. L., Shieh, H. M. & Hecht, M. H. (1995). Periodicity of polar and nonpolar amino-acids is the major determinant of secondary structure in self-assembling oligomeric peptides. *Proc. Natl Acad. Sci. USA*, **92**, 6349–6353.