doi:10.1016/j.jmb.2009.03.002 *J. Mol. Biol.* (2009) **388**, 381–389

JMB
Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER

A Relationship between mRNA Expression Levels and Protein Solubility in *E. coli*

Gian Gaetano Tartaglia*, Sebastian Pechmann, Christopher M. Dobson and Michele Vendruscolo*

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

Received 25 July 2008; received in revised form 26 February 2009; accepted 3 March 2009
Available online
10 March 2009

Edited by F. Schmid

Each step in the process of gene expression, from the transcription of DNA into mRNA to the folding and posttranslational modification of proteins, is regulated by complex cellular mechanisms. At the same time, stringent conditions on the physicochemical properties of proteins, and hence on the nature of their amino acids, are imposed by the need to avoid aggregation at the concentrations required for optimal cellular function. A relationship is therefore expected to exist between mRNA expression levels and protein solubility in the cell. By investigating such a relationship, we formulate a method that enables the prediction of the maximal levels of mRNA expression in *Escherichia coli* with an accuracy of 83% and of the solubility of recombinant human proteins expressed in *E. coli* with an accuracy of 86%.

Keywords: protein aggregation; protein solubility; protein concentration; mRNA expression levels; recombinant proteins

Introduction

The conversion of 04214 the information stored in DNA into proteins takes place through a series of steps that are highly regulated in response to the functional requirements of the cell.[1–3] One of these requirements is that proteins, once expressed, must remain soluble and avoid misfolding and aggregation in order to function effectively and to avoid cellular damage.[4] In addition to quality-control mechanisms that act at the cellular level, such as the unfolded protein response[5,6] and the heat shock response,[7,8] increasing evidence suggests that the protein sequences themselves have evolved to reduce their intrinsic propensity to aggregate.[9–12] Indeed, as a result of the evolutionary pressure to avoid aggregation, *in vivo* mRNA expression levels and *in vitro* protein aggregation rates are strongly anti-correlated.[13]

In this work, we explore further the link between mRNA expression levels and protein aggregation behaviour and show that an amino acid scale developed for characterizing the propensity of proteins to aggregate can be used to make predictions about the maximal levels of mRNA expression in *Escherichia coli*. To complement this result, we also show that aggregation propensities can be used to predict the solubility of recombinant human proteins expressed in *E. coli*, a result of considerable value in biotechnology.

Further, since it has been established that aggregation rates of proteins can be predicted using the physicochemical properties encoded in their sequences,[14–17] we investigate the extent to which such properties can be used to predict mRNA expression levels in *E. coli*. We thus introduce the CamEL (*Cam*bridge Predictor of *E*xpression *L*evels) method that enables the prediction of the maximum levels of mRNA expression in *E. coli* using the information extracted from hydropathy scales, secondary-structure propensities and co-translational factors. We then show that the CamEL approach can be used also to predict the solubility of human proteins expressed in *E. coli* with a high degree of accuracy.

Taken together, our results provide an illustration of the link between the maximal levels of mRNA expression and the solubility of the corresponding proteins in *E. coli*.

Results

Relationship between protein aggregation propensities and mRNA expression levels

The average number of mRNA molecules per cell during the log phase of bacterial growth varies

**Corresponding authors.* E-mail addresses: ggt23@cam.ac.uk; mv245@cam.ac.uk.

Abbreviation used: CamEL, *Cam*bridge Predictor of *E*xpression *L*evels.

between 10^{-1} and 10^5 during the log phase of bacterial growth[18] (Fig. 1a). This high variability arises from the different functional requirements of proteins and is tightly regulated at the cis regulatory[19] and posttranslational levels.[19,20] There is a strong bias to maintain the solubility of proteins in the cell because of the potential toxicity of misfolded and aggregated assemblies.[4,13] However, decreased solubility is the result of random mutations,[21] defective posttranslational modifications such as phosphorylation[22] and glycosylation[23] or inefficient interaction with molecular chaperones such as DnaJ and DnaK as well as GroEL and GroES.[7] As aggregation rates of proteins can be predicted using the physicochemical properties of their amino acid sequences,[14,16,17] we investigate here their relationship with expression levels on a proteomic scale.

We use a database of cytosolic proteins in the *E. coli* proteome because their abundance enables very accurate statistics to be obtained.[24] Moreover, in the highly regulated environment of the cytosol, variations of pH or ionic strength are not significant and their effects on aggregation are negligible.[25,26] We then divide protein sequences into fragments and calculate the aggregation propensity of each fragment using a scale that has been recently described.[16] In agreement with other experimental scales,[12,27] high aggregation propensities are associated with hydrophobic amino acids such as I, F, V, L, W, Y, and M and low aggregation propensities are associated with polar amino acids such as D, E, N, R, Q, H and K. As we move from protein expressed at low levels to those expressed at high levels, we observe that the number of hydrophobic fragments (agg_+) is progressively depleted in comparison with the number of polar fragments (agg_-), and as a consequence, the aggregation propensities tend to decrease (Fig. 1b).

Having made this observation, we provide an illustration of the relationship between gene expression and protein aggregation by discussing a method for discriminating between high and low mRNA expression levels using the aggregation propensities of proteins. For this purpose, we exploit the information contained in the aggregation propensities of 50 sequences associated with the highest expression levels and 50 sequences associated with the lowest expression levels in our database (Supplementary Information).[18] As the difference between the aggregation propensities of these two groups is large (Fig. 1b), we have an ideal data set for deriving the parameters required for such a method to succeed (Materials and Methods). We use the distribution of agg_+ and agg_- aggregation propensities (Fig. 1c) to fit the parameters needed to successfully discriminate between sequences associated with low and high expression levels (Materials and Methods). Using a *leave-one-out* procedure, we establish that 69% of the entries in our data set can be correctly allocated to one of the two classes (Fig. 1d). Our results indicate that low and high mRNA expression levels can be predicted rather accurately by using only the aggregation propensities of the corresponding proteins. We have shown here in quantitative terms that hydrophilic proteins are particularly associated with high

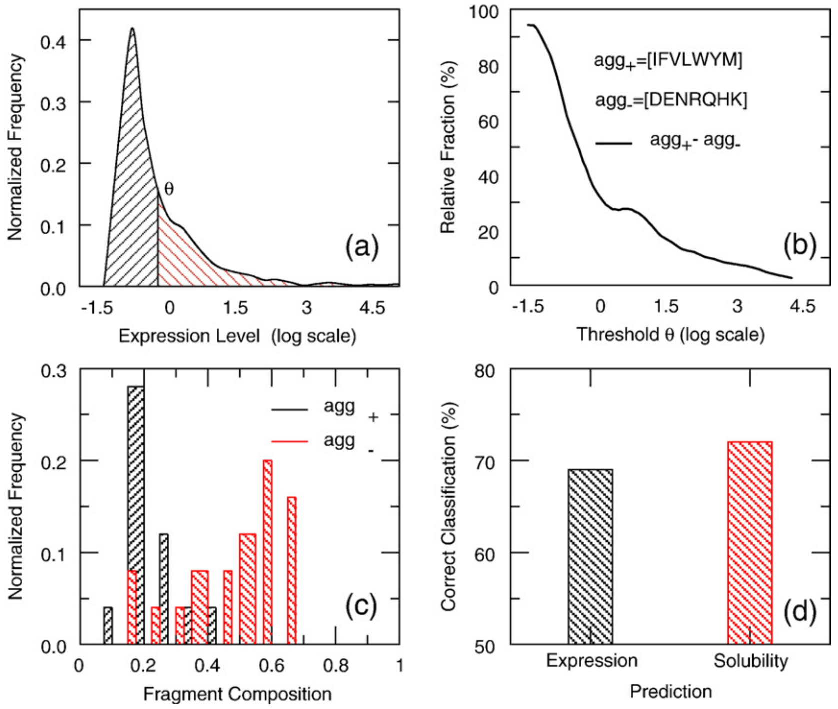

Fig. 1. Relationship between mRNA expression levels and protein aggregation propensities. (a) Distribution of mRNA expression levels of *E. coli* proteins during the log phase of bacterial growth. (b) Analysis of the aggregation propensities of protein sequences associated with mRNA expression levels above a specific threshold θ [see (a)]. For each protein, we compute the number of protein fragments characterised by high (agg_+) and low (agg_-) aggregation propensities. By moving the threshold θ from low to high expression levels, we observe that the number of low aggregation propensity fragments agg_- is enriched with respect to the number of high aggregation propensity fragments agg_+. (c) The distribution of agg_+ and agg_- fragments is shown for the 50S ribosomal protein L35 (SwissProt entry RL35_ECOLI); this gene is highly expressed (3270 mRNA copies per cell) and its corresponding protein is characterised by low hydrophobicity and high polarity. (d) The information contained in the amino acid composition of the fragments is used to formulate a method (Materials and Methods) that discriminates between high and low expression levels. The parameters are derived using a set of 100 cytosolic proteins and the method assigns the correct class with an accuracy of 69%. The same method has an accuracy of 72% in predicting the soluble fraction of human proteins expressed in *E. coli*. A leave-one-out procedure was used to establish the performance of the method (Supplementary Information).

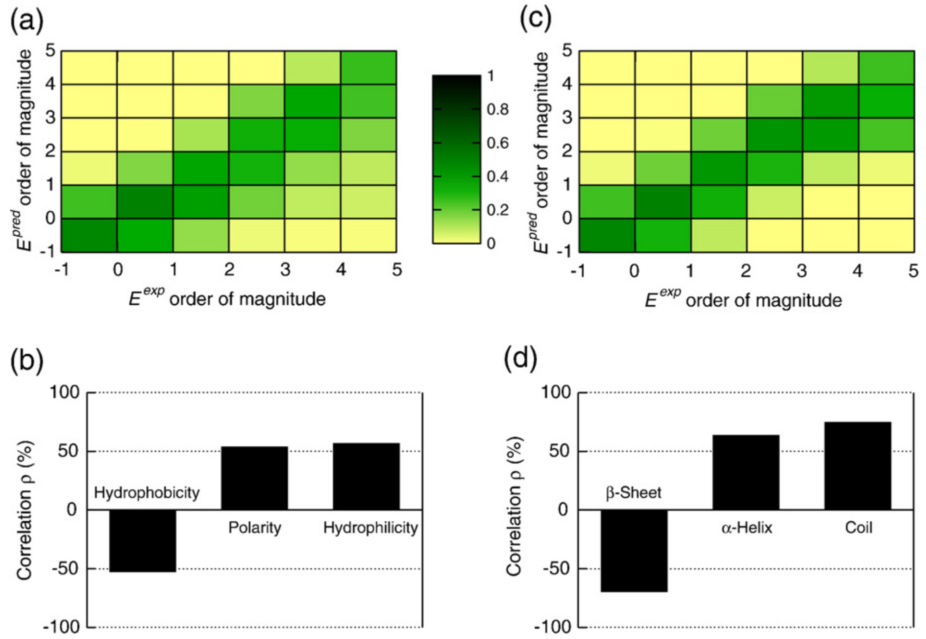

Fig. 2. Prediction of expression levels using physicochemical properties of protein sequences. The average number of mRNA molecules per cell ranges from 10^{-1} to 10^5, and mRNA expression levels are partitioned into six classes according to their order of magnitude. Shades of gray are used to indicate the frequency for the classification of experimental *versus* predicted expression levels. (a) Hydrophobicity, polarity and hydrophilicity of protein sequences are used to predict the expression levels. 58% of the predicted expression levels E^{pred} are observed to be within one order of magnitude of the experimental expression levels (E^{exp}) and 77% within two orders of magnitude. (b) Polarity and hydrophilicity show individual correlations of 54% and 57% with expression levels, respectively, while hydrophobicity shows an anti-correlation of −54% (Materials and Methods). (c) Prediction of mRNA expression levels using α-helix, β-sheet and random coil propensities. 66% of the expression levels are predicted within one order of magnitude of the experimental expression levels (E^{exp}) and 88% within two orders of magnitude. (d) α-Helical and random show individual correlations of 64% and 75% with expression levels, respectively, while β-sheet propensity shows an anti-correlation of −74% (Materials and Methods).

expression levels, a result in agreement with previous observations.[28]

To assess further the validity of this approach, we here used it to predict the soluble fractions of recombinant human proteins (clones) in *E. coli*. This case is of particular interest because a lack of solubility represents a major bottleneck in biotechnology, including functional and structural genomics projects.[29,30] For example, low success rates have been reported for the expression of eukaryotic proteins in *E. coli* and indeed only a small proportion of proteins can be successfully prepared and purified in this way.[31] From the hEx1 expression library of human proteins in *E. coli*,[32] we selected all 72 proteins with the lowest solubility and all 112 proteins with the highest solubility (Materials and Methods). We then classified these human proteins by following the procedure introduced above for distinguishing between sequences encoding high and low expression levels. By using a leave-one-out validation procedure (see Supplementary Information), we found that it is possible to discriminate between high solubility and low solubility clones with an accuracy of 72% (Fig. 1d). These findings indicate that the same physicochemical principles can be used to distinguish between high and low mRNA expression levels and between high and low solubilities of proteins in *E. coli*.

Analysis of physicochemical determinants of mRNA expression levels

In the previous section, we have described a method to discriminate between high and low expression levels of genes using the aggregation propensities of proteins. Using a database of cytosolic proteins,[18,24] we now investigate the relationship between mRNA expression levels and a variety of physicochemical properties of proteins. We formulate the CamEL method†, which enables prediction of the maximal levels of mRNA expression in *E. coli*. We focus on the order of magnitude of the prediction rather than on the specific values of the expression levels as an intrinsic error is associated with microarray data[33] and because of the stochastic nature of gene expression process itself.[34] As in the previous section, we divide protein sequences into fragments and then calculate their physicochemical properties (Supplementary Information).

We first consider a combination of physicochemical properties of the amino acid sequences, including hydrophobicity, polarity and hydrophilicity, to predict the order of magnitude of the expression levels (Materials and Methods). We find that 58% of

† http://www-vendruscolo.ch.cam.ac.uk/camel.html

the mRNA expression levels can be predicted within one order of magnitude and 84% within two orders of magnitude of the experimental values (Fig. 2a). We also observe that underestimation is more frequent that overestimation, because, in our database, the number of sequences associated with low expression levels exceeds the number associated with high expression levels; this slight asymmetry is reflected in the parametrisation of the CamEL method and, hence, in its performance. An increase in expression level is associated with an increase in polarity and hydrophilicity, while an increase in hydrophobicity correlates with a decrease in the predicted expression level (Fig. 2b and Materials and Methods).

We also find that by using a combination of α-helical, β-sheet and random coil propensities, it is possible to predict 66% of the expression levels within one order of magnitude and 88% within two orders of magnitude from the experimental values (Fig. 2c). When such propensities are considered individually, β-sheet contributions show an anti-correlation with expression levels, whereas α-helical and coil propensities have a positive correlation (Fig. 2d). We observe that most of the regions characterised by high β-sheet propensities are also aggregation prone and that an increase in polarity is linked to an increase in random coil and α-helical propensities (Supplementary Information). Overall, we find that high levels of gene expression are associated with low values of the aggregation propensity in the corresponding protein sequence.

These results show that hydropathy scales and secondary-structure propensities of amino acid sequences are quantitatively linked with the expression levels of the corresponding genes. By combining these two parameters, we achieved a correlation of 75% in predicting the mRNA expression levels within one order of magnitude of the experimental values. As the physicochemical properties considered here are also those that are important for predicting aggregation rates,[14,35,36] the observed correlations can be attributed to the relationship that has already been established between mRNA expression levels and protein aggregation rates.[13]

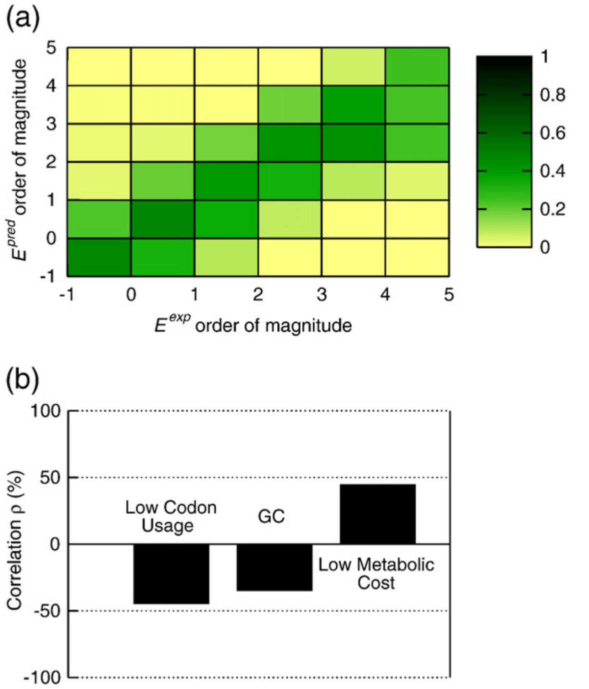

Fig. 3. Prediction of mRNA expression levels using data on GC content, low codon usage and metabolic costs. Shades of gray are used to indicate the frequencies of the classification of experimental *versus* predicted expression levels. (a) The calculated expression levels E^{pred} are predicted in 61% of cases within one order of magnitude and in 70% of cases within two orders of magnitude from the experimental expression levels E^{exp}. (b) Low codon usage bias and GC content show an anti-correlation of −35% and −45% with expression levels, respectively. Expression levels and low metabolic cost show a positive correlation of 45% (Materials and Methods). We use a database of 2800 cytosolic proteins.[24]

Translation-related determinants of mRNA expression levels

Newly synthesised polypeptide chains are particularly vulnerable to misfolding and aggregation, indicating that co-translational strategies to avoid these events must have evolved within the cell to assist the folding process.[7] A series of remarkable correlations has recently been reported for genomic sequence evolution, codon usage and mRNA levels.[20,36–41] According to the “translation robustness” hypothesis, highly expressed proteins are subjected to a pressure to fold despite transcriptional errors that lead to the accumulation of toxic species.[37,38] In agreement with this hypothesis, it has been observed that translational pausing at rare codons might induce a time delay that enables independent and sequential folding of defined portions of the nascent polypeptide chain emerging from the ribosome.[20,39,40] In analogy with punctuation marks in written language, specific codon pairs appear to dictate the timing of protein expression by slowing the translation with induced pauses.[41,42]

To take into account the translation-related contributions in the prediction of expression levels, we adapted our approach to perform predictions based on GC content, codon usage and metabolic costs (Supplementary Information). We observe that 61% of the predicted expression levels are within one order of magnitude and 85% are within two orders of magnitude of the experimental values (Fig. 3a). An interesting observation that emerges from these findings is that the expression levels increase as (a) the GC content decreases, (b) the codon usage of the sequence is optimised and (c) the metabolic costs are reduced (Fig. 3b). Intriguingly, the GC content is significantly higher for hydrophobic amino acids,[43,44] while the metabolic costs are significantly lower for hydrophilic amino acids.[45] These data suggest that an evolutionary pressure acts to avoid aggregation (Supplementary Information). We also observe that polar amino acids are more frequent in loops and tend

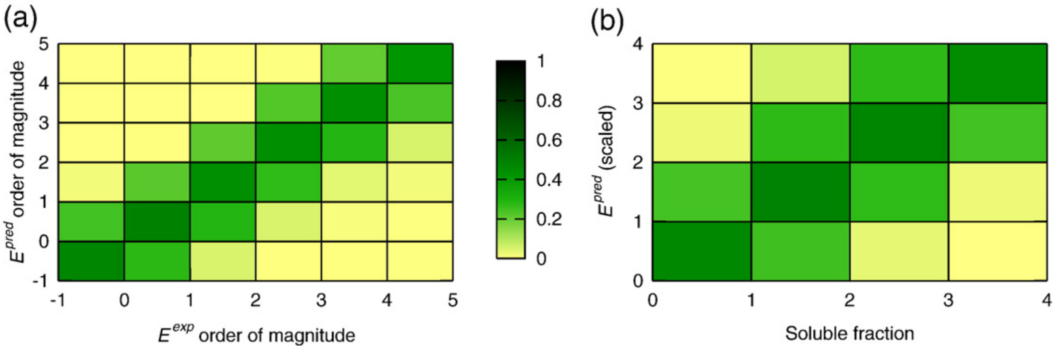

Fig. 4. Prediction of mRNA expression levels and soluble fractions of recombinant human proteins in *E. coli* by analysing the physicochemical properties of proteins and translation factors. (a) Expression of *E. coli* genes. *E. coli* mRNA expression levels are partitioned into six classes according to the order of magnitude of the molecules expressed. We predict 83% of the expression levels within one order of magnitude and 92% within two orders of magnitude from the experimental values. (b) Expression of recombinant human proteins in *E. coli*. The predicted expression levels are partitioned into four classes according to the soluble fraction score and plotted against the experimentally measured soluble fraction of recombinant human proteins. We predict the soluble fraction of recombinant proteins with an accuracy of 86% for a total of 746 proteins. Shades of gray are used to indicate the classification frequencies.

to be encoded by low-usage codons, which suggests that flexible regions and structured regions are subjected to different evolutionary pressures, as suggested by Thanaraj and Argos[20] (Supplementary Information).

Prediction of mRNA expression levels in *E. coli*

In the previous sections, we used individual physicochemical properties of amino acid sequences, such as hydropathy, secondary-structure propensities and translation factors, to predict the maximal levels of mRNA expression in *E. coli*. Using a support vector machine to combine these factors (Materials and Methods and Supplementary Information), we generated a version of the CamEL method capable of predicting 83% of expression levels within one order of magnitude and 92% within two orders of magnitude of the experimental values (Fig. 4a).

Prediction of the solubility of recombinant human proteins in *E. coli*

Since our initial aim is to investigate the relationship between mRNA expression levels and the solubility of the corresponding proteins, we apply the CamEL method to predict the soluble fraction of human proteins expressed in *E. coli*. The soluble fraction of a protein is defined here as an integer that ranges from 0 (no expression) to 3 (high soluble expression).[32] Hence, to predict this variable, we partition the output into four classes rather than into six (Materials and Methods). We find that the soluble fraction can be correctly predicted in 86% of cases for a total of 746 human proteins. These results suggest that the same physicochemical principles can be used to predict both the maximal levels of mRNA expression levels of *E. coli* proteins and the soluble fraction of human proteins expressed heterologously in *E. coli*.

Discussion and Conclusions

In this work, we have shown that it is possible by using the CamEL method to achieve an accuracy of 83% in predicting the order of magnitude of mRNA expression levels in *E. coli* by considering the physicochemical properties of the amino acid sequences of the corresponding proteins, and an accuracy of 86% in the prediction of the solubilities of recombinant human proteins in *E. coli*. Thus, although the CamEL method has been parametrised for predicting mRNA expression levels in *E. coli*, it can also be used to predict the solubility of recombinant proteins when the organism is employed as an expression system. Indeed, our findings complement and extend those recently reported for the predictions of yeast gene expression levels with an accuracy of 70% using the amino acid compositions of di- and tripeptide fragments of the corresponding proteins.[46,47]

Our results provide an illustration of the close relationship between mRNA expression levels and protein solubility, which arises from the evolutionary pressure for *E. coli* proteins to avoid aggregation when expressed at their maximal levels during the cell cycle. As a consequence of such evolutionary pressure, the link between mRNA abundance and the physicochemical properties of the sequences can be established only for the log phase of bacterial growth and not the stationary phase.[28] Since recombinant proteins expressed in *E. coli* are not subject to a similar evolutionary pressure, they often aggregate at least in part when their production is forced to take place at very high levels. Nevertheless, we have shown that the fraction of a given protein that remains soluble can be predicted from its amino acid sequence, as a consequence of the existence of a relationship between mRNA expression levels and protein solubility. It has also been observed that recombinant proteins tend to form amyloid-like aggregates in *E. coli* inclusion bodies.[48,49] This finding indicates that the propensity to form ordered aggregates[9,16] is a crucial factor in determining protein expression, as expected on the basis of this as well as previous work.[13]

Since a relationship between mRNA expression levels and protein solubility is implied by the need to avoid aggregation at the protein concentrations required for optimal cellular function,[13] it is expected to be at least to some extent specific for given organisms, since it is likely to be the result of a coevolution of the regulatory processes at the cellular level and of the protein sequences at the biochemical level. It thus remains to be established if a different set of parameters for the CamEL method needs to be obtained when a different expression system or specific stress conditions are used. Through microarray expression technology, such parametrisation can now be carried out with relative ease.[50] In the light of the limitations of existing approaches for predicting the solubility of proteins in heterologous environments,[47,51,52] the CamEL method should be of considerable value in high-throughput expression programmes.

The up-regulation and down-regulation of genes are crucial for all cellular functions and for adaptation to environmental changes.[53] Since detailed regulatory mechanisms have been established for several genes, and a general understanding of structural and dynamical properties of transcriptional networks is emerging,[54] it will be interesting to investigate whether the approach that we have described could be extended to the prediction of the changes in expression levels under varying environmental conditions. In addition, we expect that, by taking into account the subcellular localisations of the proteins,[55] it might be possible to improve yet further the quality of the predictions. Thus, the CamEL method of predicting mRNA expression levels can be expected to be complementary to those that exploit our rapidly increasing knowledge of the regulatory network at the cellular level,[19,56,57] since we look at the chemical building blocks that make up the molecules that have evolved to participate in these processes.

In summary, the results that we have described in this article provide further support for the idea that the physicochemical properties of proteins have coevolved with their cellular environments to optimise the efficiency of the biochemical processes on which all living systems depends.[13,58]

Materials and Methods

Prediction of mRNA expression levels from the aggregation propensities of the corresponding proteins

We analyze the information contained in the aggregation propensities of 50 protein sequences associated with the highest expression levels and 50 protein sequences associated with the lowest expression levels in the database of 2800 cytosolic proteins that we have used.[24] Each protein sequence is cut in 25 fragments, and the predicted mRNA expression levels are then written as a function of the distributions of aggregation-prone and aggregation-resistant properties of the fragments

$$E^{\mathrm{pred}} = \sum\nolimits_i \left[a_i D_i^{\mathrm{agg}} - b_i D_i^{agg-}\right] \qquad (1)$$

where $D^{\mathrm{agg-}}$ and $D^{\mathrm{agg+}}$ are the distribution of negative and positive aggregation propensities, respectively. The weights a and b are determined, requiring that the scores of proteins associated with high expression levels are higher than the scores of proteins associated with low expression levels

$$E^{\mathrm{pred}}(\text{high expression}) > E^{\mathrm{pred}}(\text{low expression}) \qquad (2)$$

Prediction of mRNA expression levels from the physicochemical properties of the amino acid sequences of the corresponding proteins

For each fragment k, three chemical properties are used as input of a neural network

$$\underline{i}_k = \left(i_k^1, i_k^2, i_k^3\right) \qquad (3)$$

Between the inputs $\underline{i}_k$ and the output o, we introduce a hidden layer h_k to combine the individual contributions

$$h_k = \tanh\left(\sum\nolimits_{ja} \varphi_{jak} i_j^a + \gamma_j^a\right) \qquad (4)$$

$$o = \tanh\left(\sum\nolimits_k \psi_k h_k + \delta_k\right) \qquad (5)$$

The weights φ_{jak}, γ_j^a, ψ_k and δ_k are estimated using a back-propagation algorithm. To avoid proliferation of internal variables, we reduce the number of the internal variables proportionally to the number of data points. For c=25, we observe the highest predictive power (Fig. S1a in Supplementary Information). As indicated in Eq. (4), the contribution of each fragment is associated with a different weight, which makes our model position dependent.

Estimation of the importance of individual physicochemical properties

To estimate the relationship between a physicochemical property of an amino acid sequence and the level of expression of the corresponding gene, we introduce the concept of *individual contribution*. For each fragment n_{ak}) weights related to the chemical property a

$$\begin{aligned} p_{ak} &= \sum\nolimits_j \theta(+\varphi_{jak})\theta(+\psi_j) + \theta(+\varphi_{jak})\theta(-\psi_j) \\ n_{ak} &= \sum\nolimits_j \theta(-\varphi_{jak})\theta(+\psi_j) + \theta(+\varphi_{jak})\theta(-\psi_j) \end{aligned} \qquad (6)$$

where the function θ is defined as $\theta(x)$=0 if x<0 and $\theta(x)$=1 if x>0.

The variables p_{ak} and n_{ak} correlate the output with the sign of the weights of the property a at position k. We count all the signs associated with the property a

$$\begin{aligned} p_a &= \textstyle\sum_k p_{ak} \\ n_a &= \textstyle\sum_k n_{ak} \end{aligned} \tag{7}$$

We define the correlation for the individual contribution ρ_a of the property a

$$\rho_a = (p_a - n_a)/t_a \tag{8}$$

where t_a is the total number of internal weights.

Optimisation of the number of fragments

By increasing the number of fragments from 1 to 20, we observe that the accuracy in the predictions improves from 46% to 80% (Fig. S1a in Supplementary Information). If the number of cuts is beyond 25, the performance decreases, indicating that the best accuracy permitted by the model and the data is obtained when the number of fragments is 25 (Fig. S1b in Supplementary Information).

Prediction of the order of magnitude

For the prediction of the order of magnitude of the expression levels, the experimental data set is partitioned into six classes, while for the prediction of the solubility, the data set is partitioned into four classes.

Cross-validation

The data set of cytoplasmic proteins is randomly partitioned into five subsamples requiring the condition that each partition carries the same distribution of experimental expression levels. One subsample is retained for testing, and the remaining four are used for training the algorithm. The cross-validation process is repeated five times with each of the five subsamples used exactly once as the validation data.

Combination of different predictors

Support vector machines are used to combine the information from the different algorithms trained on hydropathy properties, secondary-structure propensities and translation factors. An additional 5-fold cross-validation is used to estimate the performance of the support vector machine (see Supplementary Information).

Protein solubility and the hEx1 database

In the hEx1 library, the *soluble fraction* of a protein is defined as an integer that ranges from 0 (no soluble expression) to 3 (strong soluble expression).[32] The library contains 1287 recombinant proteins (clones), but often more than one clone is associated to a specific gene. We use a consensus of the different expression data to generate a reference data set. For instance, for the Ensembl transcript ENST00000315491, we assume that the soluble expression strength is 0, because 6 clones are associated with soluble expression strength 0 and only 1 clone has soluble expression strength 1. The database was generated using the same vector, strain and helper plasmid for all the genes (vector: pQE30NST, strain: SCS1 and helper plasmid: pSE111), which eliminates the contribution of exogenous factors to gene expression (see Supplementary Information).

Acknowledgements

This research was supported by the Wellcome Trust, the Leverhulme Trust, the Medical Research Council, the European Molecular Biology Organization and the Royal Society. The authors thank Dr. Carlos W. Bertoncini, Dr. Damian C. Crowther, Dr. Thomas R. Jahn and Dr. Leila Luheshi for interesting discussions. The CamEL method is freely available for academic use at http://www-vendruscolo.ch.cam.ac.uk/software.html.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2009.03.002

References

1. Levine, M. & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, **424**, 147–151.
2. Lackner, D. H., Beilharz, T. H., Marguerat, S., Mata, J., Watt, S., Schubert, F. *et al.* (2007). A network of multiple regulatory layers shapes gene expression in fission yeast. *Mol. Cell*, **26**, 145–155.
3. Balaji, S., Babu, M. M., Iyer, L. M., Luscombe, N. M. & Aravind, L. (2006). Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J. Mol. Biol.* **360**, 213–227.
4. Dobson, C. M. (1999). Protein misfolding, evolution and disease. *Trends Biochem. Sci.* **24**, 329–332.
5. Blond-Elguindi, S., Cwirla, S. E., Dower, W. J., Lipshutz, R. J., Sprang, S. R., Sambrook, J. F. & Gething, M. J. (1993). Affinity panning of a library of peptides displayed on bacteriophages reveals the binding specificity of BiP. *Cell*, **75**, 717–728.
6. Brewer, J. W. & Diehl, J. A. (2000). PERK mediates cell-cycle exit during the mammalian unfolded protein response. *Proc. Natl Acad. Sci. USA*, **97**, 12625–12630.
7. Hartl, F. U. & Hayer-Hartl, M. (2002). Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science*, **295**, 1852–1858.
8. McClellan, A. J., Xia, Y., Deutschbauer, A. M., Davis, R. W., Gerstein, M. & Frydman, J. (2007). Diverse cellular functions of the Hsp90 molecular chaperone uncovered using systems approaches. *Cell*, **131**, 121–135.
9. Chiti, F. & Dobson, C. M. (2006). Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.* **75**, 333–366.
10. Ventura, S. (2005). Sequence determinants of protein aggregation: tools to increase protein solubility. *Microb. Cell Fact.* **4**, 11.
11. Linding, R., Schymkowitz, J., Rousseau, F., Diella, F. & Serrano, L. (2004). A comparative study of the relationship between protein structure and beta-

aggregation in globular and intrinsically disordered proteins. *J. Mol. Biol.* **342**, 345–353.

12. Tartaglia, G. G., Pellarin, R., Cavalli, A. & Caflisch, A. (2005). Organism complexity anti-correlates with proteomic beta-aggregation propensity. *Protein Sci.* **14**, 2735–2740.
13. 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.
14. 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.
15. Fernandez-Escamilla, A., 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.
16. Sánchez de Groot, N., Pallarés, I., Avilés, F. X., Vendrell, J. & Ventura, S. (2005). Prediction of "hot spots" of aggregation in disease-linked polypeptides. *BMC Struct. Biol.* **5**, 18.
17. 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.
18. 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.
19. Beer, M. A. & Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell*, **117**, 185–198.
20. Thanaraj, T. A. & Argos, P. (1996). Ribosome-mediated translational pause and protein domain organization. *Protein Sci.* **5**, 1594–1612.
21. Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K. & Sarai, A. (2004). ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.* **32**, D120–D121.
22. Ashcroft, M., Kubbutat, M. H. & Vousden, K. H. (1999). Regulation of p53 function and stability by phosphorylation. *Mol. Cell. Biol.* **19**, 1751–1758.
23. Zhang, Y., Olsen, D. R., Nguyen, K. B., Olson, P. S., Rhodes, E. T. & Mascarenhas, D. (1998). Expression of eukaryotic proteins in soluble form in *Escherichia coli*. *Protein Expr. Purif.* **12**, 159–165.
24. Sundararaj, S., Guo, A., Habibi-Nazhad, B., Rouani, M., Stothard, P., Ellison, M. & Wishart, D. S. (2004). The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of *Escherichia coli*. *Nucleic Acids Res.* **32**, D293–D295.
25. Tartaglia, G. G. & Vendruscolo, M. (2008). The Zyggregator method for predicting protein aggregation propensities. *Chem. Soc. Rev.* **37**, 1395–1401.
26. Tartaglia, G. G., Cavalli, A., Pellarin, R. & Caflisch, A. (2005). Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Sci.* **14**, 2723–2734.
27. Pawar, A. P., Dubay, K. F., Zurdo, J., Chiti, F., Vendruscolo, M. & Dobson, C. M. (2005). Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. *J. Mol. Biol.* **350**, 379–392.
28. Vendruscolo, M. & Tartaglia, G. G. (2008). Towards quantitative predictions in cell biology using chemical properties of proteins. *Mol. BioSyst.* **4**, 1170–1175.
29. Benita, Y., Wise, M. J., Lok, M. C., Humphery-Smith, I. & Oosting, R. S. (2006). Analysis of high throughput protein expression in *Escherichia coli*. *Mol. Cell. Proteomics*, **5**, 1567–1580.
30. Todd, A. E., Marsden, R. L., Thornton, J. M. & Orengo, C. A. (2005). Progress of structural genomics initiatives: an analysis of solved target structures. *J. Mol. Biol.* **348**, 1235–1260.
31. Ventura, S. & Villaverde, A. (2006). Protein quality in bacterial inclusion bodies. *Trends Biotechnol.* **24**, 179–185.
32. Büssow, 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**, R71.
33. Lockhart, D. J. & Barlow, C. (2001). Expressing what's on your mind: DNA arrays and the brain. *Nat. Rev., Neurosci.* **2**, 63–68.
34. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, **297**, 1183–1186.
35. Tartaglia, G. G., Cavalli, A., Pellarin, R. & Caflisch, A. (2004). The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein Sci.* **13**, 1939–1941.
36. Bemporad, F., Calloni, G., Campioni, S., Plakoutsi, G., Taddei, N. & Chiti, F. (2006). Sequence and structural determinants of amyloid fibril formation. *Acc. Chem. Res.* **39**, 620–627.
37. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. (2005). Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA*, **102**, 14338–14343.
38. Eames, M. & Kortemme, T. (2007). Structural mapping of protein interactions reveals differences in evolutionary pressures correlated to mRNA level and protein abundance. *Structure*, **15**, 1442–1451.
39. Komar, A. A. (2008). A pause for thought along the co-translational folding pathway. *Trends Biochem. Sci.* **34**, 16–24.
40. Krasheninnikov, I. A., Komar, A. A. & Adzhubei, I. A. (1991). Nonuniform size distribution of nascent globin peptides, evidence for pause localization sites, and a contranslational protein-folding model. *J. Protein Chem.* **10**, 445–453.
41. Irwin, B., Heck, J. D. & Hatfield, G. W. (1995). Codon pair utilization biases influence translational elongation step times. *J. Biol. Chem.* **270**, 22801–22806.
42. Kittle, D. (2006). Radical changes in the engineering of synthetic genes for protein expression. *BioPharm. Int.* **19**, 12–18.
43. Gu, X., Hewett-Emmett, D. & Li, W. (1998). Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. *Genetica*, **102–103**, 383–391.
44. Epstein, R. J., Lin, K. & Tan, T. W. (2000). A functional significance for codon third bases. *Gene*, **245**, 291–298.
45. Akashi, H. & Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc. Natl Acad. Sci. USA*, **99**, 3695–3700.
46. Raghava, G. P. S. & Han, J. H. (2005). Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein. *BMC Bioinformatics*, **6**, 59.
47. Idicula-Thomas, S. & Balaji, P. V. (2005). Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in *Escherichia coli*. *Protein Sci.* **14**, 582–592.

48. Ami, D., Natalello, A., Taylor, G., Tonon, G. & Maria Doglia, S. (2006). Structural analysis of protein inclusion bodies by Fourier transform infrared microspectroscopy. *Biochim. Biophys. Acta*, **1764**, 793–799.
49. Carrió, M., González-Montalbán, N., Vera, A., Villaverde, A. & Ventura, S. (2005). Amyloid-like properties of bacterial inclusion bodies. *J. Mol. Biol.* **347**, 1025–1037.
50. Hoheisel, J. D. (2006). Microarray technology: beyond transcript profiling and genotype analysis. *Nat. Rev., Genet.* **7**, 200–210.
51. Wilkinson, D. L. & Harrison, R. G. (1991). Predicting the solubility of recombinant proteins in *Escherichia coli*. *Biotechnology (N Y)*, **9**, 443–448.
52. Smialowski, P., Martin-Galiano, A. J., Mikolajka, A., Girschick, T., Holak, T. A. & Frishman, D. (2007). Protein solubility: sequence based prediction and experimental verification. *Bioinformatics*, **23**, 2536–2542.
53. Tarassov, K., Messier, V., Landry, C. R., Radinovic, S., Serna Molina, M. M., Shames, I. *et al.* (2008). An in vivo map of the yeast protein interactome. *Science*, **320**, 1465–1470.
54. Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A. & Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–312.
55. Shav-Tal, Y., Singer, R. H. & Darzacq, X. (2004). Imaging gene expression in single living cells. *Nat. Rev., Mol. Cell Biol.* **5**, 855–861.
56. Nguyen, D. H. & D'haeseleer, P (2006). Deciphering principles of transcription regulation in eukaryotic genomes. *Mol. Syst. Biol.* **2**, 0012.
57. Bussemaker, H. J., Foat, B. C. & Ward, L. D. (2007). Predictive modeling of genome-wide mRNA expression: from modules to molecules. *Annu. Rev. Biophys. Biomol. Struct.* **36**, 329–347.
58. Tartaglia, G. G. & Caflisch, A. (2007). Computational analysis of the *S. cerevisiae* proteome reveals the function and cellular localization of the least and most amyloidogenic proteins. *Proteins*, **68**, 273–278.