doi:10.1016/j.jmb.2011.12.005 J. Mol. Biol. (2012) 421, 237–241

Contents lists available at www.sciencedirect.com

Journal of Molecular Biology

journal homepage: http://ees.elsevier.com.jmb

COMMUNICATION

Sequence-Based Prediction of Protein Solubility

Federico Agostini [1], Michele Vendruscolo [2*] and Gian Gaetano Tartaglia [1*]

[1]*Centre for Genomic Regulation (CRG) and Universitat Pompeu Fabra (UPF), Dr. Aiguader, 88, Barcelona 08003, Spain*
[2]*Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK*

Received 13 October 2011; received in revised form 1 December 2011; accepted 4 December 2011
Available online 9 December 2011

Edited by S. Radford

Keywords:
protein aggregation; protein solubility; protein folding; *E. coli* proteome

In order to investigate the relationship between the thermodynamics and kinetics of protein aggregation, we compared the solubility of proteins with their aggregation rates. We found a significant correlation between these two quantities by considering a database of protein solubility values measured using an *in vitro* reconstituted translation system containing about 70% of *Escherichia coli* proteins. The existence of such correlation suggests that the thermodynamic stability of the native states of proteins relative to the aggregate states is closely linked with the kinetic barriers that separate them. In order to create the possibility of conducting computational studies at the proteome level to investigate further this concept, we developed a method of predicting the solubility of proteins based on their physicochemical properties.

Protein folding and aggregation are processes in competition in the cellular environment.[1–5] When proteins fail to fold, they tend to form aggregates, whose presence is associated with a variety of human conditions, including Alzheimer's and Parkinson's diseases.[1,2,6,7] Since it is becoming clear that the amino acid sequences of proteins encode not just their folding but also their aggregation, there is great interest in identifying the amino acid code responsible for the aggregation process.[8–16] Considerable advances have been made in establishing the relationship between the physicochemical properties of proteins and their aggregation rates, thus supporting the view that the kinetics of protein aggregation are encoded in the amino acid sequences.[8–12] Since, however, it has also been suggested that the native states of proteins are metastable against aggregation,[17,18] it is important to understand also how the thermodynamics of protein aggregation are specified in the sequence. A difficult challenge for these studies is the scarcity of proteome-level data on protein solubility, which makes it difficult to analyze the relationship between the relative stability of the native and aggregate states.

To address this problem, we took advantage of a recent study in which the solubility of about 70% of the *Escherichia coli* proteins was experimentally measured *in vitro*.[19] By analyzing a range of methods to predict protein aggregation rates, the authors of that study found that protein solubility is not correlated with the intrinsic aggregation rates, that is, the rates of conversion between the unfolded and aggregated states of proteins.[19] They thus suggested that a quantitative analysis of protein solubility requires the consideration of the protection against aggregation provided by the native state.[19] By taking up such as suggestion, we consider here whether the solubility of proteins correlates with the aggregation propensity of proteins, when the aggregation is estimated from the native state.[12] The data set employed in this work is

*Corresponding authors. E-mail addresses: mv245@cam.ac.uk; gian.tartaglia@crg.eu.

Abbreviation used: SVM, support vector machine.

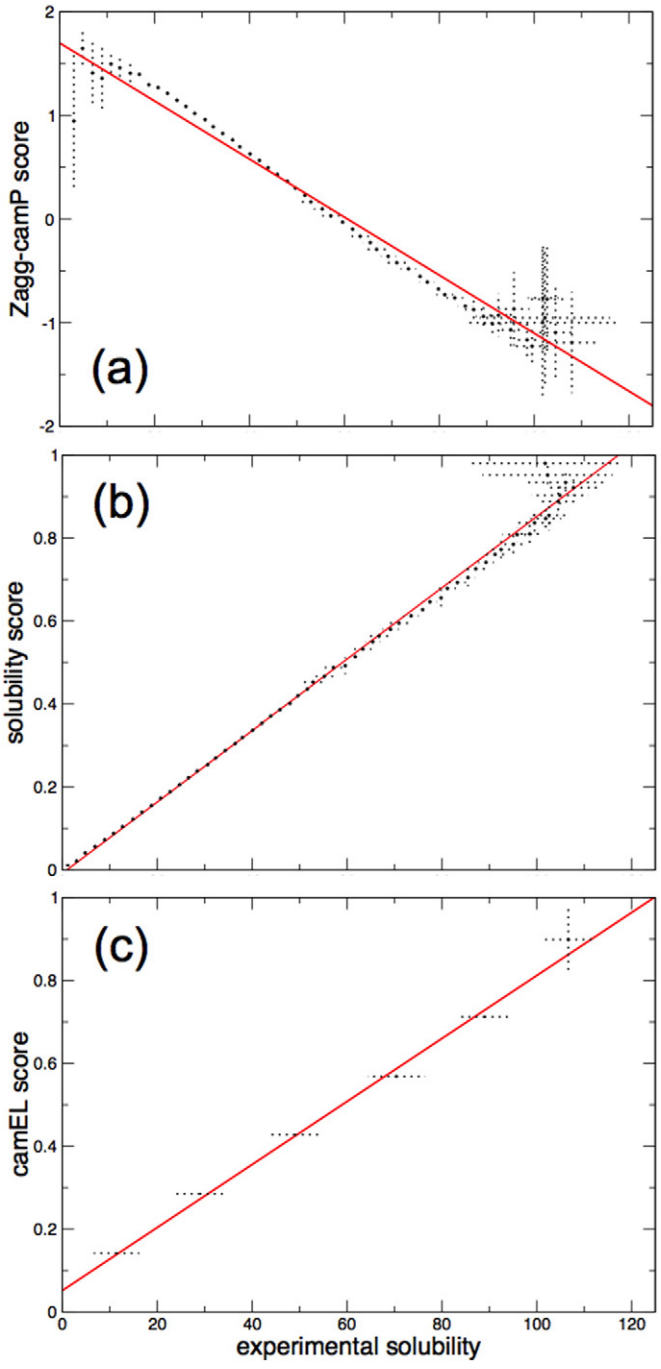

Fig. 1. Correlation between experimental[19] and predicted solubility scores: (a) aggregation rates from native states, calculated using the Zyggregator method;[12] (b) solubility scores, calculated using the CCSOL method discussed in this work; and (c) maximal expression level scores, calculated using the CamEL method.[20] The output of the CamEL method is a discrete variable ranging from 1 to 6 (here normalized to 1).

the one provided by Niwa *et al.*,[19] and protein identifiers were collected from the Ensembl Bacteria Database†. Our results indicate that protein aggregation rates and protein solubility values are highly correlated if the propensity of aggregation is calculated from the folded state[12] (Fig. 1a).

These results are intriguing since the aggregation propensity scores provide a prediction of the rate at which proteins aggregate, but they do not represent a direct prediction of the critical concentration of proteins, that is, their solubility, which is the parameter measured by Niwa *et al.*[19] The finding that protein aggregation rates are correlated with critical concentrations suggests the existence of a link between the thermodynamics and kinetics of protein aggregation, which would arise because the

† http://bacteria.ensembl.org/index.html

solubility of proteins depends on the relative stability of the folded and aggregate states, whereas the aggregation rates depend on free-energy barriers between these states. It is worth observing that, also in the case of protein folding, the rate at which this process takes place has been linked with the thermodynamic properties of the folded and unfolded states.[21,22]

In order to investigate further this phenomenon, we developed a predictor of protein solubility based on the physicochemical properties of amino acid sequences. For each sequence within the data set, we calculated the profiles of 28 physicochemical properties collected through a literature search (Fig. 2a and Supplementary Material). Physicochemical profiles were generated by a window of seven amino acids sliding from the N- to the C-terminus of the protein sequence. We split the original data set of 3043 proteins[19] into three subsets containing the most soluble (1081 entries, "head set"), least soluble (1078 entries, "tail set") and all the other proteins (884 entries), respectively.

In a first step, we built a support vector machine (SVM) to identify properties that allow the best discrimination between the "head" and "tail" sets. This SVM recognizes the main features that differentiate two groups of protein sequences and uses this criterion to select them. In the selection process, the SVM compares each protein of one data set with all the proteins in the other data set. Proteins scoring above a given threshold (80%) are discriminated and removed from the original data set for the next

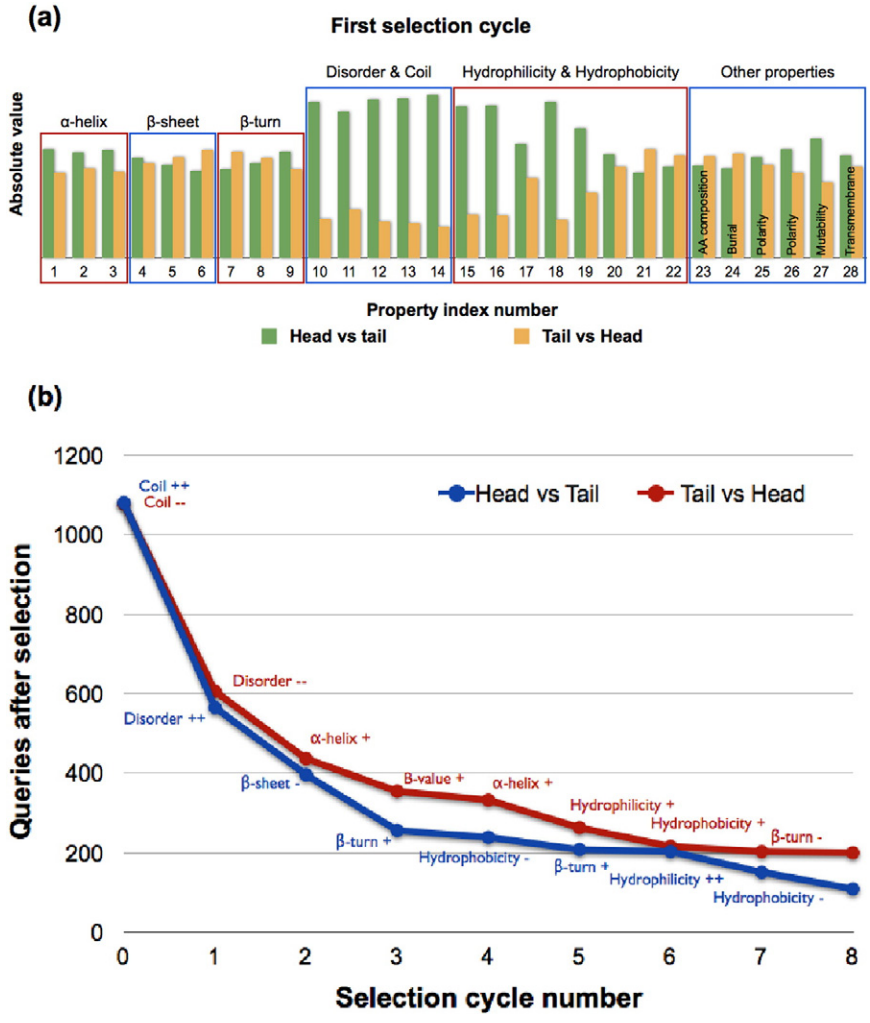

Fig. 2. Comparison of the average values of 28 amino acid propensity scales for the *E. coli* proteins considered in this work.[19] (a) Physicochemical profiles are calculated for each property, and comparisons are made between the "head" and "tail" data sets. (b) Graphical representation of the properties selected during eight iterations: in each cycle, the best-discriminated sequences are removed for a new round of analysis. Plus and minus signs indicate whether a positive or negative selection is applied to a data set [(+) positive selection, (+ +) positive selection and steep gradient, (−) negative selection, (− −) negative selection and steep gradient].

iterative round (Fig. 2b). After eight iterations, only a small number of entries (200) remained un-discriminated, and a number of physicochemical characteristics (11) were collected (Fig. S1). We found that disorder, coil and hydrophilicity propensities best discriminate between "head" and "tail" (Fig. 2a and Supplementary Material). It should be mentioned that polar residues are associated with high coil and disorder propensities: P, E, S, K and Q are the most disorder-prone residues,[23] and N, D, G, H and P are the most coil-prone ones.[21-27] For this reason, coil and disorder are selected in our method as determinants of protein solubility. In agreement with this finding, we observe that disorder predictions carried out with DisEMBL REM465 (i.e., of missing assignment of electron density) show a correlation of 45% with the experimental solubility measured by Niwa *et al.*[19] (data not shown). Nevertheless, it is worth mentioning that disordered proteins contain more hydrophilic regions but still retain potential for aggregation because their hydrophobic regions cannot be buried to the solvent and are consequently available for interaction.[12]

In a second step, we combined a number of physicochemical properties into a SVM to predict protein solubility. In order to reduce the number of variables and identify those that give the strongest signal, we generated $\sum_k \binom{11}{k} = 2^{11} = 2048$ SVMs (all the combinations of 11 scales) and ranked them according to their performances upon cross-validation. In our cross-validation, one subsample of the original data set is retained for testing, and the remaining nine are used for training the algorithm. The cross-validation process was repeated 10 times with each of the 10 subsamples used exactly once as the validation data. At the end of the process, we identified six properties: coil/disorder,[21-27] hydrophobicity,[25] hydrophilicity,[26] β-turn,[27,28] α-helix[27] (Fig. 1b and Supplementary Material). The method is available online‡.

Further insight about the relationship between solubility and aggregation rates (Fig. 1a) is provided by considering the correlation that was recently reported between protein aggregation rates and mRNA expression levels, which arises as a consequence of the stringent requirement for proteins to remain soluble in order to function at the concentrations at which they are expressed in the cell.[17,29–33] On the basis of this observation, we demonstrated that it is possible to estimate the solubility of recombinant human proteins in *E. coli* from the corresponding maximal mRNA expression levels.[20] Here, we used this approach to provide an alternative prediction of the solubility scores provided by Niwa *et al.*[19] finding a very high correlation (Fig. 1c). Taken together, these results suggest that kinetics and thermodynamics may play equally important roles in determining the stability of native states against aggregation in living systems, which is consistent with the view that the native forms of proteins can be only metastable.[17,18]

In summary, in this work, we have developed a method to predict the solubility of proteins, which, used in combination with existing methods for predicting aggregation rates, should open the possibility of carrying out proteome-level studies of the relationship between thermodynamics and kinetics of protein aggregation.

‡ http://tartaglialab.crg.cat/ccsol.html

Acknowledgements

We are grateful to Prajwal Ciryam, Giulia Tomba, Davide Cirillo and Hideki Taguchi for stimulating discussions. We thank Toni Hermoso Pulido for setting up the CCSOL web server.

Supplementary Data

Supplementary data to this article can be found online at doi:10.1016/j.jmb.2011.12.005

References

1. Dobson, C. M. (2003). Protein folding and misfolding. *Nature*, **426**, 884–890.
2. Balch, W. E., Morimoto, R. I., Dillin, A. & Kelly, J. W. (2008). Adapting proteostasis for disease intervention. *Science*, **319**, 916–919.
3. Jahn, T. R. & Radford, S. E. (2008). Folding *versus* aggregation: polypeptide conformations on competing pathways.. *Arch. Biochem. Biophys.* **469**, 100–117.
4. Hartl, F. U., Bracher, A. & Hayer-Hartl, M. (2011). Molecular chaperones in protein folding and proteostasis. *Nature*, **475**, 324–332.
5. Tartaglia, G. G. & Vendruscolo, M. (2010). Proteome-level interplay between folding and aggregation propensities of proteins. *J. Mol. Biol*, **402**, 919–928.
6. Selkoe, D. J. (2003). Folding proteins in fatal ways. *Nature*, **426**, 900–904.
7. 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.
8. 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.
9. 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. Biotech.* **22**, 1302–1306.

10. 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.
11. Conchillo-Sole, O., de Groot, N. S., Aviles, F. X., Vendrell, J., Daura, X. & Ventura, S. (2007). Aggrescan: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinf.* **8**.
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. Belli, M., Ramazzotti, M. & Chiti, F. (2011). Prediction of amyloid aggregation *in vivo*. *EMBO Rep.* **12**, 657–663.
14. Bryan, A. W., Menke, M., Cowen, L. J., Lindquist, S. L. & Berger, B. (2009). BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis. *PLoS Comp. Biol*, **5**, e1000333.
15. Maurer-Stroh, S., Debulpaep, M., Kuemmerer, N., de la Paz, M. L., Martins, I. C., Reumers, J. *et al.* (2010). Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat. Methods*, **7**, 237–242.
16. Thompson, M. J., Sievers, S. A., Karanicolas, J., Ivanova, M. I., Baker, D. & Eisenberg, D. (2006). The 3D profile method for identifying fibril-forming segments of proteins. *Proc. Natl Acad. Sci. USA*, **103**, 4074–4078.
17. 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.
18. Baldwin, A. J., Knowles, T. P. J., Tartaglia, G. G., Fitzpatrick, A. W., Devlin, G. L., Shammas, S. L. *et al.* (2011). Metastability of native proteins and the phenomenon of amyloid formation. *J. Am. Chem. Soc.* **133**, 14160–14163.
19. 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.
20. 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.
21. Plaxco, K. W., Simons, K. T. & Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994.
22. Dinner, A. R. & Karplus, M. (2001). The roles of stability and contact order in determining protein folding rates. *Nat. Struct. Biol.* **8**, 21–22.
23. Campen, A., Williams, R. M., Brown, C. J., Meng, J. W., Uversky, V. N. & Dunker, A. K. (2008). Top-idp-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept. Lett.* **15**, 956–963.
24. 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.
25. Engelman, D. M., Steitz, T. A. & Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Bioph. Biophys. Chem.* **15**, 321–353.
26. Hopp, T. P. & Woods, K. R. (1981). Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl Acad. Sci. USA*, **78**, 3824–3828.
27. Deleage, G. & Roux, B. (1987). An algorithm for protein secondary structure prediction based on class prediction. *Protein Eng.* **1**, 289–294.
28. Levitt, M. (1978). Conformational preferences of amino-acids in globular proteins. *Biochemistry*, **17**, 4277–4284.
29. Tartaglia, G. G. & Vendruscolo, M. (2009). Correlation between mRNA expression levels and protein aggregation propensities in subcellular localisations. *Mol. BioSys.* **5**, 1873–1876.
30. Vendruscolo, M. & Tartaglia, G. G. (2008). Towards quantitative predictions in cell biology using chemical properties of proteins. *Mol. BioSys.* **4**, 1170–1175.
31. Castillo, V., Espargaro, A., Gordo, V., Vendrell, J. & Ventura, S. (2010). Deciphering the role of the thermodynamic and kinetic stabilities of SH3 domains on their aggregation inside bacteria. *Proteomics*, **10**, 4172–4185.
32. Castillo, V., Grana-Montes, R. & Ventura, S. (2011). The aggregation properties of *Escherichia coli* proteins associated with their cellular abundance. *Biotech. J.* **6**, 752–760.
33. Espargaro, A., Castillo, V., de Groot, N. S. & Ventura, S. (2008). The *in vivo* and *in vitro* aggregation properties of globular proteins correlate with their conformational stability: the SH3 case. *J. Mol. Biol.* **378**, 1116–1131.