Structure
Ways & Means

Prediction of Local Structural Stabilities of Proteins from Their Amino Acid Sequences

Gian Gaetano Tartaglia,[1] Andrea Cavalli,[1] and Michele Vendruscolo[1],*
[1] Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom
*Correspondence: mv245@cam.ac.uk

DOI 10.1016/j.str.2006.12.007

SUMMARY

Hydrogen exchange experiments provide detailed information about the local stability and the solvent accessibility of different regions of the structures of folded proteins, protein complexes, and amyloid fibrils. We introduce an approach to predict protection factors from hydrogen exchange in proteins based on the knowledge of their amino acid sequences without the inclusion of any additional structural information. These results suggest that the propensity of different regions of the structures of globular proteins to undergo local unfolding events can be predicted from their amino acid sequences with an accuracy of 80% or better.

The information contained in the amino acid sequences determines the folding of globular proteins into their native structures (Anfinsen, 1973), their specificities for interaction with other molecules (Janin and Chothia, 1976), and their lifetimes and stability with respect to the unfolded state (Warshel and Levitt, 1976). However, the reliable prediction of the structure of a protein from the knowledge of the sequence remains one of the most challenging tasks in structural biology. Recently, significant advances have been made, as testified by the increasingly accurate results of the CASP experiments, and in many cases structural models at about 4–5 Å resolution or better can be obtained (Moult, 2005; Schueler-Furman et al., 2005). These results suggest that the local stability, the structural flexibility, and the degree of exposure to solvent of different regions of the structure of a protein may also be directly predicted from the sequence. Because this type of information can be obtained through hydrogen exchange measurements (Woodward and Hilton, 1980; Bai et al., 1993; Chamberlain et al., 1996; Clarke and Itzhaki, 1998; Fersht, 1999; Englander, 2000), in this study we introduce an approach that uses only the knowledge of the amino acid sequence of a protein to predict protection factors obtained from equilibrium hydrogen exchange experiments.

The protection factor for residue i, $P_i = k_i^{int}/k_i$, is the ratio of the intrinsic rate, k_i^{int}, observed in an unstructured peptide (Bai et al., 1993; Hvidt and Nielsen, 1966), to the observed amide hydrogen exchange rate, k_i. In the case in which an amide hydrogen can exchange only when the protein is substantially unfolded, the local stability is equal to the global stability and the amide is said to be undergoing "global" exchange. By contrast, the so-called "local" exchange occurs through localized fluctuations of the structure and can be applied to study native state fluctuations. Whether an amide hydrogen is exchanging locally or globally must be determined experimentally, by using denaturant dependence (Bai and Englander, 1996; Chu et al., 2002; Itzhaki et al., 1997), temperature dependence (Bai and Englander, 1996), or mutagenesis (Neira et al., 1997).

Although experimental methods for measuring protections factors are well established, with the advent of structural genomics initiatives (Vitkup et al., 2001) and proteomics analysis (Pandey and Mann, 2000), it would be convenient to establish reliable theoretical approaches to predict protection factors without carrying out experimental measurements. In the last years, a variety of computational approaches has been proposed to predict protection factors from the knowledge of the structures of proteins (Hilser and Freire, 1996; Sheinerman and Brooks, 1998; Bahar et al., 1998; Garcia and Hummer, 1999; Viguera and Serrano, 2003; Dixon et al., 2004). For example, it has been recently shown that experimental protection factors arising from local exchange can be approximated with an accuracy of ~85% by using the following phenomenological equation (Vendruscolo et al., 2003; Best and Vendruscolo, 2006):

$$\ln P_i = b_c N_i^c + b_h N_i^h. \qquad (1)$$

Equation 1 is based on an interpretation of equilibrium hydrogen exchange measurements (Woodward and Hilton, 1980; Clarke and Itzhaki, 1998; Englander, 2000) in terms of the properties of the structure; protection from hydrogen exchange is assumed to arise either from burial or from hydrogen bonding (Vendruscolo et al., 2003; Best and Vendruscolo, 2006; Gsponer et al., 2006). N_i^c is the contribution from burial, and N_i^h is the number of hydrogen bonds for the amide hydrogen of residue i (Vendruscolo et al., 2003). The parameters b_c and b_h give the free energy contributions of creating a van der Waals contact or a hydrogen bond, respectively. In the subsequent paragraphs, we refer to the values of a protein structure as the "ln*P* profile."

In this work, we present a method (CamP; http://www-almost.ch.cam.ac.uk/camp.php) of predicting

Table 1. Comparison between Different Predictions of Protection Factors

Protein Name	C_{NE}	C_{RE}	C_{DE}	C_{PE}	C_{BE}	BH
Apo α lactalbumin	50	40	−13	−23	−22	1HFZ
Equine lysozyme	66	69	−17	−44	−43	1JSF
Human lysozyme T70N	62	62	−22	−40	−23	1JSF
Snake venom CTXIII	63	35	−17	+14	−23	1H0J
Cytochrome C	67	61	−49	−47	−10	1HRC
α spectrin	71	65	−37	−53	−30	1BK2
Ovomucoid protein	69	41	−27	−33	−20	4OVO
Leucine zipper	55	48	−79	−88	−35	1CE9
Ferricytochrome C-551	45	47	−45	−22	−20	2EXV
Snake venom CBTX	50	37	−30	−12	−25	1ONJ
Llama antibody	50	29	−10	−30	−30	1DZB
Chemotactic CheY	52	n/a	−16	−27	−10	1A0O

C, correlation between predictions provided by methods x and y; N, protection factor predicted by CamP; E, experimental protection factors; R, I-SITES/HMMSTR/ROSETTA (http://www.bioinfo.rpi.edu/bystrc/hmmstr/server.php) (Bystroff and Shao, 2002); D, DisEMBL (http://dis.embl.de/) (Linding et al., 2003); p, PONDR (http://www.pondr.com/) (Obradovic et al., 2006); B, B factors; BH, homolog from which the B factors are taken, which is the one that, according to PSI-BLAST (Altschul et al., 1997), has the highest sequence similarity with the protein used in the hydrogen exchange experiment; n/a, I-SITES/HMMSTR/ROSETTA failed to provide a structure.

protection factors directly from the amino acid sequence that does not require any knowledge of the native structure of a protein. We tested the performance of the method over a set of 12 proteins (Table 1). We find a good agreement between predicted and experimental protection factors (Lacroix et al., 1997; Russell et al., 2003; Sivaraman et al., 2000; Sadqi et al., 1999; Wand et al., 1986; Swint-Kruse and Robertson, 1996; Morozova et al., 1995; Goodman and Kim, 1991; Perez et al., 2001; Wijesinha-Bettoni et al., 2001) with a correlation in the range of 50%–70% (Figure 1; Table 1, column C_{NE}; Figure S1, see the Supplemental Data available with this article online). Although the method has been parameterized to predict EX2 exchange, it also gives good results for EX1 exchange, as in the case of the C-terminal region of horse cytochrome C (Figure 1) (Milne et al., 1998). Remarkably, these results enable the prediction of regions of a structure that are protected from hydrogen exchange with an accuracy in the range 80%–100%; we define a region to be protected if $\ln p > 5$ for all of its amide groups. The method that we present can also be used, at least in principle, for predicting the local stabilities of specific regions of the structure of a protein in cases in which standard experimental methods of detection of hydrogen exchange may become of difficult applicability, either because of peak overlapping in NMR spectra, or because the exchange is too fast, as for example in the case of the extended flexible region including helix D (residues 85–105) in CheY (Figure 1). More studies, however, will be needed to establish the reliability of the predictions provided by the CamP method in these cases. The recent introduction of NMR techniques capable of measuring very rapid exchange reactions (Schanda and Brutscher, 2005) will enable additional data to be acquired and a better parameterization of the CamP method to be developed.

The CamP method provides a prediction of the coefficients of the Fourier transform of the ln$P$ profile, rather than of the profile itself. These coefficients are predicted with a neural network trained on a set of 2000 structures selected from the Protein Data Bank (Berman et al., 2000) according to their ClustalW alignment score (Thompson et al., 1994) with the query sequence; sequences with a score above 50% are discarded. The number of proteins taken from the database was chosen to be 2000 to optimize the convergence of the conjugate gradient algorithm during the network training (Nissen, 2003). In the fitting procedure, the protection factors of the selected structures are calculated by using the formula of Equation 1. To enable the neural network to identify unstructured regions, we include 100 random sequences generated with the amino acid composition of natively unfolded proteins in the training data set (Romero et al., 2001); the ln$P$ values associated with these sequences are assigned random values smaller than 5. To ensure that there is no overfitting, the protein data set is shuffled and divided into two parts containing the same number of protein structures (i.e., 1000 + 1000 proteins); the first part of the data set is used to train the network, and the second is used to test it. We searched for the optimal configuration of the neural network by varying the number of input and output neurons and by checking the correlation between calculated and predicted protection factors in the testing set. The amino acid composition of nonoverlapping segments of l residues of the protein is used as input. The size of the segments and the number of coefficients of the discrete Fourier transform were chosen by exploring values in the intervals 5–20 and 10–100, respectively. We found the best solution when the number of input neurons is $I = 20L/l$, where L is the length of the sequence and 20 is the length of the list of frequencies (one for each amino acid type) in a segment of l residues. The number of output neurons is $O = bL$, where $b = 0.75$. With these values we obtained an average correlation of ~90% for the training set and ~76% for the testing set; the correlation in the testing set is stable and larger than 55% if the number of output neurons is $O > 20$. Interestingly, the optimal length of 7 residues is also in agreement with an estimate based on an NMR analysis of unfolded states (Schwalbe et al., 1997), and with a survey of structural databases that showed that the average size of structured elements in a protein is typically 6–10 amino acids (Berndt, 1996). The learning rates, the connection rates, the number of hidden neurons, and the training error are set to 0.7, 1, $(I + O)/2$, and 0.0025, respectively. In addition, we divide the input neurons into two sets, according to the polar or nonpolar character of the amino acid linked to a particular neuron.

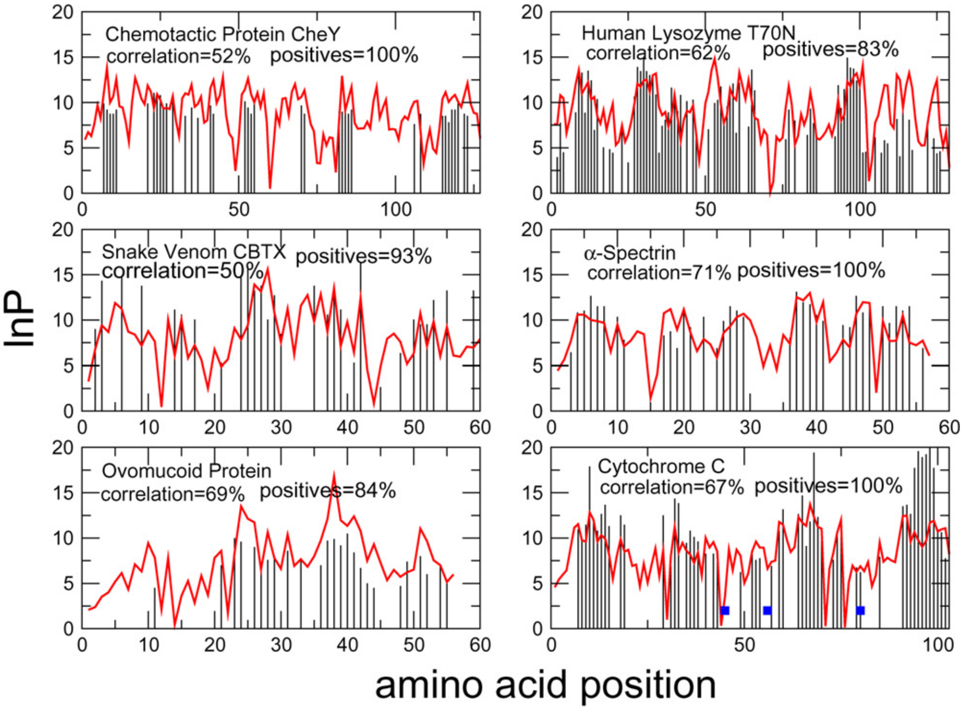

Figure 1. Comparison between Predicted and Experimental Protection Factors

CamP predictions (red lines) are presented for six proteins that have been characterized experimentally in detail (black bars) (Lacroix et al., 1997; Russell et al., 2003; Sivaraman et al., 2000; Sadqi et al., 1999; Wand et al., 1986; Swint-Kruse and Robertson, 1996). The percentage of positives is defined as the fraction of predicted and experimental protection factors that are found in the same ranges (ln$p$ < 5 for nonprotected residues; ln$p$ > 5 for protected residues); values not measured by experiments are excluded from counting. For cytochrome c, the proteolytic sites (Fontana et al., 1995) are shown as blue squares.

We find that the overall (i.e., training and testing set) correlation drops to 45% when neurons corresponding to nonpolar residues are deleted, while neglecting polar contributions gives a correlation of 55%. Thus, nonpolar residues give a slightly larger contribution to the ln$P$ profile.

We also present an alternative way to predict protection factors from the sequence of a protein. In this second method, a model for the structure is first generated from the sequence, and the protection factors are then predicted from the model structure by using Equation 1. Several procedures are available for the generation of structural models from the knowledge of the sequence, and in this work we used I-SITES/HMMSTR/ROSETTA (Bystroff and Shao, 2002). In all cases except one (the T70N mutational variant of human lysozyme), the predictions by CamP were better than those obtained by applying Equation 1 to the structures predicted by I-SITES/HMMSTR/ROSETTA (Table 1, columns C_{NE} and C_{RE}).

As crystallographic B factors are often used to infer the local flexibility of a folded state (Halle, 2002; Zoete et al., 2002), we analyzed the correlation (Table 1, column C_{BE}) between the experimental protection factors and the experimental B factors (Table 1, column BH). In all cases, the correlations between experimental protection factors and B factors are relatively weak, consistent with the view that protection factors mainly probe larger-amplitude fluctuations than B factors (Miller and Dill, 1987). It is also well known that the presence of proteolytic sites in proteins can be correlated with solvent exposure and flexibility of regions of 8–10 residues of the polypeptide chain (Hubbard, 1998; Fontana et al., 2004). In the case of horse cytochrome C, which has been subjected to a limited proteolysis study by thermolysin in 50% aqueous (v/v) TFE at neutral pH (Fontana et al., 1995) that identified a major cut at peptide bond 56–57 and additional but minor cleavages at peptide bonds 45–46 and 80–81, we compared the regions for which low protection factors are predicted or measured experimentally with the proteolytic sites. As expected, these proteolytic sites, which are in long loops between native helices, are in regions of low (ln$p$ < 6.5) protection factors (Figure 1).

It is also interesting to compare the prediction of protection factors with the predictions of intrinsic disorder (Table 1, columns C_{DE} and C_{PE}). The latter predictions should identify regions of the polypeptide chain that have a tendency to undergo significant structural fluctuations. The weak correlations found in this case suggest that the intrinsic propensity for being unfolded is strongly modulated by the interactions in the folded state to define the local fluctuations probed by hydrogen exchange measurements.

In conclusion, in this work we have shown that it is possible to predict with good accuracy protection factors directly from amino acid sequences. A web server for the computation of protection factors with the CamP method is available at the university of Cambridge (http://www-almost.ch.cam.ac.uk/camp.php). These results indicate that the intrinsic propensity to fold of different parts of the amino acid sequences are identifiable without any structural measurement, and thus may help the development of protein design as well as protein fold predictions methods. In addition, they are particularly interesting in light of the recent recognition that protein aggregation is often initiated by local structural fluctuations that expose regions of the sequence with a high propensity to form intermolecular interactions (Chiti and Dobson, 2006).

Supplemental Data

Supplemental Data include a comparison between predicted and experimental protection factors for six additional proteins and are available at http://www.structure.org/cgi/content/full/15/2/139/DC1/.

ACKNOWLEDGMENTS

We thank Robert Best, Jane Clarke, Chris Dobson, Joerg Gsponer, Amol Pawar, Danny Hsu, and Chris Waudby for discussions. This work was supported by the Swiss National Science Foundation, grant PBZHA-112735 (G.G.T.); the European Union (A.C.); the Leverhulme Trust (G.G.T. and M.V.); and the Royal Society (M.V.).

Received: July 13, 2006
Revised: December 5, 2006
Accepted: December 28, 2006
Published: February 13, 2007

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Nucleic Acids Res. *25*, 3389–3402.

Anfinsen, C. (1973). Science *181*, 223–230.

Bahar, I., Wallqvist, A., Covell, D.G., and Jernigan, R.L. (1998). Biochemistry *37*, 1067–1075.

Bai, Y., and Englander, S.W. (1996). Proteins *24*, 145–151.

Bai, Y.W., Milne, J.S., Mayne, L., and Englander, S.W. (1993). Proteins *17*, 75–86.

Berman, H.M., Westbrook, J., Feng, Z., Gillirand, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). Nucleic Acids Res. *28*, 235–242.

Berndt, K.D. (1996). Protein Secondary Structure (Stockholm: Karolinska Institute).

Best, R.B., and Vendruscolo, M. (2006). Structure *14*, 97–106.

Bystroff, C., and Shao, Y. (2002). Bioinformatics *18*, S54–S61.

Chamberlain, A.K., Handel, T.M., and Marqusee, S. (1996). Nat. Struct. Biol. *3*, 782–787.

Chiti, F., and Dobson, C.M. (2006). Annu. Rev. Biochem. *75*, 333–366.

Chu, R.A., Pei, W., Takei, J., and Bai, Y. (2002). Biochemistry *41*, 7998–8003.

Clarke, J., and Itzhaki, L.S. (1998). Curr. Opin. Struct. Biol. *8*, 112–118.

Dixon, R.D.S., Chen, Y., Ding, F., Khare, S.D., Prutzman, K.C., Shaller, M.D., Campbell, S.L., and Dokholyan, N.V. (2004). Structure *12*, 2161–2171.

Englander, S.W. (2000). Annu. Rev. Biophys. Biomol. Struct. *29*, 213–238.

Fersht, A.R. (1999). Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding (New York: W.H. Freeman & Co.).

Fontana, A., Zambonin, M., De Filippis, V., Bosco, M., and Polverino de Laureto, P. (1995). FEBS Lett. *362*, 266–270.

Fontana, A., Polverino de Laureto, P., Frare, B.S.E., Picotti, P., and Zambonin, M. (2004). Acta Biochim. Pol. *51*, 299–321.

Garcia, A.E., and Hummer, G. (1999). Proteins *36*, 175–191.

Goodman, E.M., and Kim, P.S. (1991). Biochemistry *30*, 11615–11620.

Gsponer, J., Hopearuoho, H., Whittaker, S.B.-M., Spence, G.R., Moore, G.R., Paci, E., Radford, S.E., and Vendruscolo, M. (2006). Proc. Natl. Acad. Sci. USA *103*, 99–104.

Halle, B. (2002). Proc. Natl. Acad. Sci. USA *99*, 1274–1279.

Hilser, V.J., and Freire, E. (1996). J. Mol. Biol. *262*, 765–772.

Hubbard, S.J. (1998). Biochim. Biophys. Acta *1382*, 191–206.

Hvidt, A., and Nielsen, S.O. (1966). Adv. Protein Chem. *21*, 287–386.

Itzhaki, L.S., Neira, J.L., and Fersht, A.R. (1997). J. Mol. Biol. *270*, 89–98.

Janin, J., and Chothia, C. (1976). J. Mol. Biol. *100*, 197–211.

Lacroix, E., Bruix, M., Lopez-Herandez, E., Serrano, L., and Rico, M. (1997). J. Mol. Biol. *271*, 472–487.

Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., and Russell, R.B. (2003). Structure *11*, 1453–1459.

Miller, D.W., and Dill, K.A. (1987). Protein Sci. *196*, 641–656.

Milne, J.S., Mayne, L., Roder, H., Wand, A.J., and Englander, S.W. (1998). Protein Sci. *7*, 739–745.

Morozova, L.A., Haynie, D.T., Arico-Meundel, C., Dael, H.V., and Dobson, C.M. (1995). Nat. Struct. Biol. *2*, 871–875.

Moult, J.R. (2005). Curr. Opin. Struct. Biol. *15*, 285–289.

Neira, J.L., Itzhaki, L.S., Otzen, D.E., Davis, B., and Fersht, A.R. (1997). J. Mol. Biol. *270*, 99–110.

Nissen, S. (2003). Implementation of a Fast Artificial Neural Network Library (FANN), http://leenissen.dk/fann/ (Copenhagen: University of Copenhagen, Department of Computer Science).

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., and Dunker, A.K. (2006). Proteins *61*(S7), 176–182.

Pandey, A., and Mann, M. (2000). Nature *405*, 837–846.

Perez, J.M., Renisio, J.G., Prompers, J.J., van Platerink, C.J., Cambillau, C., Darbon, H., and Frenken, L.G. (2001). Biochemistry *40*, 74–83.

Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., and Dunker, A.K. (2001). Proteins *42*, 38–48.

Russell, B.S., Zhong, L., Bigotti, M.G., Cutruzzola, F., and Bren, K.L. (2003). J. Biol. Inorg. Chem. *8*, 156–166.

Sadqi, M., Casares, S., Abril, M.A., Lapez-Mayorga, O., Conejero-Lara, F., and Freire, E. (1999). Biochemistry *38*, 8899–8906.

Schanda, P., and Brutscher, B. (2005). J. Am. Chem. Soc. *127*, 8014–8015.

Schueler-Furman, O., Wang, C., Bradley, P., Misura, K., and Baker, D. (2005). Science *310*, 638–642.

Schwalbe, H., Fiebig, K.M., Buck, M., Jones, J.A., Grimshaw, S.B., Spencer, A., Glaser, S.J., Smith, L.J., and Dobson, C.M. (1997). Biochemistry *36*, 8977–8991.

Sheinerman, F.B., and Brooks, C.L. (1998). Proc. Natl. Acad. Sci. USA *95*, 1562–1567.

Sivaraman, T., Kumar, T.K., Hung, K.W., and Yu, C. (2000). Biochemistry *39*, 8705–8710.

Swint-Kruse, L., and Robertson, A.D. (1996). Biochemistry *35*, 171–180.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). Nucleic Acids Res. *22*, 4673–4680.

Vendruscolo, M., Paci, E., Dobson, C.M., and Karplus, M. (2003). J. Am. Chem. Soc. *125*, 15686–15687.

Viguera, A.R., and Serrano, L. (2003). Proc. Natl. Acad. Sci. USA *100*, 5730–5735.

Vitkup, D., Melamud, E., Moult, J., and Sander, C. (2001). Nat. Struct. Biol. *8*, 559–566.

Wand, A.J., Roder, H., and Englander, S.W. (1986). Biochemistry *25*, 1107–1114.

Warshel, A., and Levitt, M. (1976). J. Mol. Biol. *106*, 421–437.

Wijesinha-Bettoni, R., Dobson, C.M., and Redfield, C. (2001). J. Mol. Biol. *307*, 885–898.

Woodward, C.K., and Hilton, B.D. (1980). Biophys. J. *32*, 561–575.

Zoete, V., Michielin, O., and Karplus, M. (2002). J. Mol. Biol. *315*, 21–52.