doi:10.1016/j.jmb.2005.02.068 J. Mol. Biol. (2005) 348, 507–512

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

COMMUNICATION

Protein Folding Rates Estimated from Contact Predictions

Marco Punta[1,2]* and Burkhard Rost[1,2,3]*

[1]*CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, USA*

[2]*Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St Nicholas Avenue, New York NY 10032, USA*

[3]*NorthEast Structural Genomics Consortium (NESG) Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, USA*

**Corresponding authors*

Folding rates of small single-domain proteins that fold through simple two-state kinetics can be estimated from details of the three-dimensional protein structure. Previously, predictions of secondary structure had been exploited to predict folding rates from sequence. Here, we estimate two-state folding rates from predictions of internal residue–residue contacts in proteins of unknown structure. Our estimate is based on the correlation between the folding rate and the number of predicted long-range contacts normalized by the square of the protein length. It is well known that long-range order derived from known structures correlates with folding rates. The surprise was that estimates based on very noisy contact predictions were almost as accurate as the estimates based on known contacts. On average, our estimates were similar to those previously published from secondary structure predictions. The combination of these methods that exploit different sources of information improved performance. It appeared that the combined method reliably distinguished fast from slow two-state folders.

Keywords: two-state proteins; folding rate; long-range order; contact predictions

Multi-state folding rate inversely proportional to protein length

Advances in the experimental and theoretical study of the dynamics of protein folding have improved our understanding of the phenomenon over the last few years.[1] Various theories and simulations suggest a surprisingly simple relation between the number of residues in a protein, its length L, and the rate at which it folds.[2–6] Basically, this relation is of the form:

$$\log(k_f) \propto C_1 L^{C_2} \quad (1)$$

where k_f is the experimental folding rate, L is the length of the protein, and C_1 and C_2 are simple constants. For proteins that fold through "transition states",[7,8] all values of C_2 between 0 and 1 give a good estimate for the multi-state transition through intermediates, i.e. the longer the protein the slower the transition.[9] In contrast to multi-state transitions, equation (1) does not hold for single-domain two-state folders, i.e. proteins that fold without intermediates directly into their native three-dimensional (3D) structure. In other words, protein length does not describe the transition rates of direct folding.

Two-state folding rates correlate with secondary structure

In solution, native secondary structure such as helices can form even in the unfolded state.[10] This

Abbreviations used: *LRO*, long-range order; PROFcon, system for prediction of residue–residue contact in single-chain proteins (unpublished results); PDB, Protein Data Bank; PROFphd, system for prediction of 1D structure; L, length of protein, i.e. number of its residues; LRO_{pred}, number of long-range contacts predicted by method introduced here; k_f, experimental folding rate; two-state folders, proteins that fold without intermediates; multi-state folders, proteins that fold through intermediate states.

E-mail addresses of the corresponding authors: punta@cubic.bioc.columbia.edu; rost@columbia.edu

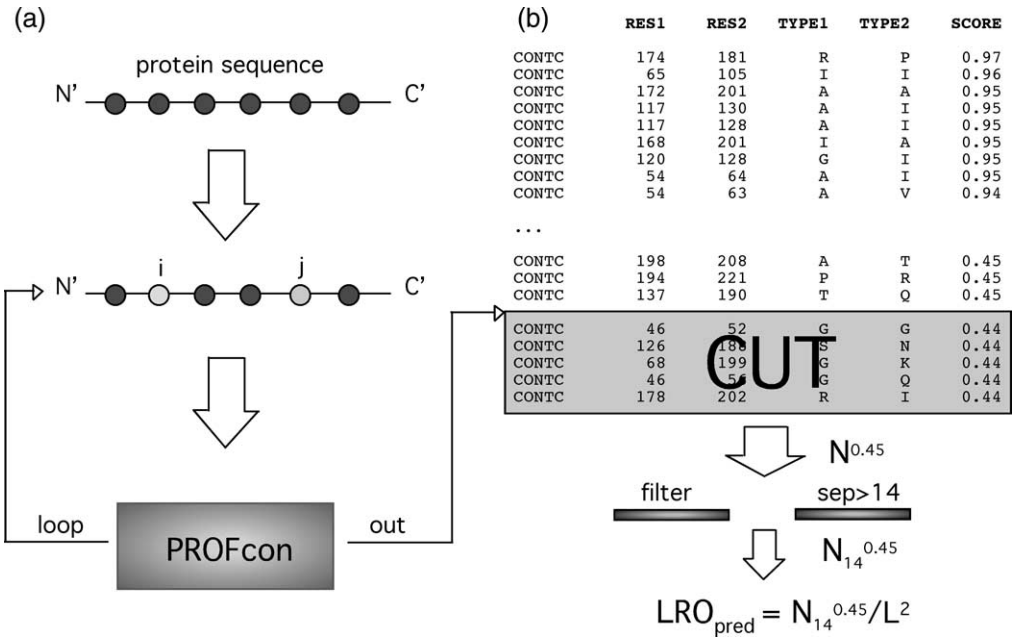

Figure 1. Sketch of underlying method. (a) PROFcon predicts the probability of a spatial contact between each pair of residues *ij* in the protein. The output of PROFcon is a number between 0 and 1, with scores closer to 1 indicating a higher probability for the pair to be in contact. Iterating over all possible residue pairs in the protein produces a list (b) of scores. By fixing a cut-off on the output score (we used 0.45), all pairs ranking below the cut-off were discarded. Next, pairs with sequence separations ≤ 14 were eliminated. The remaining number of pairs divided by L^2 (L is the protein length) was our estimate for the number of long-range contacts (LRO_{pred}; equation (4)).

implies that regular secondary structure might be a key player in determining the rate of folding. Indeed, two recent methods estimate folding rates directly from secondary structure content. George Rose and collaborators[11] observed that folding rates correlate very well with the overall secondary structure composition in three states (helix, strand, other) assigned from 3D co-ordinates through the programs DSSP[12] and PROSS.[11] Ivankov & Finkelstein[13] have introduced the concept of an "effective length of a folding chain" that is defined as the length of the protein L minus the number of residues in helical conformation, plus the number of helices; more precisely:[13]

$$L_{\mathrm{eff}} = L - L_{\mathrm{H}} + C_3 N_{\mathrm{H}}, \quad \text{and} \quad L_{\mathrm{eff}}^{\mathrm{P}} = (L_{\mathrm{eff}})^{\mathrm{P}} \qquad (2)$$

where L_{H} is the number of residues in helical conformation, N_{H} is the number of helices and $C_3 \geq 0$ is a simple constant parameter to be optimized. The effective length, taken to the power by any value between $P = 0.1$ and 0.7, correlates with two-state and multi-state protein folding rates. Remarkably, the folding rates correlate almost as well with the predicted (PSIPRED[14]) as with the observed (DSSP[12]) helical content. Folding rates can therefore be estimated directly from sequence, i.e. without explicit knowledge of experimental 3D structures.

Two-state folding rates correlate with long-range order (*LRO*)

Here, we introduce a new approach to the prediction of two-state protein folding rates from sequence alone. Our method relies on predictions of residue–residue contacts to tap into another correlation, first reported by Gromiha & Selvaraj,[15] namely that between folding rates and the long-range order (*LRO*).[15] *LRO* is defined as:

$$LRO = \frac{N_{12}}{L}, \quad N_{12} = \sum_{|i-j|>12} \delta_{ij}^{3\mathrm{D}}$$
$$\delta_{ij}^{3\mathrm{D}} = \begin{cases} 1, & \text{if } d_{ij} \leq 0.8\ \mathrm{nm} \\ 0, & \text{else} \end{cases} \qquad (3)$$

where L is the protein length, and N_{12} is the number of residues that are in spatial contact (d_{ij} is the spatial distance of C^{α} atoms) and are more than 12 sequence positions apart. Gromiha & Selvaraj[15] observed that choosing the sequence separation threshold to be exactly 12 residues resulted in the highest correlation between two-state folding rates and *LRO* for a set of 23 two-state folders. Not surprisingly, *LRO* anti-correlates with the helical composition because in helices many residues saturate their "contactability" through short-range contacts.

PROFcon accurately predicts the *LRO* from sequence

PROFcon is a neural network trained to predict intra-chain residue–residue contacts.[16] For each pair of internal residues *ij*, PROFcon predicts the probability that *i* and *j* are in spatial contact (closer than 0.8 nm for C^{β} atoms). One remarkable and

unexpected feature of PROFcon is that it can predict the overall number of contacts in a protein more accurately than any simple function (our unpublished results). In order to achieve this, we consider all the $N(T)$ contacts predicted above a threshold of T, i.e. the $N(T)$ most probable predictions (Figure 1; PROFcon is available online through PredictProtein†[17]). Using this protein-specific threshold in predicting contact maps also improves the contact predictions directly.[16] In the context of two-state folding rates, the relevant finding is that our predictions allow the distinction between two proteins that both have L residues but differ in their numbers of contacts. The difference between predicting the number of contacts in a protein through its length[18–21] and our method is crucial in this context because protein length correlates very poorly with two-state folding rates.[9] The next step was to combine our prediction for the number of contacts with the correlation between the observed number of long-range contacts in two-state folders (*LRO*) and their folding rates. We defined the following quantity:

$$LRO_{\mathrm{pred}} = N_S^T/L^2 \qquad (4)$$

where N_S^T is the number of pairs predicted by PROFcon with a score $\geqq T$ and separated by at least S sequence positions; L is the protein length. The normalization factor L^2, in contrast to L in equation (3), was chosen because the number of contacts predicted by the raw PROFcon networks is proportional to L^2.

There are two free parameters to be chosen in our number of predicted long-range contacts (LRO_{pred}; equation (4)), the sequence separation S and the threshold T in the probability of our PROFcon for considering long-range contacts. In this case, we simply chose $S=12$ in analogy to the optimal value found for the *LRO*[15] (equation (3)), and $T=0.5$, i.e. considered all residue pairs for which the PROFcon prediction for contact was higher than the prediction for non-contact. The number of long-range contacts predicted in this way (LRO_{pred}) correlated with the long-range order (*LRO*; Table 1). This correlation was significantly higher for shorter proteins. Different choices of S and T gave qualitatively similar results, i.e. the correlation was robust with our *ad hoc* choice. PROFcon performs better for shorter than for longer proteins; this may be the reason why the correlation between LRO_{pred} and *LRO* was higher for shorter proteins. Since most proteins experimentally known to fold directly (two-state transition) are short, this problem is not severely limiting our ability to estimate two-state folding rates.

Data set and parameter optimization

We used the set of 37 two-state folders introduced by Ivankov & Finkelstein.[13] These proteins are not sequence-unique, in fact, at *HSSP* values <0,[22–24] this set is reduced to 31 proteins. The results for the correlation are similar for the entire and the sequence-unique subset. Furthermore, homologous proteins may differ in their folding rates. For example, the SH3 domain in human Fyn (PDB[25,26] identifier, 1shf:A[27]) and the SH3 domain of the p85 alpha subunit of phosphatidylinositol 3-kinase (1pnj[28]) have similar sequence (*HSSP* value $=0.27$); however, their folding rates differ substantially: for 1shf_A $\log(k_f)=2.0$, and for 1pnj $\log(k_f)=-0.5$ (k_f is the experimentally derived folding rate). In order to simplify the comparison to the previous results,[13] we therefore reported our performance on the full data set. It is also important to note that one protein (acylphosphatase, 2acy[29]) was used to train PROFcon; 13 others were sequence-similar to proteins used for training. Removing all these proteins from our set of two-state folders did not alter any of the results discussed below (data not shown). Note, furthermore, that our re-capitulation of the method introduced by Ivankov & Finkelstein ($L_{\mathrm{eff}}^{\mathrm{P}}$; Table 2) was based on our secondary structure predictions from PROFphd[30–32] rather than on those from PSIPRED[14] and ALB[33] used by Ivankov & Finkelstein. Again, this technical detail appeared not to have altered any results, since the PROFphd predictions yielded results similar to those obtained by the methods used previously[13] (data not shown). The reported optimal estimates from LRO_{pred} were obtained for the following choices of the parameters: $S=14$ and $T=0.45$. For $L_{\mathrm{eff}}^{\mathrm{P}}$ (equation (2)), we used $P=0.1$ and $C_3=1$.

Table 1. Correlation between predicted long-range contacts and long-range order

L	N_{prot}	$R(LRO_{\mathrm{pred}}, LRO)$
≤ 150	199	0.69
150–250	211	0.53
250–400	226	0.49

Scores: L, sequence length (number of residues in protein) interval chosen to group data; N_{prot}, number of proteins in a sequence-unique subset of proteins from the PDB within the given length interval; $R(LRO_{\mathrm{pred}}, LRO)$, correlation between predicted long-range contacts and long-range order ($S=12$ and $T=0.5$; equation (4)). Data set: taken from the EVA version of the largest sequence-unique subset as of December 2003.[39,40] All proteins in the set have X-ray structures at resolutions <0.25 nm. No pair in the set has levels of sequence similarity with *HSSP* values >0[22,23] to any other protein (this corresponds to $<20\%$ sequence identity for long alignments).

Predicted long-range order correlates with folding rates

The "effective length" ($L_{\mathrm{eff}}^{\mathrm{P}}$; equation (2)) predicted the folding rates remarkably well with a correlation of 0.70 in the jack-knife test, which was almost as high from sequence alone as the correlation between *LRO* and folding rates from 3D structures (Table 2). Note for comparison that the back-check, i.e. the value obtained after fitting $L_{\mathrm{eff}}^{\mathrm{P}}$

† www.predictprotein.org

Table 2. Correlation between estimated and experimental folding rates

N_{prot}	$R(L^P_{eff}, K_f)$	$D(L^P_{eff}, K_f)$	$R(LRO, K_f)$	$D(LRO, K_f)$	$R(LRO_{pred}, K_f)$	$D(LRO_{pred}, K_f)$
37	0.70 (−0.74)	0.96	0.78 (−0.80)	0.81	0.61 (−0.68)	0.98
36	**0.68 (−0.74)**	**0.99**	**0.78 (−0.81)**	**0.80**	**0.74 (−0.78)**	**0.86**

Scores: N_{prot}, number of proteins; $K_f = \log(k_f)$, logarithm of the folding rate k_f; $R(x,K_f)$ correlations between estimated (x) and observed (K_f) logarithm of folding rate; $D(x,K_f)$ average differences from the actual K_f, e.g. $\sum_i |LRO_{pred}(i) - \log(k_f^i)|/N_{prot}$, where N_{prot} was the overall number of proteins in the dataset under consideration. Methods: $x = L^P_{eff}$ (equation (2)) is our implementation of Ivankov & Finkelstein,[13] $x = LRO$ the long-range order (equation (3)), and $x = LRO_{pred}$ our prediction of long-range contacts. Data set: all proteins were taken from a previous work;[13] lower rows give results for subsets of the first set. Values in parentheses are for back-check correlation, i.e. the values obtained by the fit using all proteins, rather than by determining the parameters from the fit on different proteins and testing on a protein left out (jack-knife). Note that values in parentheses most likely over-estimate performance; they are given for comparison with other work only.

to all experimental rates, was −0.74 as reported.[13] The correlation between our predicted long-range contacts LRO_{pred} and the folding rate was markedly lower (0.61 for jack-knife and −0.68 for back-check; Table 2). However, when considering the sum over the differences between estimate and predictions as a measure for the performance instead of the correlation then both the predicted effective length, L^P_{eff}, and the predicted long-range contacts, LRO_{pred}, reached rather similar levels (Table 2). For the 37 proteins the correlation between LRO_{pred} and L^P_{eff} reached 0.47; we observed a similar number (0.45) when testing the correlation between the two on a much larger data set of 199 proteins shorter than 150 residues that had been used to test our contact prediction method PROFcon[16] (same set as used for Table 1).

Results statistically significant

In order to establish that the correlation achieved by our method was not due to the small data set, we carried out two different tests. Firstly, we calculated the probability that a correlation of −0.68 (Figure 2) could be achieved by chance. Toward this end, we randomly assigned the 37 values of the LRO_{pred} to the 37 experimental folding rates (taken as logarithms) and calculated the correlation between these two sets (i.e. the random pairs of prediction/observation). We repeated this operation 10^6 times; the correlation was $>|0.68|$ only five times (absolute values). By this model, the probability for a correlation to exceed 0.68 in our data set therefore is 5×10^{-6}. Secondly, we estimated the standard error in our estimate for the correlation between our prediction and the observed folding rates by bootstrapping[34] the 37 pairs of predicted/observed rates. The average was −0.67 with a standard deviation of 0.12. Even the lower limit (the average minus standard plus the deviation, i.e. $-0.67+0.12=-0.55$) had a chance of being random of $<4.6\times10^{-4}$. Clearly then, the correlation between predicted and experimental folding rates was statistically significant.

Predictions more accurate for short proteins

The Lyme disease antigen Vlse of *Borrelia burgdorferi* (1l8w[35]) was an extreme outlier in the distribution of our predictions (Figure 2). This 341 residue protein is by far the longest protein in our dataset; the next longest was cyclophilin A (1lop[36]) with 164 residues, and the average over the entire set was 84 residues. Obviously, our method failed

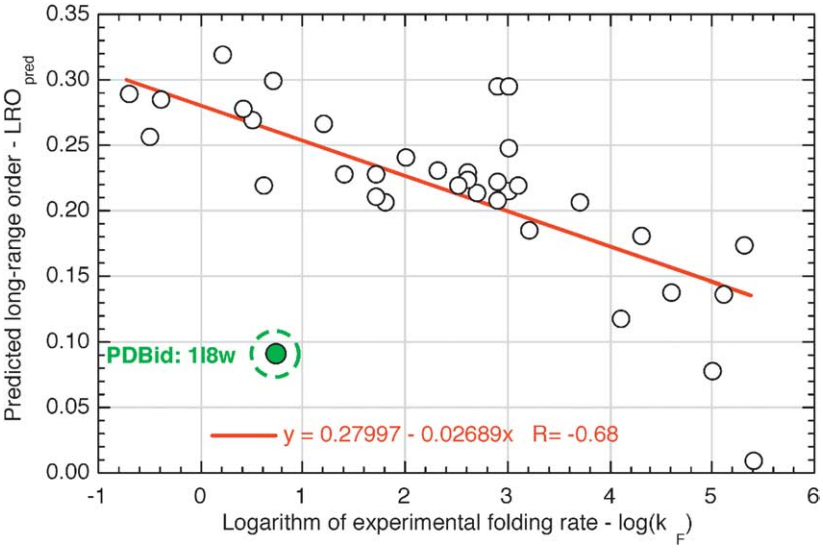

Figure 2. Regression line for the comparison of the predicted number of long-range contacts (LRO_{pred}; $S=14$ and $T=0.45$; equation (4)) and the logarithm of the observed two-state folding rates on a set of 37 two-state folders. The overall correlation coefficient was $R=-0.68$. The green circle labels the outlier, antigen Vlse (1l8w[35]).

for proteins much longer than the average domain length (around 100 residues[37,38]). Excluding this outlier left us with 36 proteins for which the LRO_{pred} (predicted long-range contacts) predicted folding rates more accurately than the $L^{\mathrm{P}}_{\mathrm{eff}}$ (effective length) measured both by correlation and mean deviation (Table 2, in bold). In fact, for these proteins our estimates from sequence alone were almost as accurate as the estimates from the full details of 3D structures (*LRO*). Although the helical content and *LRO* are related, we observed some degree of non-redundancy between the predictions based on contacts (LRO_{pred}) and those based on secondary structure ($L^{\mathrm{P}}_{\mathrm{eff}}$). By simply compiling the arithmetic average over both, we improved the estimate of folding rates to a jack-knife correlation of 0.73 (for all 37 proteins) and to a deviation sum of 0.89. In other words, the performance was better than that of any of the two individual methods that predicted two-state folding rates from sequence alone.

Implications for understanding folding?

Two-state folding rates are closely related to the content in local, regular secondary structure,[11] in particular to that in α-helices.[13] Our results seem to suggest that although the α-helical content is crucial for determining two-state folding rates, some other mechanisms might play an important role. The extreme argument in point is highlighted by the observation that, when considering two-state folders that have a significant content of beta strands (i.e. all-beta; alpha/beta and alpha+beta; 27 proteins in our dataset) the correlation between the effective length (equation (2)) and the folding rate becomes insignificant (0.13 in a jack-knife experiment), while the correlation between the long-range order and the folding rates remains considerable (>0.5 in a jack-knife experiment) for both the lookup from 3D structures (equation (3)) and for the prediction from sequence (equation (4)). Do our results then favor any model of folding over any other? We believe that our evidence was not clear and conclusive enough to answer that question in the affirmative.

Conclusions

We did not find new evidence concerning the question of what are the determinants of two-state folding rates. However, we have shown that estimates from local secondary structure and long-range contacts both somehow contribute independent information in a predictive sense. Our estimates are based on contact predictions that in turn rely mostly on local sequence features. Therefore, our results do not clearly falsify the assumption that folding rates are determined largely by local factors. Most importantly, even methods that predict internal residue–residue contacts at seemingly low levels of accuracy contain enough relevant information to predict two-state folding rates almost as well as the entirely correct experimentally observed contact map. We therefore challenge the suggestion that *de novo* predictions of inter-residue contact maps have been significantly under-appreciated.

Acknowledgements

Thanks to Jinfeng Liu and Megan Restuccia (both Columbia) for computer assistance, to Guy Yachdav (Columbia) for integrating the program into an Internet server, and to Dariusz Przybylski (Columbia) and Murad Nayal (Columbia) for important discussions. Thanks to Dmitry N. Ivankov and Alexey V. Finkelstein (both Institute of Protein Research, Pushchino) for providing us with crucial data. This work was supported by the grants RO1-GM64633-01 from the National Institutes of Health (NIH) and RO1-LM07329-01 from the National Library of Medicine (NLM). Last, but not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases.

References

1. Mirny, L. & Shakhnovich, E. (2001). Protein folding theory: from lattice to all-atom models. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 361–396.
2. Finkelstein, A. V. & Badretdinov, A. (1997). Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. *Fold. Des.* **2**, 115–121.
3. Galzitskaya, O. V., Ivankov, D. N. & Finkelstein, A. V. (2001). Folding nuclei in proteins. *FEBS Letters*, **489**, 113–118.
4. Thirumalai, D. (1995). From minimal models to real proteins:time scales for protein folding kinetics. *J. Phys.* **5**, 1457–1469.
5. Gutin, A. M., Abkevich, V. V. & Shakhnovich, E. I. (1996). Chain length scaling of protein folding time. *Phys. Rev. Letters*, **77**, 5433–5436.
6. Koga, N. & Takada, S. (2001). Roles of native topology and chain-length scaling in protein folding: a simulation study with a Go-like model. *J. Mol. Biol.* **313**, 171–180.
7. Ewbank, J. J. & Creighton, T. E. (1992). Protein folding by stages. *Curr. Opin. Struct. Biol.* **2**, 347–349.
8. Ewbank, J. J., Creighton, T., Hayer-Hartl, M. K. & Hartl, F. U. (1995). What is the molten globule? *Nature Struct. Biol.* **2**, 10.
9. Galzitskaya, O. V., Garbuzynskiy, S. O., Ivankov, D. N. & Finkelstein, A. V. (2003). Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. *Proteins: Struct. Funct. Genet.* **51**, 162–166.
10. Prieto, J. & Serrano, L. (1997). C-capping and helix stability: the Pro C-capping motif. *J. Mol. Biol.* **274**, 276–288.
11. Gong, H., Isom, D. G., Srinivasan, R. & Rose, G. D. (2003). Local secondary structure content predicts folding rates for simple, two-state proteins. *J. Mol. Biol.* **327**, 1149–1154.

12. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
13. Ivankov, D. N. & Finkelstein, A. V. (2004). Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 8942–8944.
14. Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202.
15. Gromiha, M. M. & Selvaraj, S. (2001). Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. *J. Mol. Biol.* **310**, 27–32.
16. Punta, M. & Rost, B. (2005). Toward good 2D predictions in proteins. *Bioinformatics*. In the press.
17. Rost, B., Yachdav, G. & Liu, J. (2004). The PredictProtein server. *Nucl. Acids Res.* **32**, W321–W326.
18. Goebel, U., Sander, C., Schneider, R. & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Struct. Funct. Genet.* **18**, 309–317.
19. Olmea, O. & Valencia, A. (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold. Des.* **2**, S25–S32.
20. Olmea, O., Rost, B. & Valencia, A. (1999). Effective use of sequence correlation and conservation in fold recognition. *J. Mol. Biol.* **293**, 1221–1239.
21. Fariselli, P., Olmea, O., Valencia, A. & Casadio, R. (2001). Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins: Struct. Funct. Genet.* **Suppl**, 157–162.
22. Sander, C. & Schneider, R. (1991). Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.
23. Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94.
24. Mika, S. & Rost, B. (2003). UniqueProt: creating representative protein sequence sets. *Nucl. Acids Res.* **31**, 3789–3791.
25. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R. *et al*. (1977). The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.
26. Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K. *et al*. (2002). The Protein Data Bank. *Acta Crystallog. sect. D*, **58**, 899–907.
27. Noble, M. E., Musacchio, A., Saraste, M., Courtneidge, S. A. & Wierenga, R. K. (1993). Crystal structure of the SH3 domain in human Fyn; comparison of the three-dimensional structures of SH3 domains in tyrosine kinases and spectrin. *EMBO J.* **12**, 2617–2624.
28. Booker, G. W., Gout, I., Downing, A. K., Driscoll, P. C., Boyd, J., Waterfield, M. D. & Campbell, I. D. (1993). Solution structure and ligand-binding site of the SH3 domain of the p85 alpha subunit of phosphatidylinositol 3-kinase. *Cell*, **73**, 813–822.
29. Thunnissen, M. M., Taddei, N., Liguri, G., Ramponi, G. & Nordlund, P. (1997). Crystal structure of common type acylphosphatase from bovine testis. *Structure*, **5**, 69–79.
30. Rost, B. (2005). How to use protein 1D structure predicted by PROFphd. In *The Proteomics Protocols Handbook* (Walker, J. E., ed.), pp. 879–908, Humana, Totowa, NJ.
31. Rost, B. (2001). Protein secondary structure prediction continues to rise. *J. Struct. Biol.* **134**, 204–218.
32. Rost, B. (1996). PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol.* **266**, 525–539.
33. Ptitsyn, O. B. & Finkelstein, A. V. (1983). Theory of protein secondary structure and algorithm of its prediction. *Biopolymers*, **22**, 15–25.
34. Efron, B.; & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall/CRC, Boca Raton, FL.
35. Eicken, C., Sharma, V., Klabunde, T., Lawrenz, M. B., Hardham, J. M., Norris, S. J. & Sacchettini, J. C. (2002). Crystal structure of Lyme disease variable surface antigen VlsE of *Borrelia burgdorferi*. *J. Biol. Chem.* **277**, 21691–21696.
36. Konno, M., Ito, M., Hayano, T. & Takahashi, N. (1996). The substrate-binding site in *Escherichia coli* cyclophilin A preferably recognizes a *cis*-proline isomer or a highly distorted form of the *trans* isomer. *J. Mol. Biol.* **256**, 897–908.
37. Liu, J. & Rost, B. (2004). CHOP proteins into structural domains. *Proteins: Struct. Funct. Genet.* **55**, 678–688.
38. Liu, J., Hegyi, H., Acton, T. B., Montelione, G. T. & Rost, B. (2005). Automatic target selection for structural genomics on eukaryotes. *Proteins: Struct. Funct. Genet.* **56**, 188–200.
39. Eyrich, V., Martí-Renom, M. A., Przybylski, D., Fiser, A., Pazos, F., Valencia, A. *et al*. (2001). EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics*, **17**, 1242–1243.
40. Koh, I. Y. Y., Eyrich, V. A., Marti-Renom, M. A., Przybylski, D., Madhusudhan, M. S., Narayanan, E. *et al*. (2003). EVA: evaluation of protein structure prediction servers. *Nucl. Acids Res.* **31**, 3311–3315.

Edited by M. Levitt

(Received 29 November 2004; received in revised form 4 February 2005; accepted 15 February 2005)