15. Saxton, W. M. *et al. J. Cell Biol.* **99,** 2175–2186 (1984).
16. Blose, S. H., Meltzer, D. I. & Feramisco, J. R. *J. Cell Biol.* **98,** 847–858 (1984).
17. Wehland, J. & Willingham, M. C. *J. Cell Biol.* **97,** 1476–1490 (1983).
18. Cowan, N. J., Dobner, P. R., Fuchs, E. V. & Cleveland, D. W. *Molec. cell. Biol.* **3,** 1738–1745 (1983).
19. Ponstingl, H., Krauhs, E., Little, M. & Kempf, T. *Proc. natn. Acad. Sci. U.S.A.* **78,** 2757–2761 (1981).
20. Lemischka, I. R., Farmer, S. R., Racaniello, V. R. & Sharp, P. A. *J. molec. Biol.* **150,** 101–120 (1981).
21. Elliott, E. M., Henderson, G., Sarangi, F. & Ling, V. *Molec. cell. Biol.* **6,** 906–913 (1986).
22. Lopata, M. A. & Cleveland, D. W. *J. Cell Biol.* **105,** 11707–11720 (1987).
23. Sullivan, K. F. *A. Rev. Cell Biol.* **4,** 687–716 (1988).
24. Joshi, H. C. & Cleveland, D. W. *J. Cell Biol.* **109,** 663–673 (1989).
25. Smith, P. K. *et al. Analyt. Biochem.* **150,** 76–85 (1985).

ACKNOWLEDGEMENTS. We thank B. Oakley and Y. Zheng for *Drosophila* and human γ-tubulin cDNAs, J. Wood for access to his microinjection equipment, D. Aitken of Carl Zeiss for advice and assistance, K. Wallis and D. Murphy for purified tubulin, I. Szilak for the sequence of the mouse γ-tubulin cDNA clone, D. Compton for human pancreatic epithelial cells (CF-PAC1), W. Sale for discussion and E. Chávez for preparing the manuscript. This work has been supported by grants to H.C.J. from the American Cancer Society, the NIH, and the University Research Committee of Emory University and by a grant to D.W.C. from the NIH.

Assessment of protein models with three-dimensional profiles

Roland Lüthy*, James U. Bowie† & David Eisenberg†

The Molecular Biology Institute and Department of Chemistry and Biochemistry, UCLA, Los Angeles, California 90024–1570, USA

As methods for determining protein three-dimensional (3D) structure develop, a continuing problem is how to verify that the final protein model is correct. The revision of several protein models to correct errors[1–6] has prompted the development of new criteria for judging the validity of X-ray[7–9] and NMR[10,11] structures, as well as the formation of energetic[12–14] and empirical methods[15,16] to evaluate the correctness of protein models. The challenge is to distinguish between a mistraced or wrongly folded model, and one that is basically correct, but not adequately refined. We show that an effective test of the accuracy of a 3D protein model is a comparison of the model to its own amino-acid sequence, using a 3D profile[16], computed from the atomic coordinates of the structure 3D profiles of correct protein structures match their own sequences with high scores. In contrast, 3D profiles for protein models known to be wrong score poorly. An incorrectly modelled segment in an otherwise correct structure can be identified by examining the profile score in a moving-window scan. The accuracy of a protein model can be assessed by its 3D profile, regardless of whether the model has been derived by X-ray, NMR or computational procedures.

The method, outlined in Fig. 1, measures the compatibility of a protein model with its sequence, using a 3D profile. Each residue position in the 3D model is characterized by its environment[17] and is represented by a row of 20 numbers in the profile. These numbers are the statistical preferences (called 3D–1D scores) of each of the 20 amino acids for this environment. Environments of residues are defined by three parameters: the area of the residue that is buried; the fraction of side-chain area that is covered by polar atoms (O and N); and the local secondary structure. The 3D profile score S for the compatibility of the sequence with the model is the sum, over all residue positions, of the 3D–1D scores for the amino-acid sequence of the protein. As described below, the compatibility of segments of the sequence with their 3D structures can be assessed by plotting, against sequence number, the average 3D–1D score in a window of 21 residues.

For 3D protein models known to be correct, the 3D profile score S for the amino-acid sequence of the model is high. This is illustrated in Fig. 2 where the scores of all very well determined structures are indicated by large dots. By contrast, the profile score S for the compatibility of a wrong 3D protein model with its sequence is often low, as shown by the seven squares in Fig. 2 and discussed below. The 3D profile scores for four models that are largely misfolded all fall below the dotted line. Three other faulty structures that contain some correct as well as incorrect segments have 3D profile scores that fall between the two lines.

The profile score of a model depends on its size and its validity. Profile scores of correct models increase with the length of the protein, simply because more positive residue preferences are added into the sum. Small proteins may also score somewhat less than large ones because they have a smaller fraction of internal residues and hence give less informative profiles. Structures of the same length determined by X-ray and NMR tend to have comparable scores. The scores for computationally based models vary, depending on their accuracy. The deliberately misfolded models of Novotny *et al.*[12] have low scores because the environments of residues in the incorrect 3D structures are not compatible with the residues in the corresponding positions of the sequence. By contrast, models based on structures having

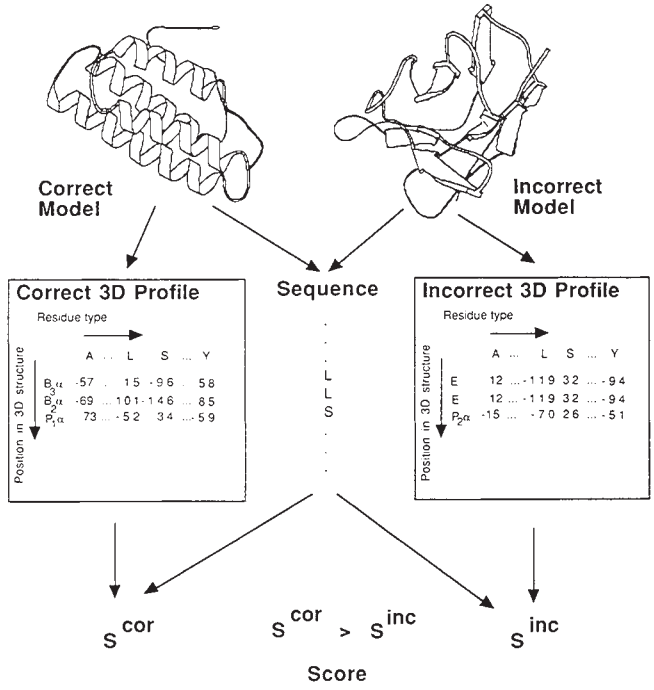

FIG. 1 The use of 3D profiles to verify protein models, illustrated with the correct and misfolded structures of haemerythrin, both having 113 residues. The model on the left is the X-ray-derived structure[29]. A 3D profile calculated from its coordinates matches the sequence of haemerythrin with a score $S^{cor}=38$. The right hand model is the misfolded haemerythrin model of Novotny *et al.*[12]. A 3D profile calculated from it matches its sequence poorly, with score $S^{inc}=15$. The actual profile consists of 113 rows (one for each position of the folded protein). In this schematic example only 3 rows are shown, those for positions 33 (where the residue is L), 34 (L) and 35 (S) (single-letter amino-acid code). The first column of the profile gives the environmental class of the position, computed from the coordinates of the model[17], and the next 20 columns give the amino-acid preferences (called 3D–1D scores) at that position. In this schematic example, there are only four columns of 3D–1D scores shown, those for residues, A, L, S and Y. In the correct profile on the left, position 33 is computed to be in the buried polar α-helical class ($B_3\alpha$); positions 34 and 35 are computed to be in the buried moderately polar α-helical class ($B_2\alpha$) and the partially buried α-helical polar class, $P_1\alpha$, respectively. The scores for the residues L, L and S for these three positions are 15, 101 and 34. The profile for the misfolded model assigns positions 33, 34 and 35 to the environmental classes E, E and $P_2\alpha$, giving 3D–1D scores for the residues L, L and S of −119, −119 and 26, respectively. That is, in the incorrect structure, the leucine residues are exposed, giving low 3D–1D scores and leading to a summed total score S^{inc} which is much smaller than S^{cor} when all other 3D–1D scores are summed along with the three shown here.

† To whom correspondence should be addressed.
* Present address: Swiss Institute for Experimental Cancer Research, Ch. des Boveresse 155, 1066 Epalinges s/Lausanne, Switzerland.

TABLE 1 3D profile scores of correct and incorrect/initial protein models with their own amino-acid sequences

Model	Correct	Incorrect/initial	Length (residues)
Rubisco small subunit	55.0 (3RUB)*†	15.1 (ref. 23)	123
Ferredoxin of *A. vinelandii*	45.9 (4FD1)* (ref. 2)	10.8 (2FD1) (ref. 24)	106
Ig κ V-region	47.0‡	6.9‡	113
Haemerythrin	37.9‡	14.9‡	113
ras p21	79.8(2P21)* (ref. 5)	50.5 (ref. 25)	177
EcoRI	116.5 (ref. 6)	89.3 (ref. 26)	261
HIV protease	53.9(3HVP)*	33.7 (ref. 27)	97

Notice that correct models match their own sequences with higher scores than do incorrect models. Scores of incorrect/initial models are shown in Fig. 2 by squares.

* Brookhaven Protein Data Bank codes[28].

† P. Curmi *et al.*, manuscript in preparation.

‡ Energy minimized models, described in ref. 12.

closely related amino-acid sequences, such as those for cyclic-AMP-dependent protein kinase (Brookhaven Protein Data Bank model 2APK)[18], insulin-like growth factor (1GF1, 2GF1)[19], apolipoprotein D (1APD)[20], and protein C inhibitor (2PAI)[21], have high scores comparable to many X-ray and NMR models. This contrast suggests that 3D profiles can distinguish between correct and misfolded models, however developed, and will be useful in assessing computational models and *de novo* protein designs. If this is so, then none of the three δ haemolysin models[22] deposited with the Brookhaven Protein Data Bank (1DHL; the lowest circles of Fig. 2) is correct.

Seven examples of profiles from misfolded models are given in Table 1 and illustrated by profile window plots in Fig. 3. One example is that of the small subunit of Rubisco, which was traced essentially backwards from a poor electron-density map[23]. The profile calculated from this mistraced model gives a score of only 15 when matched to the sequence of the small subunit of Rubisco. This score is well below the value of about 58 expected for a correct structure of this length (123 residues). In fact, the profile for the correct model of the small subunit of tobacco Rubisco (P. Curmi, H. Schreuder, D. Cuscio, R. Sweet and D.E., manuscript in preparation) matches its sequence with a score of 55. Also, when the score for the mistraced small subunit is plotted as a function of the sequence, as in the profile window plot of Fig. 3, the average score is often below the value of 0.1 and dips below zero at two points. As a control, 71 well-refined X-ray structures (having *R*-factors below 0.20, and resolutions of 2 Å or finer) were also examined by profile window plots. None of these profile window plots dipped below zero. Other faulty models whose profile window plots are shown in Fig. 3 include the mistraced ferredoxin from *Azotobacter vinelandii*[24], the deliberately misfolded protein models of Novotny *et al.*[12], the initial model of *ras* p21 protein, which had two β strands interchanged[25], the initial model of *Eco*RI endonuclease[26] which had several improperly built segments, and the initial model of the human immunodeficiency virus (HIV) protease[27] which was improperly interpreted at the C terminus. All of the incorrect models have profile window plots that are largely or completely below the plots for the correct models, and in all cases the plots for the incorrect models signal a problem by approaching or falling below a score of zero.

Of the examples of Fig. 3, *ras* p21 presents the most interesting case of profile window plots. In the initial model of *ras* p21 protein[25], β strands 1 and 3 had been interchanged along with associated loops, corresponding to changes in the vicinity of residues 1–11 and 46–65. The window plot reveals the problem in the region 46–65, but because of the averaging procedure, misses the problem at the N terminus. Also the initial model contained a one-residue misregistration between residues 98–107, and this problem is reflected by the dip in the average score in this region. Ramachandran diagrams are not so effective in revealing the faulty segments: in the initial structure, 33% of all non-glycine residues lie outside the allowed regions, but only 32% of non-glycine residues in the faulty segments lie outside. Similarly for the deliberately misfolded V-region domain[12] of Table 1 and Fig. 3, the Ramachandran diagram appears normal, even though the structure is completely wrong.

Our experience with profile assessment of faulty models can be summarized as follows. Models that are largely incorrect can often be detected by low overall profile scores (Table 1; Fig. 2).

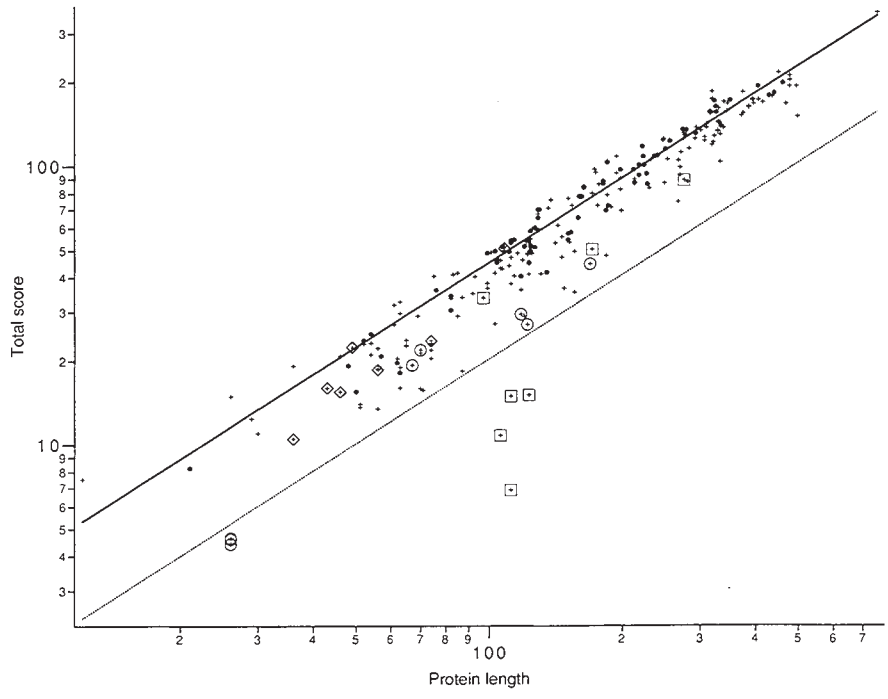

FIG. 2 3D profile scores (+) for protein coordinate sets in the Brookhaven protein data bank[28], as a function of sequence length, on a log–log scale. All + signs that are not enclosed represent X-ray determinations. ●, Scores for highly refined X-ray determinations; these are structures determined at resolutions of at least 2Å and with *R*-factors <20%. The heavy line is fit by least squares to these highly refined structures. ⊕, Scores for NMR structures; ⊕, scores for computationally determined models; ⊞, misfolded structures: these correspond to the entries of Table 1 and also the lower curves of Fig. 3. Notice that severely misfolded structures and some computational models have scores below $0.45 \times S^{calc}$ shown by the dotted line. Environmental classes for 3D profiles of oligomeric proteins were generally computed from oligomeric structures, rather than protomers. The difference is that the accessible surface areas of residues positioned at interfaces are greater for the protomers, producing a poorer fit of the profile to the sequence. This is insignificant for large structures, but important for small structures. As an extreme case, the profile for the 12-residue designed protein α1 (ref. 30) (the + at the left of the figure) matches its sequence only when the molecule is surrounded by its neighbours as in the crystal structure. Profile scores are plotted for coordinate sets in the January 1991 release (most recent entry when multiple entries are present) plus NMR coordinates in the April 1991 release.

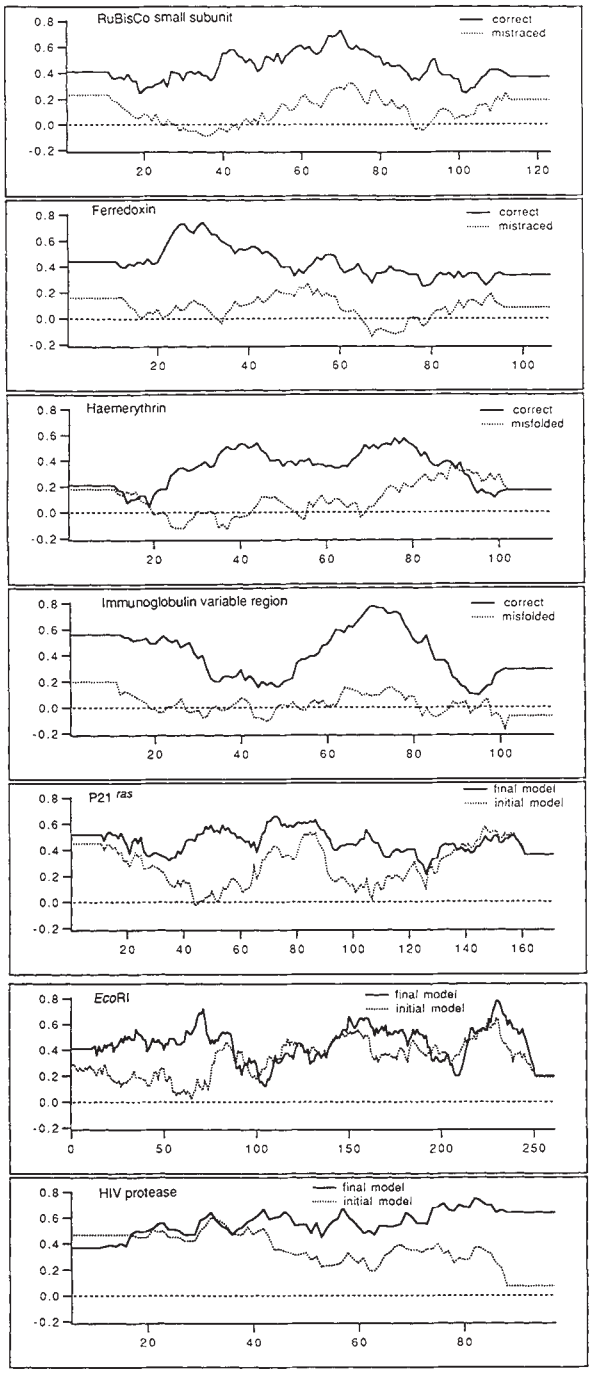

FIG. 3 Profile window plots for various incorrect and partially incorrect protein models, compared with window plots for their correct counterparts. The vertical axis gives the average 3D–1D score for residues in a 21-residue sliding window, the centre of which is at the sequence position indicated by the horizontal axis. Scores for the first 9 and the final 9 sequence positions have no meaning A window length of 21 residues strikes a useful balance between smoothing fluctuations and localizing the error.

Models that contain a small number of improperly built segments can be detected by low scoring regions in a profile window plot (Fig. 3). But not all faulty regions are always evident directly from the profile, particularly if the misbuilt regions are at the termini, where they are obscured by the windowing procedure. In practice, however, once a profile reveals any improperly built region in a protein, reinspection of the model may uncover other problems. In short, Figs 2 and 3 suggest that a correct protein model can be verified by the compatibility of its 3D profile with its sequence, and that a misfolded model can often be detected by its incompatibility.

An advantage of using 3D profiles for testing models is that the profiles have not themselves been used in the determination of the structure. In X-ray analysis, the final model is generally obtained by systematic variation of the model, forced by minimization of a function closely related to the *R*-factor. The *R*-factor is then also used as a criterion for assessing the accuracy of the model. Thus in essence, the model is judged by a test that is inherent in its formulation. Similarly, in construction of models by molecular mechanics and molecular dynamics, the final model is arrived at by minimization of energy. If no experimental structure is available for comparison, energetic analysis enters into both the determination and the evaluation of the structure. By contrast, the 3D profile method for model evaluation proposed here relies only on a comparison of the model to its own amino-acid sequence. Thus the present method not only seems effective in assessing protein models, but it also avoids the circularity that is to some extent inherent in the common methods of evaluation of 3D protein models. □

Received 8 November; accepted 4 December 1991.

1. Knight, S., Andersson, I. & Brändén, C. I. *Science* **244,** 702–705 (1989).
2. Stout, G. H., Turley, S., Sieker, L. C. & Jensen, L. H. *Proc. natn. Acad. Sci. U.S.A.* **85,** 1020–1022 (1988).
3. Lebioda, L., Stec, B. & Brewer, J. M. *J. biol. Chem.* **264,** 3685–3693 (1989).
4. Wlodawer, A. *et al. Science* **245,** 616–621 (1989).
5. Tong, L., Milburn, M. V., de Vos, A. M. & Kim, S. H. *Science* **245,** 244 (1989).
6. Kim, Y., Grable, J. C., Love, R., Greene, P. J. & Rosenberg, J. *Science* **249,** 1307–1309 (1990).
7. Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. *Acta crystallogr.* **A47,** 110–119 (1991).
8. Bränden, C.-I. & Jones, T. A. *Nature* **343,** 687–689 (1990).
9. Brünger, A. T. *Nature* **355,** 472–475 (1992).
10. Gonzalez, J. A., Rullmann, C., Bonvin, A. M. J. J., Boelens, R. & Kaptein, R. *J. magn. Reson.* **91,** 659–664 (1991).
11. Nilges, M., Habazettl, J., Brünger, A. T. & Holak, T. A. *J. molec. Biol.* **219,** 499–510 (1991).
12. Novotny, J., Bruccoleri, R. & Karplus, M. *J. molec. Biol.* **177,** 787–818 (1984).
13. Eisenberg, D. & McLachlan, A. D. *Nature* **319,** 199–203 (1986).
14. Novotny, J., Rashin, A. A. & Bruccoleri, R. *Proteins struct. funct. Genet.* **4,** 19–30 (1988).
15. Baumann, G., Frömmel, C. & Sander, C. *Protein Eng.* **2,** 329–334 (1989).
16. Hendlich, M. *et al. J. molec. Biol.* **216,** 167–180 (1990).
17. Bowie, J. U., Lüthy, R. & Eisenberg, D. *Science* **253,** 164–170 (1991).
18. Weber, I. T., Steitz, T. A., Bubis, J. & Taylor, S. S. *Biochemistry* **26,** 343–351 (1987).
19. Blundell, T. L., Bedarkar, S. & Humbel, R. E. FASEB J. **42,** 2592 (1983).
20. Peitsch, M. C. & Boguski, M. S. *New Biol.* **2,** 197–206 (1990).
21. Kuhn, L. A. *et al. Proc. natn. Acad. Sci. U.S.A.* **87,** 8506–8510 (1990).
22. Raghunathan, G., Seetharamulu, P., Brooks, B. R. & Guy, H. R. *Proteins struct. funct. Genet.* **8,** 213–225 (1990).
23. Chapman, M. S. *et al. Science* **241,** 71–74 (1988).
24. Gosh, D. *J. molec. Biol.* **158,** 73–109 (1982).
25. de Vos, A. M. *et al. Science* **239,** 888–893 (1988).
26. McClarin, C. *et al. Science* **234,** 1526 (1986).
27. Navia, M. *et al. Nature* **347,** 615–620 (1989).
28. Bernstein, F. C. *et al. J. molec. Biol.* **112,** 535–542 (1977).
29. Stenkamp, R. E., Sieker, L. C. & Jensen, L. H. *Acta crystallogr.* **B38,** 784–792 (1978).
30. Hill, C. P., Anderson, D. H., Wesson, L., DeGrado, W. F. & Eisenberg, D. *Science* **249,** 543–546 (1990).

ACKNOWLEDGEMENTS. We thank S.-H. Kim, P. Fitzgerald, J. Rosenberg and J. Novotny for undeposited coordinates and discussions. We thank the NIH for financial support.