

Crystal structure of a dimeric chymotrypsin inhibitor 2 mutant containing an inserted glutamine repeat

YU WAI CHEN^{*†}, KELVIN STOTT^{*†}, AND MAX F. PERUTZ^{‡§}

^{*}Centre for Protein Engineering and [‡]Laboratory of Molecular Biology, Medical Research Council Centre, Hills Road, Cambridge CB2 2QH, United Kingdom

Contributed by Max F. Perutz, December 14, 1998

ABSTRACT We have constructed mutants of chymotrypsin inhibitor 2 with short glutamine repeats inserted into its inhibitory loop. These mutants oligomerize when expressed in *Escherichia coli*. The dimer of a mutant with four glutamines now has been crystallized, and its structure has been solved by molecular replacement by using the wild-type monomer as a search model. The structure of each half of the dimer is found to be the same as that of the wild-type monomer, except around the glutamine insertion. It was proposed that the components of the oligomers are held together by hydrogen bonds between the main-chain and side-chain amides of the glutamine repeats. Instead, they appear to form by swapping domains on folding in *E. coli*, and the glutamine repeats connecting the components of the dimers are disordered.

Over the past 5 years, a growing number of dominantly inherited neurodegenerative diseases, including Huntington's disease and several different forms of spinocerebellar ataxia, have been linked to abnormally expanded CAG repeats in the coding regions of certain genes. The diseases are collectively known as polyglutamine-expansion diseases because the CAG repeats code for stretches of polyglutamine in the affected proteins. Progressive neurodegeneration begins in late adulthood if a single glutamine repeat exceeds a critical length, usually ≈ 40 residues, and begins earlier and becomes more severe as the repeat gets even longer in successive generations of affected individuals. The proteins causing the different diseases are unrelated except for the glutamine repeats. It is likely, therefore, that the diseases share a common pathogenic mechanism that depends only on the presence of an abnormally expanded glutamine repeat and not on its protein context. Together with the autosomal dominant genetics of the diseases, this mechanism implies that abnormally expanded glutamine repeats are toxic. Several groups have confirmed this by demonstrating the cytotoxicity of long glutamine repeats in various model systems (1, 2).

Neither the normal function of glutamine repeats in proteins nor the mechanism by which their abnormally expanded counterparts cause progressive neurodegeneration is known. Perutz *et al.* proposed that glutamine repeats in proteins can associate with each other by forming polar zippers in which β -strands are held together by hydrogen bonds between their main-chain and side-chain amides (3). They went on to suggest that the threshold effect of the polyglutamine-expansion diseases might reflect a delicate balance of two counteracting factors that determine whether a stable polar zipper is formed: the entropy gain or loss of the glutamines versus that of the solvent molecules. Abnormally expanded glutamine repeats form stable polar zippers because the increase in entropy caused by the liberation of many bound water molecules offsets the entropy loss caused by the formation of a rigid secondary structure, and this leads to the proteins' aggregation, which

causes progressive neurodegeneration; by contrast, shorter glutamine repeats do not form stable polar zippers because the entropy loss cannot be offset by the liberation of fewer solvent molecules (4).

To test the initial hypothesis of Perutz *et al.* (3) that glutamine repeats in proteins can associate with each other by forming polar zippers, we inserted a 10-glutamine repeat into the inhibitory loop of chymotrypsin inhibitor 2 (CI2), a naturally monomeric protein, to see whether this caused the protein to associate into oligomers (5). This mutant, CI2-Q10i, did indeed form dimers and trimers as well as monomers, and the circular dichroism spectra of these oligomers indicated that they had associated by the formation of polar zippers. However, they were so stable that they could be dissociated into monomers only by denaturation of the protein. Therefore, loop-insertion mutants with shorter glutamine repeats were constructed in the hope that a dynamic equilibrium could be established. These mutants also formed extremely stable oligomers, and it became clear that some other mechanism must be responsible for their formation. Oligomerization has now been shown to occur by domain swapping (K.S., P. Scamborova, T. Galvao, and M.F.P., unpublished results), but polar zipper interactions have not been ruled out. Here, we set out to determine the crystal structure of one of the mutant oligomers, to see whether such interactions could be observed. We tried to purify the monomers, dimers, and trimers of all of these loop-insertion mutants for crystallization. The CI2-Q10i monomer gave crystals that diffracted only to low resolution (K.S., unpublished observations), but the CI2-Q4i dimer produced two suitable crystal forms. The structure of the hexagonal form has been solved and is described below.

MATERIALS AND METHODS

Crystallization. The protein CI2-Q4i consists of the residues Gly-Gln₄-Gly-Met (GQQQQGM) inserted into the inhibitory loop of truncated CI2 (first 20 disordered residues removed and replaced by an N-terminal methionine) immediately after the native residue Met59. For residue numbering, we adopt the wild-type CI2 convention for the convenience of structure comparison. We used "N domain" (residues 21 to 57) and "C domain" (residues 61 to 83) to identify the two fragments that constitute a globular unit (pseudomonomer) of the dimer. For discussion purpose, the residues inserted after Met59 are identified as Gly59A, Gln59B, Gln59C, Gln59D, Gln59E, Gly59F, and Met59G, to comply with the Protein Data Bank convention (6, 7). The CI2-Q4i mutant protein was expressed in *Escherichia coli* strain NM554 by using plasmid pCI2-Q4i (K.S., P. Scamborova, T. Galvao, and M.F.P., unpublished results). The dimeric fraction of this mutant was purified as

Abbreviation: CI2, chymotrypsin inhibitor 2.

[†]Y.W.C. and K.S. contributed equally to this work.

Data deposition: The atomic coordinates and structure factors have been deposited in the Protein Data Bank, Biology Department, Brookhaven National Laboratory, Upton, NY 11973 (PDB ID codes 1CQ4 and R1CQ4SF).

[§]To whom reprint requests should be addressed.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

PNAS is available online at www.pnas.org.

described for the corresponding 10-glutamine loop insertion mutant CI2-Q10i (5) and was crystallized at 4°C by the hanging drop method. The drops were prepared by mixing 2 μ l of the purified dimer at 25 mg/ml with an equal volume of crystallization buffer (30% wt/vol polyethylene glycol 400, 1.0 M lithium sulfate, and 1 mM calcium chloride, in 0.1 M Tris-HCl at pH 7.5), which also was used as the well buffer (1 ml volume). Crystal growth typically took 4 weeks. A single hexagonal crystal measuring $0.3 \times 0.3 \times 0.2$ mm was selected by using a 0.4-mm loop and was mounted in a cryostream at 100 K for data collection; the transfer of crystals had to be done extremely quickly to avoid fragmentation of the crystal. No cryo protectant was added because the high concentration of polyethylene glycol 400 in the crystallization medium made it viscous. The dimeric CI2-Q4i also crystallizes in a cubic form (K.S., unpublished observations).

Structure Determination. The hexagonal crystals belong to the space group P622, with cell dimensions of $a = b = 68.27$ Å and $c = 60.83$ Å. This is similar to the wild type, which also crystallizes in P622, with $a = b = 69.02$ Å but a smaller unit cell dimension c of 52.89 Å (8). The crystallization conditions of CI2-Q4i were similar to those of the wild-type CI2 (9). The crystal parameters suggest that there is one molecule of CI2-Q4i per asymmetric unit, with a solvent content of 53% and Matthew's coefficient, $V_m = 2.6$ Å³/dalton. A single crystal was used to collect two continuous segments of data (40 and 72 frames, respectively, with 1° oscillation per frame) at the X11 beamline in Hamburg. Diffraction images were processed with MOSFLM (10) and then were scaled and reduced with programs from the CCP4 suite (11). The data were significantly anisotropic: while the crystal diffracts strongly beyond 1.8 Å along h and k , the diffraction intensities fall off rapidly beyond 2.4 Å along l . Ten percent of the data were set aside for cross-validation and were omitted from all stages of refinement. The structure was solved by the molecular replacement method (12) with the program AMORE (11, 13, 14). The search model was constructed from wild-type CI2 (PDB ID code 2CI2) (6, 7), with the first two disordered residues at the N terminus (residues I19 and I20) as well as the residues flanking the insertion site (residues I54–I63) removed. A clear solution was found with an R -factor of 0.42 and a correlation coefficient of 70.3 for data in the range of 8–2.5 Å. The R -factor dropped to 0.39 after rigid-body refinement.

Refinement. The structure was refined with REFMAC (11, 15), with alternating cycles of manual rebuilding. The inclusion of anisotropic scaling in subsequent steps proved to be essential. A bulk solvent model also was used. The computer graphics program O (16) was used for rebuilding, aided by σ_A -weighted maps with $2mF_o - DF_c$ and $mF_o - DF_c$ coefficients (17). ARP (18) was used to add and refine water molecules automatically. After four cycles, the refinement converged, and the resulting model has an R -factor of 0.24 and R_{free} of 0.30, using all data to 1.8 Å. The stereochemistry of the final model was checked with PROCHECK (19) and was found to be good. The relatively high R -factors are probably attributable to the disorder of $\approx 14\%$ (10 residues of 71) of all protein atoms. The structural model of CI2-Q4i consists of residues 21 to 57 and 61 to 83, which are called the N and C domains. Most residues in the model have clear electron density. A high average real-space correlation coefficient of 0.89 (20) was calculated with O (16), which shows that the model correlates well with the x-ray data. No electron density was observed for residues 20, 58, 59, 59A–59G, and 60. Despite extensive density modification techniques using the programs DM and DMMULTI (11, 21), including multiple crystal form averaging with non-crystallographic symmetry (the cubic form has three monomers in the asymmetric unit), the connectivity between residue 57 and 62 could not be found. In case the space group of CI2-Q4i was of a lower symmetry than P622—i.e., trigonal—and the hexameric rings could actually be trimers of domain-

Table 1. Statistics of data processing and refinement

Data processing	
Resolution, Å	22-1.8 (1.9-1.8)
Number of measurements	86441
Number of unique reflections	8153
Multiplicity	10.6 (10.8)
Completeness of data, %	99.7 (100)
R_{merge}	0.074 (0.286)
Refinement	
Resolution, Å	22-1.8 (1.9-1.8)
Number of reflections in refinement/cross-validation	7358 (795)
Number of protein atoms	474
Number of water molecules	55
Number of SO ₄ atoms	5
R -factor	0.24 (0.24)
R_{free}	0.30 (0.35)
Overall anisotropic B factors, scaling	
$F_o - F_c$	
B_{11}	-6.1
B_{22}	-9.0
B_{33}	23.9
B_{12}	-5.1
Average B factor, Å	32.2
rms deviations	
Bond distance, Å	0.016
Angle distance, Å	0.034
Overall error in coordinates from R_{free} , Å	0.15
PROCHECK validation	
Ramachandran plot (most favored/additional allowed)	94.0/6.0
Overall G -factor	-0.2

R_{merge} is defined as $\Sigma |I - \langle I \rangle| / \Sigma I$. R -factor = $\Sigma ||F_o| - |F_c|| / \Sigma |F_o|$, where $|F_o|$ and $|F_c|$ are the observed and calculated structure factors, respectively. R_{free} is the same as R -factor but calculated with 10% of the data excluded from refinement.

swapped adjacent dimers, in the same layer, processing the data in P622 would have averaged out the dimeric linkage electron density. We also processed them in P312 and P321. We then used the final model from the P622 data to generate a suitable dimeric model for each of these two trigonal space groups and calculated $2mF_o - DF_c$ and $mF_o - DF_c$ maps. No electron density was observed for the loops in either space group. We also found that both the R -factor and R_{free} were very

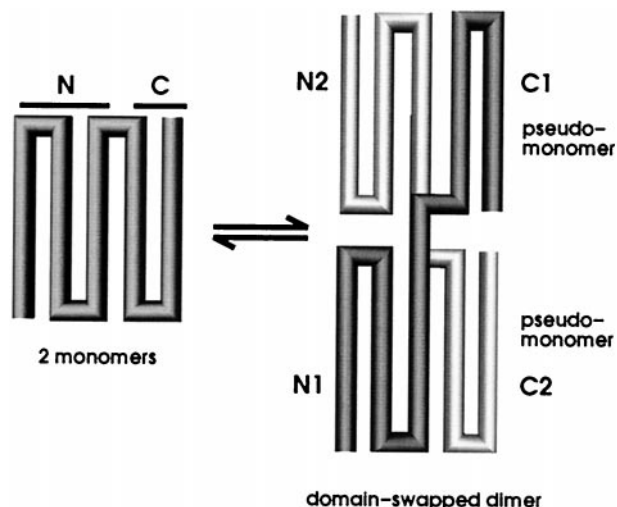


FIG. 1. Diagrammatic sketch of how monomers associate to form a domain-swapped dimer. N and C denote N-terminal and C-terminal domains, respectively.

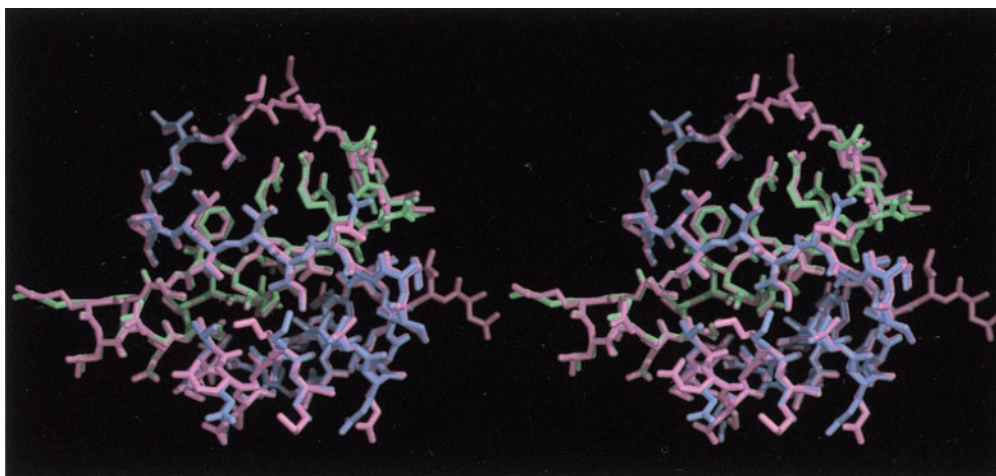


FIG. 2. The crystal structure of the pseudomonomer of CI2-Q4i compared with the wild-type CI2 monomer. The structure of CI2-Q4i is colored purple for the N domain (residues 21 to 57) and pale green for the C domain (residues 61 to 83). Wild-type CI2 is colored pink. The rms displacement between all main-chain atoms of the crystal structures of CI2-Q4i and wild-type CI2, when all common residues are superimposed, is 0.20 Å (that for all atoms is 0.88 Å). Figs. 2, 3, and 4 in this work were generated with BOBSRIPT (30), which is an enhanced version of MOLSCRIPT (31), and were rendered with RASTER3D (32).

similar ($R = 0.23$, $R_{\text{free}} = 0.27$ for P312; $R = 0.24$, $R_{\text{free}} = 0.26$ for P321) in both trigonal space groups: the two trigonal space groups are indistinguishable, confirming that it is more accurate to refine the structure in P622.

RESULTS AND DISCUSSION

Each member of the dimer is made up of two domains, one from each member; one domain lies N-terminal and the other

C-terminal to the inserted residues. We shall call the members pseudomonomers (Fig. 1). Like wild-type CI2, the dimer crystallizes in space group P622. The unit cell contains one pseudomonomer in the asymmetric unit (Fig. 2). The pseudomonomers must therefore be related by a dyad axis of symmetry. All of the secondary structure elements (an α -helix and a four-stranded mixed parallel and antiparallel β -sheet) and the overall fold of the wild type are preserved (8, 22). In both wild-type CI2 and CI2-Q4i, the proteins are arranged in layers

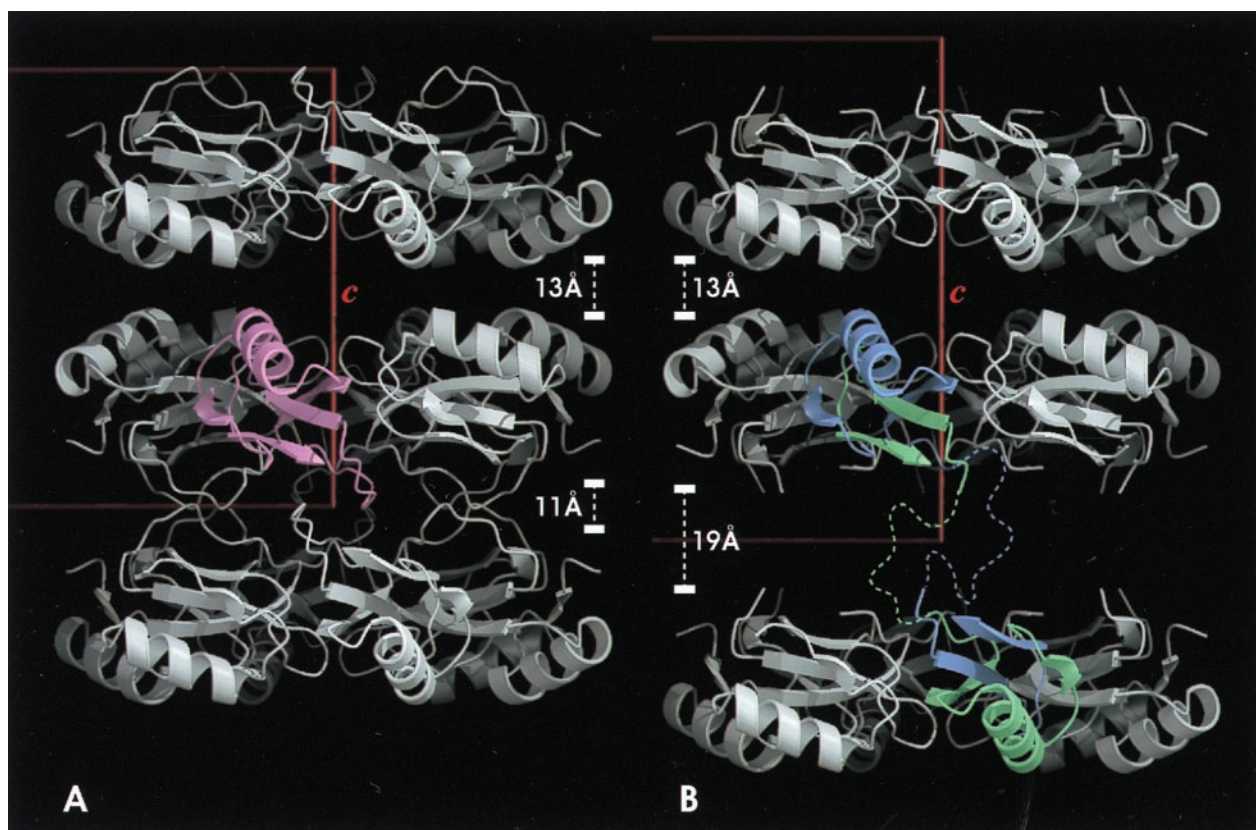


FIG. 3. The quaternary structure and crystal packing of wild-type CI2 (A) compared with that of dimeric CI2-Q4i (B). Three hexameric ring layers are shown. In B, a CI2-Q4i dimer is colored, and in A, a CI2 monomer is colored. In the CI2-Q4i structure, one of the spaces between the layers is expanded to 19 Å, compared to 11 Å in the wild-type. The color coding is the same as in Fig. 2. The dotted lines represent the inferred positions of the disordered residues. The red boxes are the unit cells. This view is along the *b* axis.

of hexameric rings. In the wild type, one of the spaces between the layers contains the inhibitory loop residues packing against symmetry-related loop residues of the next layer. This space is 11 Å wide and contains extensive hydrophobic interactions (Fig. 3*A*). The corresponding space in the mutant structure is 19 Å wide (Fig. 3*B*) and contains no distinct peaks of electron density. The alternate spaces are 13 Å wide in both the wild type and the mutant structures (Fig. 3*A* and *B*). Contacts in the 13 Å layer consist of very few van der Waals interactions between the symmetry-related α -helices containing residues Glu33, Glu34, and Lys37; the weakness of these interactions may explain the fragility of the crystals along this plane.

The crystal structure of CI2-Q4i does not reveal how the dimer was formed because the loops connecting the N and C domains are disordered. There are two possible mechanisms that would leave the structure of each component of the dimer apparently undisturbed: either domain swapping (for reviews, see refs. 23 and 24) or concatenation of loops on folding. Both mechanisms would account for the pseudomonomers being kept together so firmly that only denaturation can separate them. Biochemical studies showed that CI2-Q4i is domain-swapped rather than concatenated (K.S., P. Scamborova, T. Galvao, and M.F.P., unpublished results). The trace of the polypeptide chain before and after the disordered loop region and the distances between these two residues from two neighboring 2-fold symmetry-related pseudomonomers shows that the only arrangement consistent with the 2-fold symmetry and the distance and volume required to accommodate the bulky inserted residues is that shown in Fig. 3*B*. This places the symmetry-related pseudomonomers of the dimer in adjacent layers, which are linked covalently by the inserted residues. The space is 19 Å wide, shorter than the 24 Å that would be the length of the seven inserted residues if they formed an extended β -strand. The inserted residues are disordered. Insertion of residues other than glutamines into the CI2 loop also induced oligomerization, which implies that extension of an external loop alone is sufficient to induce misfolding by domain swapping (K.S., P. Scamborova, T. Galvao, and M.F.P., unpublished results). The seven residues introduced into the CI2-Q4i mutant contribute an addition of $\approx 10\%$ of protein atoms. This was easily tolerated. CI2-Q4i folds like two fragments (residues 20–57 and residues 61–83) docking to-

gether to form the pseudomonomer. Similarly, wild-type CI2 (truncated, residues 20–83) can be cleaved into two fragments [CI2 (20–59) and CI2 (60–83)], and these fragments can fold into one structure indistinguishable to the wild type and to our mutant (25, 26).

We found not the results we had hoped for, which was oligomerization by regular hydrogen bonds formed between neighboring β -strands as in a synthetic poly-L-glutamine (3), but instead we found oligomerization by domain swapping. Does this leave anything that we can learn from our results about diseases caused by expansion of glutamine repeat? Insertion of several bulky polar glutamines into a flexible loop does not destroy the structure of our model protein. CI2 loop-insertion mutants have stabilities and folding rates similar to those of the wild type (27). Our mutant CI2-Q4i protein cannot fulfill its normal function of inhibiting serine proteinases because the insertion is at its essential Met59, but if the insertion occurred at some nonessential external loop in a protein, preservation of its structure would allow the protein to function. The puffer fish Huntingtin homologue has a repeat of only four glutamines in the position of the long repeats in human Huntingtin (28), which shows that the ends of the glutamine repeat cannot be farther than 4×3.4 Å, or 14 Å, apart. Extension of the glutamine repeat beyond this length may be harmless, as long as its structure remains random. If the repeats form a hairpin, it may loosen the structure either by pulling the two ends too close together or by inserting itself between the β -strands of another molecule as happens in mutant serine proteinase inhibitors (29). So far, there is no evidence for domain-swapping in any of the proteins affected by polyglutamine diseases.

The authors thank Dr. Ashley Buckle for his help with data collection. This work was supported by a research grant from the Wellcome Trust.



FIG. 4. The $2mF_o - DF_c$ electron density near the insertion site of CI2-Q4i. The map is contoured at 1σ . No connectivity is seen between residues 57 and 61. Color coding is as in Fig. 2. Water molecules are represented by cyan spheres.

- Onodera, O., Roses, A. D., Tsuji, S., Vance, J. M., Strittmatter, W. J. & Burke, J. R. (1996) *FEBS Lett.* **399**, 135–139.
- Ordway, J. M., Tallaksen-Greene, S., Gutekunst, C. A., Bernstein, E. M., Cearley, J. A., Wiener, H. W., Dure, L. S. T., Lindsey, R., Hersch, S. M., Jope, R. S., *et al.* (1997) *Cell* **91**, 753–763.
- Perutz, M. F., Johnson, T., Suzuki, M. & Finch, J. T. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 5355–5358.
- Perutz, M. F. (1996) *Curr. Opin. Struct. Biol.* **6**, 848–858.
- Stott, K., Blackburn, J. M., Butler, P. J. & Perutz, M. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 6509–6513.
- Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Commission of the International Union of Crystallography, Bonn), pp. 107–132.
- Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. E., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112**, 535–542.
- McPhalen, C. A. & James, M. N. (1987) *Biochemistry* **26**, 261–269.
- McPhalen, C. A., Evans, C., Hayakawa, K., Jonassen, I., Svendsen, I. & James, M. N. (1983) *J. Mol. Biol.* **168**, 445–447.
- Leslie, A. G. W., Brick, P. & Wonacott, A. T. (1986) in *Daresbury Laboratory Information Quarterly for Protein Crystallography* (SERC Daresbury Laboratory, Warrington, United Kingdom), Vol. 18, pp. 33–39.
- Collaborative Computational Project, No. 4 (1994) *Acta Crystallogr. D* **50**, 760–763.
- Rossmann, M. G. (1972) in *International Science Review Series*, ed. Klein, L. (Gordon and Breach, New York).
- Navaza, J. & Saludjian, P. (1997) *Methods Enzymol.* **276**, 581–594.
- Navaza, J. (1994) *Acta Crystallogr. A* **50**, 157–163.
- Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997) *Acta Crystallogr. D* **53**, 240–255.
- Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991) *Acta Crystallogr. A* **47**, 110–119.
- Read, R. J. (1986) *Acta Crystallogr. A* **42**, 140–149.
- Lamzin, V. S. & Wilson, K. S. (1997) *Methods Enzymol.* **277**, 269–305.

19. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993) *J. Appl. Crystallogr.* **26**, 283–291.
20. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991) *Acta Crystallogr. A* **47**, 110–119.
21. Cowtan, K. (1994) *Jnt. CCP 4 ESF-EACBM Newslett. Protein Crystallogr.* **31**, 34–38.
22. Clore, G. M., Gronenborn, A. M., Kjaer, M. & Poulsen, F. M. (1987) *Protein Eng.* **1**, 305–311.
23. Xu, D., Tsai, C.-J. & Nussinov, R. (1998) *Protein Sci.* **7**, 533–544.
24. Bennett, M. J., Schlunegger, M. P. & Eisenberg, D. (1995) *Protein Sci.* **4**, 2455–2468.
25. Neira, J. L., Davis, B., Ladurner, A. G., Buckle, A. M., de Prat Gay, G. & Fersht, A. R. (1996) *Fold Des.* **1**, 189–208.
26. de Prat Gay, G. & Fersht, A. R. (1994) *Biochemistry* **33**, 7957–7963.
27. Ladurner, A. G. & Fersht, A. R. (1997) *J. Mol. Biol.* **273**, 330–337.
28. Baxendale, S., Abdulla, S., Elgar, G., Buck, D., Berks, M., Micklem, G., Durbin, R., Bates, G., Brenner, S., Beck, S., *et al.* (1995) *Nat. Genet.* **10**, 67–76.
29. Carrell, R. W. & Stein, P. E. (1996) *Biol. Chem. Hoppe-Seyler* **377**, 1–17.
30. Esnouf, R. M. (1997) *J. Mol. Graph. Model.* **15**, 132–136.
31. Kraulis, P. (1991) *J. Appl. Crystallogr.* **24**, 946–950.
32. Merritt, E. A. & Bacon, D. J. (1997) *Methods Enzymol.* **277**, 505–524.