Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

WHEATSHEAF: an algorithm to average protein structure ensembles

David Thomas[a]‡ and Annalisa Pastore[b]*

[a]Biological NMR Unit, Institute for Clinical Research, University of Birmingham Medical School, Birmingham B15 2TT, England, and [b]Division of Molecular Structure, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, England

‡ Present address: Scientific Software Solutions, 9 Stanley Street, Glasgow G41 1JA, Scotland.

Correspondence e-mail: apastor@nimr.mrc.ac.uk

Received 15 June 2004
Accepted 28 October 2004

A new algorithm is described that forms a single structure representative of ensembles of structures from files in the format used by the Protein Data Bank. A first attempt is made by averaging in the space spanned by bond lengths, inter-bond rotations and symmetry-multiplied dihedral rotations. This normally produces well formed regular secondary-structure elements, but the intervening less well ordered regions are often distorted because of the invalidity of averaging large rotations about divergent axes. For this reason, the algorithm includes a second stage that pulls the interatomic distances towards more fully representative values. Results produced by this method have proved better as judged by conventional quality checks than any input structure in nearly all cases tested so far, especially for the backbone, and much better than those produced by commonly used alternative methods.

1. Introduction

Many experimental and computational techniques in frequent use nowadays in the protein structure community output their results in the form of an ensemble of structures. Although the differences between structures in the ensemble can be of interest, being interpreted as signs of flexibility, motion or indeterminacy, it is nevertheless often more convenient to study a single representative structure. Intuitively, this is most naturally performed by averaging, and the variances of atomic positions are often retained and conveniently represented by numbers calculated by analogy to the thermal *B* factor so familiar to X-ray and neutron crystallographers. Indeed, many files deposited in the Protein Data Bank contain an averaged structure. Unfortunately, these average structures are usually derived by inappropriately averaging the Cartesian coordinates of the atoms in an Eulerian alignment of the structures in the ensemble (a simultaneous translational and proper rotational rigid-body superposition), resulting in poor and unrepresentative molecular geometries. Here, we describe a better way to average based on the more appropriate variables of bond lengths, inter-bond rotations[1] and symmetry-multiplied dihedral rotations, which necessarily produces superior results because the stereochemistry can be maintained.

The name of our algorithm arose spontaneously as a result of seeing results typified by Fig. 2(*b*) in Diamond (1992), which formally addresses the difficult (and possibly ill-posed) related question of optimal direct Eulerian alignment of more than two incongruent structures: when one aligns a short section of a chain, the tails that are not aligned diverge, as do the straws radiating from the binding around a wheatsheaf. We have addressed this effectively, if informally, with an algorithm that indirectly achieves essentially the same aim.

Our method is complementary to and complemented by the fine yet infrequently used cluster analysis of Diamond (1995), which can provide a clear insight into the variabilities in the ensemble of structures.

[1] Common usage is to refer to the angle subtended between two bonds radiating from the same atom as a bond angle. We use a simpler unit-free mathematical representation different from the chemical convention, for which we prefer the more precise geometric term of inter-bond rotation.

2. Methods

WHEATSHEAF is written in DEC Fortran, a well known *de facto* standard allowing structured data types and a wide choice of high-quality compilers that generate efficient machine code and helpful diagnostic messages. In all cases, we used deeply nested DO loops and tried to avoid IF statements, both choices being made in the light of long experience to minimize the risk of introducing errors.

The program first reads in the file containing the multiple determinations of the structure. Even this computationally undemanding step is optimized for speed. The input file must conform exactly to the format used by the Protein Data Bank for structures determined by NMR spectroscopy.

After reading the input file, the program determines the bonding pattern from the atom names and evaluates the squares of the bond lengths and unnormalized odd and even terms proportional to the sines and cosines of the inter-bond and dihedral rotations. These are then averaged, the structure is rebuilt from the averages and the α-carbon positions are output as the first attempt at a single representative structure. The computational time required to perform this highly optimized stage is negligible. The order of the rotational symmetry of the dihedrals is also determined from the atom names and is used to multiply the corresponding rotations in the internal representation of the structure; the symmetry-multiplied rotations are divided by the same factor on output and the symmetry-related atoms offset from each other by the corresponding fraction of one whole turn. This technique has the advantage of involving no decisions (*i.e.* break points), so it introduces no artefacts and does not interfere with convergence and results in output structures with perfect imposed symmetry, although the internal representation of the structures does at first sight look a bit odd. The geometrical effect of the symmetry handling is most easily described using phenylaniline as an example: the dihedral rotations about the dyadic $C^{\beta}-C^{\gamma}$ bond are doubled so that the two C^{δ} and C^{ε} atoms (and associated H atoms) move into coincidence, with the obvious result that the C atoms of the side chain no longer represent a hexagon, but something rather more reminiscent of the pan-shaped constellation Ursa major in the northern sky. This technique can result in collisions with other side chains because the plane of the group moves too, but it rarely necessitates special handling.

The initial average is quite striking in that regular secondary-structure elements (α-helices and β-strands) are usually well formed and clearly representative of the structures in the input file. The same cannot be said of β-turns, loops and less well ordered regions, however. These usually distort, sometimes strongly, so that one often sees a structure in which the regular elements appear very good, but the structure as a whole seems to have caved in or fallen apart. This result was anticipated because we knew from the outset that whilst rotations about closely clustered axes can meaningfully be averaged (as applies to consistently well formed secondary-structure elements), large rotations in three dimensions generally can not because they neither commute nor form a vector space, the effects of which become noticeable as the axes of the rotations being averaged diverge from one another by more than a few degrees, as is usually the case in loop and disordered regions. Averaging is unproblematic in two dimensions, where only one axis of rotation is possible. This fact is used to great effect in the program to simplify, speed up and improve the accuracy of the calculations, because individual bond rotations can be represented exactly in a two-dimensional symmetry-based formalism. It also means that despite the entire algorithm working by manipulation of rotations, there is no reference in the program to any angles or trigonometric functions. Some square roots do remain, however, and appear to be unavoidable. The equations used to handle rotations efficiently and accurately are either copied verbatim or deduced from Thomas (1990).

The second stage of the program is an iteration to adjust the dihedral rotations in such a way that every interatomic distance in the reconstructed output model is drawn towards the mean value of the matching distances in all of the input models in a manner correctly weighted to take into account the variance of that mean,[2] compensated (as formally correct) for the slight skew caused by representing distances as their squares. In the case of side chains with a dyad it is sometimes necessary to use an alternative formula because the artificially induced collisions from the symmetry handling can cause the variance of the distances to exceed the mean-squared distances themselves, in which case the normal deskewing formula breaks down. This actually occurs extremely rarely and is the only case in the entire program where special handling is used. The iteration is a multi-dimensional extension of the famous method of Joseph Raphson, almost universally misattributed to Isaac Newton (Thomas & Smith, 1990). It relies on knowing the rates of change (*i.e.* first derivatives) of the square of every interatomic distance with respect to every dihedral rotation. These are evaluated explicitly from the analytical formula, which is, thanks to the avoidance of trigonometric functions, remarkably simple (though this may not be clear to readers unfamiliar with the specialized notation; see Appendix *A*). This simplicity is fortunate, because the total number of these first derivatives is so large that there is no possibility using currently available computers of storing either them or a useful number of intermediate results. Because of their sheer number, their generation consumes the major fraction of the total running time of the program despite careful optimization and also necessitates the use of double-precision arrays. (Rounding errors rise as the square-root of the number of terms being summed.) The derivatives are weighted by a term representing the precision with which their associated dihedral rotation was first determined, which has the formally correct and desirable effect of helping the well determined regular secondary-structure elements move as almost rigid bodies whilst the less well determined regions of intervening structure move more freely. The inversion of the matrix equation required to solve the generalization of Raphson's method consumes nearly all of the remaining running time and is performed by a modification of an implementation of the conjugate-gradient algorithm (Hestenes & Stiefel, 1952) kindly donated by Robert Diamond many years ago. The adjustments are performed in a hierarchical sequence that improves the running speed because fewer atoms are involved in the rapidly moving early stages: at first only the C^{α} atoms are considered, then the heavy backbone atoms; the heavy side-chain atoms are then introduced and finally the H atoms are added. Convergence is rapid, with shifts normally decreasing quadratically, as they should with Raphson's method, although in some cases the structure must refold first because the initial average is inside-out or tangled in some way. The final adjustment of H atoms is normally accomplished in one cycle, which is fortunate since the computational expense of the 'all distances between all atoms' *versus* 'all distances between all atoms in all models' approach is hard to justify at this point, but we do it anyway on the grounds of simplicity and reliability. (The extension to H atoms involves nothing more than the safe and effortless technique of increasing the range of an outer loop by one.) Even so, the total

[2] In principle, each input model can also be weighted differently, preferably using a weight chosen by the spectroscopist, but this elementary modification has not yet been made for want of appropriate weights. In fact, the generation of justifiable weights may pose a serious challenge to the NMR community.

Table 1
Root-mean-square deviations (Å) of aligned bundles and results from *WHAT CHECK*.

In the table 'average' denotes the results appertaining to our new single representative structure, 'bundle' refers to the ensemble of raw structures and 'best' refers to the best structure in that ensemble.

(*a*) Values for all atoms and for backbone.

	All atoms	Backbone
2fmr	1.815	1.199
1fa3	1.537	0.901
1d8b	1.492	0.869
1dlx	1.319	0.833
1tin	1.534	0.825
1vig	1.973	1.131

(*b*) NewQua.

	Average	Bundle	Best
2fmr	−2.83	−3.99	−4.19
1fa3	−0.74	−0.62	−0.73
1d8b	0.62	0.05	−0.06
1dlx	−1.73	−2.45	−2.40
1tin	−4.53	−5.02	−4.58
1vig	−4.09	−4.32	−4.11

(*c*) OldQua.

	Average	Bundle	Best
2fmr	−1.337	−1.580	−1.435
1fa3	−1.174	−1.127	−1.081
1d8b	−0.658	−0.826	−0.825
1dlx	−0.872	−0.962	−1.021
1tin	−1.905	−1.994	−2.088
1vig	−1.559	−1.786	−1.820

(*d*) First-generation packing quality.

	Average	Bundle	Best
2fmr	−2.092	−3.513	−2.339
1fa3	−1.686	−2.092	−1.451
1d8b	−0.394	−1.686	−0.811
1dlx	−0.930	−0.811	−1.302
1tin	−3.513	−2.649	−3.970 (poor)
1vig	−2.649	−0.930	−3.301

(*e*) Second-generation packing quality.

	Average	Bundle	Best
2fmr	−2.826	−3.868 (poor)	−4.184 (bad)
1fa3	−0.735	−0.598	−0.099
1d8b	0.620	−0.512	−0.287
1dlx	−1.735	−2.568	−2.398
1tin	−4.643 (bad)	−4.439 (bad)	−4.580 (bad)
1vig	−4.090 (bad)	−4.487 (bad)	−4.110 (bad)

(*f*) Appearance of Ramachandran plot.

	Average	Bundle	Best
2fmr	−1.345	−3.578 (poor)	−3.701 (poor)
1fa3	−0.959	−2.107	−2.046
1d8b	−1.599	−2.370	−2.492
1dlx	−4.095 (bad)	−5.023 (bad)	−5.494 (bad)
1tin	−4.107 (bad)	−5.271 (bad)	−5.412 (bad)
1vig	−4.609 (bad)	−5.511 (bad)	−5.486 (bad)

running time is typically only a few minutes on currently available workstations or personal computers. (DJT normally uses a single 3.06 GHz Pentium 4 processor with 512 Mbytes of memory and the Intel Fortran compiler, with impressive combined performance.)

The code compiles successfully using the DEC VAX/VMS Fortran, DEC Alpha OpenVMS Fortran, SGI Irix f77 (when limited to 96 residues because of system frame-size limitations), Portland Group f77 (Linux) and Intel Fortran (Linux) compilers. A basic implementation (without diagnostic and intermediate outputs) is freely available from the *WHAT IF* website (http://swift.cmbi.kun.nl/WIWWWI/, option NMR) and the latest executables of the full implementation are available from Scientific Software Solutions (e-mail dr_dj_thomas@yahoo.co.uk for details).

The memory requirements of the program pose no problems on easily affordable current computers. The computational time for small proteins is dominated by the N^3 algorithm needed to solve the matrix equations (where N represents the size of the protein), but for larger proteins this is overtaken by the time taken to set up the equations, which is proportional to $M \times N^4$ (where M is the number of models). Because of this, the *WHAT IF* website implementation is currently limited to 128 models and 128 residues to prevent individual users from crippling the service, but these will be increased as faster processors become available.

The above completes the description of the parts of the program necessary to calculate the representative structure, but an extra section was added to calculate root-mean-square deviations of atomic positions, which occupy the *B*-factor field in the output file (see Appendix *A*).

It is also possible to make use of the final covariances of the iteratively determined dihedral rotations, which give a very sharply delineated indication of 'hot spots' of apparent flexibility. Indeed, they may ultimately be of greater practical import than the root-mean-square deviations, which indicate only cumulative positional indeterminacy. However, we do not currently output them for want of an accepted file format for this bond-related information.

Although we regard the current version of the program as essentially finished, we anticipate future developments to handle nucleic acids as well as polypeptides.

3. Results

We have tested the performance of our program on published and unpublished protein structures specified by as few as seven and as many as 101 input models and containing between 65 and 120 residues and have subjected both the input files and output files to the *WHAT CHECK* quality-control tools in the *WHAT IF* program (Hooft *et al.*, 1996; Vriend, 1990). As can be seen from Table 1, in most cases the overall quality of the representative structure is judged to be better (more positive figures) than for any of the input structures. Exceptions can occur when the best structures in the ensembles are particularly good, so that combining them with the others can degrade them. 1fa3 is the only case in our test set for which the overall quality ratings (OldQua and NewQua) for the representative structure fail to beat the best structure in the ensemble, though the difference is practically insignificant.

In all cases tested so far, the appearance of the Ramachandran plot is judged to have improved significantly, even when the best structure is particularly good.

The planarity of aromatic rings is reported as unusually good: this is, of course, because their symmetry is imposed exactly.

It should be noted, however, that some files include significant local structural incompatibilities, in which case the representative structure may contain corresponding regions of poor stereochemistry which can adversely affect the more specific tests of packing quality included in *WHAT CHECK*, also shown in Table 1. When it occurs, we recommend the use of Robert Diamond's cluster-analysis

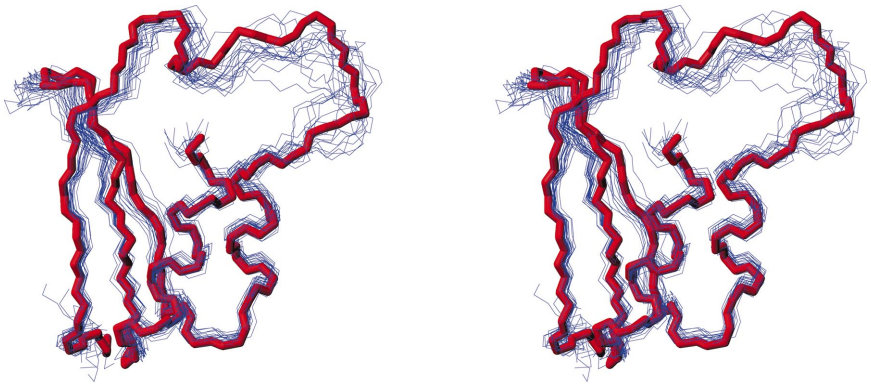

Figure 1
Backbone trace of the first KH module of FMR1.

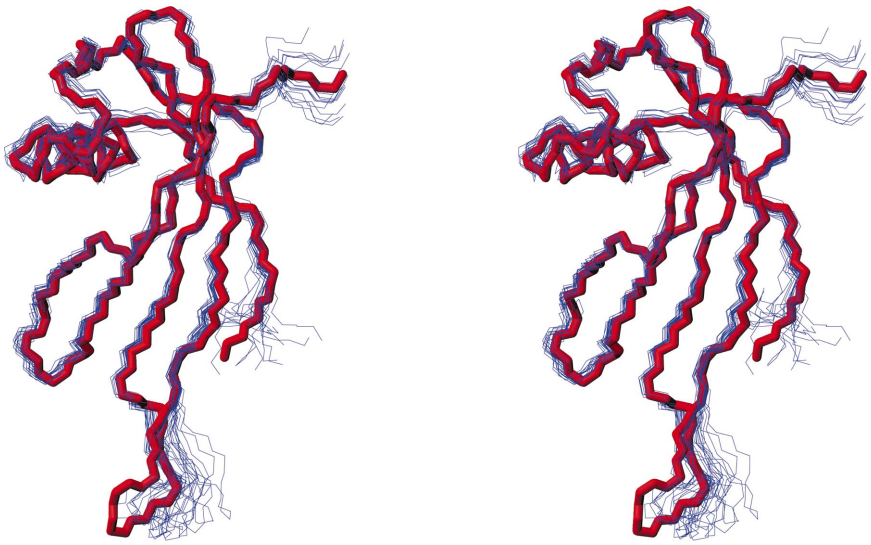

Figure 2
Backbone trace of MNEI.

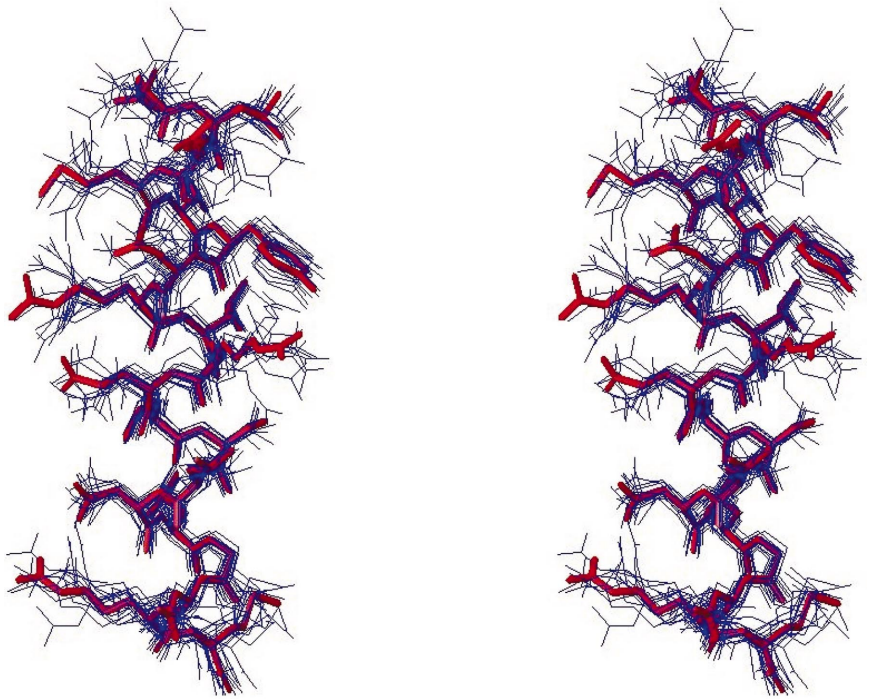

Figure 3
Backbone trace of residues 15–33 of the HRDC domain of *S. cerevisiae* RecQ helicase.

program (Diamond, 1995) either as an additional tool to sort the input models into more compatible subsets to be input into our program or else as an illuminating and trustworthy alternative approach.

Fig. 1 shows the backbone trace of our representative structure for the first KH module of FMR1 (PDB code 2fmr; Musco *et al.*, 1997). We choose this as an example because the large floppy loop would be represented particularly badly by the inappropriate method of averaging Cartesian coordinates, so it demonstrates particularly clearly the superiority of our method. Fig. 2 shows MNEI (PDB code 1fa3; Spadaccini *et al.*, 2000), an engineered sweet-tasting protein related to monellin, which is more typical. Fig. 3 shows residues 15–33 of the HRDC domain from *Saccharomyces cerevisiae* RecQ helicase (PDB code 1d8b; Liu *et al.*, 1999) including the heavy side-chain atoms.

4. Discussion

The greatest success of the algorithm is undoubtedly with the backbone, although if it is used to average closely similar structures, especially those arising from simulations of molecular dynamics, the same success is also obtained for the side chains. Under such circumstances, distance restraints that are satisfied in all input models will necessarily also be satisfied in the output model. However, there is no such guarantee if the input models disagree strongly, because of the non-linearities inherent in the method. We are aware of an artefact that can occur when models are averaged whose side chains are in disarray: they can cause the backbone to distort, which is immediately obvious when the input models are aligned onto the output. With the full implementation this is also easy to spot from the intermediate results. With the limited *WHAT IF* implementation it is possible to submit just the heavy backbone atoms first to ensure that the geometry cannot be distorted in this way and then to submit the full structure, which enables the performance to be checked even in the absence of the facility to inspect superposed structures. As Professor Gerrit Vriend has pointed out, when the purpose is to create a good structure rather than merely making one representative of the ensemble, it is a relatively easy matter to build the side chains onto the representative backbone, especially when the original distance restraints are available.

Sutcliffe (1993) explicitly investigated the utility of representing ensembles by single structures and concluded that with the algorithms available at that time it was still preferable to use the entire ensemble when the facility existed. It will be interesting so see whether experience with

WHEATSHEAF or other more recently developed algorithms will alter that conclusion.

APPENDIX *A*
Mathematical details

The most important new equations are the following derivatives of the squared interatomic distances with respect to the odd and even parameters, $\mathcal{O}$ and $\mathcal{E}$, specifying the intervening dihedral rotations,

$$\frac{\partial\langle dd\rangle}{\partial\mathcal{E}} = -2\langle pq\rangle\!\rangle\langle qr\rangle \frac{\mathcal{O}}{\mathcal{E}^2+\mathcal{O}^2} \quad (1)$$

$$\frac{\partial\langle dd\rangle}{\partial\mathcal{O}} = +2\langle pq\rangle\!\rangle\langle qr\rangle \frac{\mathcal{O}}{\mathcal{E}^2+\mathcal{O}^2}. \quad (2)$$

Here, $p\rangle$ and $r\rangle$ are the two limbs that rotate relative to each other about the dihedral bond $q\rangle$, $d\rangle = p\rangle + q\rangle + r\rangle$ is the interatomic vector and $q\rangle\!\rangle q$ is the dimensionless rank 2 skew-symmetric operator generating positive rotations in the plane normal to $q\rangle$. The two terms $\langle dd\rangle$ and $\langle pq\rangle\!\rangle\langle qr\rangle$ are just scalars (*i.e.* numbers) with the dimensions of length squared. The notation and derivation of these equations follow Thomas (1990).

The calculation of the root-mean-square deviations necessitates the Eulerian alignment of the input structures, which is performed by the elegant and efficient quaternion method of Diamond (1988) weighted by standard atomic weights. We calculate the eigenvalues of Diamond's 4 × 4 cumulant matrix **P** [his equation (22)] accurately from the explicit analytic solution of its (quartic) characteristic polynomial; the eigenvector $(\lambda, \mu, \nu; \sigma)$ corresponding to the largest eigenvalue is then calculated and would be (if normalized) the quaternion required. It is used to specify an ordinary 3 × 3 proper rotation matrix

$$\frac{1}{\lambda^2+\mu^2+\nu^2+\sigma^2}\begin{pmatrix} \lambda^2+\sigma^2-\mu^2-\nu^2 & 2\mu\lambda-2\nu\sigma & 2\nu\lambda+2\mu\sigma \\ 2\lambda\mu+2\mu\sigma & \mu^2+\sigma^2-\nu^2-\lambda^2 & 2\nu\mu-2\lambda\sigma \\ 2\lambda\nu-2\mu\sigma & 2\mu\nu+2\lambda\sigma & \nu^2+\sigma^2-\lambda^2-\mu^2 \end{pmatrix}$$

with which the pre-centred input structures are brought into accurate alignment. (We agree with Robert Diamond about the reliability and accuracy of this method.) The root-mean-square deviations of the positions of the aligned atoms are then found in the normal way and are output in the *B*-factor field of the output file. Diamond's method assumes uniform statistics and it is possible to improve the alignment by upweighting well aligned regions and downweighting badly aligned regions. We intend to incorporate this improvement in later versions of the program.

We owe an enormous debt of gratitude to Bob Diamond for many mathematical insights over the years and for establishing results beyond our own capabilities and thank Barry Levine, Gert Vriend, David Neuhaus, Michael Sattler and Bernd Simon for support, help, advice and access to unpublished ensembles, respectively. Some testing was facilitated by the kind award of an honorary fellowship to DJT by the University of Birmingham, which gave useful access to SGI computers from the Medical Research Council Bioinformatics Project housed in the Glaxo–Wellcome Biocomputing Laboratory of the Schools of Biosciences and Medicine.

References

Diamond, R. (1988). *Acta Cryst.* A**44**, 211–216.
Diamond, R. (1992). *Protein Sci.* **1**, 1279–1287.
Diamond, R. (1995). *Acta Cryst.* D**51**, 127–135.
Hestenes, M. R. & Stiefel, E. (1952). *J. Res. Natl Bur. Stand.* **49**, 409–436.
Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Nature (London)*, **381**, 272–272.
Liu, Z., Macias, M. J., Bottomley, M. J., Stier, G., Linge, J. P., Nilges, M., Bork, P. & Sattler, M. (1999). *Structure*, **7**, 1557–1566.
Musco, G., Kharrat, A., Stier, S., Fraternali, F., Gibson, T. J., Nilges, M. & Pastore, A. (1997). *Nature Struct. Biol.* **4**, 712–716.
Spadaccini, R., Crescenzi, O., Tancredi, T., De Casamassimi, N., Saviano, G., Scognamiglio, R., Di Donato, A. & Temussi, P. A. (2000). *J. Mol. Biol.* **305**, 505–514.
Sutcliffe, M. J. (1993). *Protein Sci.* **2**, 936–944.
Thomas, D. J. (1990). *Acta Cryst.* A**46**, 321–343.
Thomas, D. J. & Smith, J. M. (1990). *Notes Rec. R. Soc. Lond.* **44**, 151–167.
Vriend, G. (1990). *J. Mol. Graph.* **8**, 52–56.