electronic reprint

Automated detection of disulfide bridges in electron density maps using linear discriminant analysis

Thomas R. Ioerger

Automated detection of disulfide bridges in electron density maps using linear discriminant analysis

Thomas R. Ioerger

Department of Computer Science, Texas A&M University, TAMUS 3112, College Station, TX 77843, USA. Correspondence e-mail: ioerger@cs.tamu.edu

The ability to recognize disulfide bridges automatically in electron density maps would be useful to both protein crystallographers and automated model-building programs. A computational method is described for recognizing disulfide bridges in uninterpreted maps based on linear discriminant analysis. For each localized spherical region in a map, a vector of rotation-invariant numeric features is calculated that captures various aspects of the local pattern of density. These features values are then input into a linear equation, with coefficients computed to optimize discrimination of a set of training examples (disulfides and non-disulfides), and compared with a decision threshold. The method is shown to be highly accurate at distinguishing disulfides from non-disulfides in both the original training data and in real (experimental) electron density maps of other proteins.

1. Introduction

Disulfide bridges, formed through covalent cross-linking between cysteine side chains, are an important component of protein structure, constituting the primary exception to the linear nature of the peptide backbone. Disulfide bridges occur in approximately one out of every four proteins currently listed in the PDB (23%, using the PDBSelect non-redundant subset), tending to be more prevalent in small proteins (<100 amino acids) or secreted proteins (which must retain stability in a reducing environment), with no apparent bias for α, β, α/β, or $\alpha + \beta$ fold classes. Generally, disulfides are not thought to participate in enzymatic functions, but rather to add structural stability and reduce entropy due to internal thermal motions (Creighton, 1988), and are occasionally involved in coordination of metal ions, such as the classic FeS cluster. For these reasons, cysteines are among the most highly conserved residues, and their pattern in a sequence can often be used to help identify members of a protein super-family.

An automated method for recognizing disulfide bridges in electron density maps would be useful for protein model building, whether it is done manually, using graphical software such as *O* (Jones *et al.*, 1991), or with an automated model-building program, such as *ARP/wARP* (Perrakis *et al.*, 1999), *Resolve* (Terwilliger, 2000), *XPowerFit* (Oldfield, 1996), *Maid* (Levitt, 2001), or *TEXTAL* (Ioerger & Sacchettini, 2003). In an uninterpreted map, disulfide bridges introduce ambiguities in the density by creating non-linear connections which could potentially cause errors in backbone tracing. Despite the presence of sulfur atoms in the cysteines, there is no intrinsic characteristic of a disulfide bridge that can be observed directly in a density map. (At higher resolution, the sulfur atoms can sometimes be recognized by their bulbous density,

due to the larger orbitals; however, at lower resolutions, *e.g.* > 2.5 Å, electron density generally looks the same regardless of atom type.) An automated routine for localizing disulfide bridges in a map would help the crystallographer make correct decisions in tracing the backbone and building side chains. Furthermore, it would facilitate the process of sequence identification by anchoring local fragments on both sides to regions in the amino acid sequence containing cysteines. Clearly, prior knowledge of the location of disulfide bridges in an electron density map would benefit automated model-building programs as well.

Disulfide bridges can be detected automatically in electron density maps using pattern-recognition techniques. A method based on feature extraction and linear discriminant analysis is described. First, numeric feature vectors are calculated that characterize patterns of density in local regions of a map, and then a linear discriminant model is applied to the vectors to estimate the degree to which the regions resemble disulfides. The parameters of the linear discriminant are derived from a representative training set of density patterns associated with known disulfides and non-disulfides. The hypothesis is that the local shape of the density, for example, the typical spiral or twist, can be used to determine the presence of a disulfide. However, part of the challenge is that the shape of a disulfide might be confused with other short, non-branched structures, such as lysine or methionine side chains, or even glycine backbones. Therefore, the extracted features must be adequate to capture the disulfide pattern and to discriminate it from all other possibilities. The results demonstrate that this linear discriminant approach can be used to identify disulfide bridges with high accuracy in uninterpreted electron density maps, even in the presence of a moderate amount of noise (*e.g.* phase error) in experimental data.

2. Background

Disulfide bridges are formed from the covalent linkage between the S_γ atoms of two cysteine residues, arbitrarily separated in sequence. The length of a disulfide, as measured by $C_\alpha—C_\alpha$ distance, typically ranges between 4.5 and 6.5 Å (Thornton, 1981). Disulfides come in one of three common conformations: a left-handed twist, a right-handed spiral, and a less-frequent extended (straight) conformation, which is slightly longer than the other two (up to 7.5 Å) (Richarson, 1981). These left- and right-handed conformations are determined by the dihedral angle across the $S_\gamma—S_\gamma$ bond (restricted to $\pm 90°$ due to steric constraints), and the rest of the atoms adopt a corresponding low-energy position, resulting in the preference for a spiral or twist conformation.

Disulfide bridges can be recognized visually in an uninterpreted map based on their local density pattern, or shape, as shown in Fig. 1. To emulate this ability in an automated procedure, both feature extraction and linear discriminant analysis can be used. The analysis is based on a spherical region of density centered at the midpoint of a hypothetical $S_\gamma—S_\gamma$ bond of a disulfide and extending out 4–5 Å in radius: just enough to cover the C_α atoms in the backbones. Numeric features can be computed that capture salient aspects of the shape of the density in the region that are characteristic of disulfides. However, it is essential that these features be rotation-invariant, such that their values would not change if the region were rotated arbitrarily in three dimensions, since disulfide bridges may occur in any orientation in space. Such a set of rotation-invariant features has been developed for the *TEXTAL* model-building program, and is described by Holton *et al.* (2000). The features include such geometrical properties as distance-to-center-of-mass and moments of inertia, as well as statistical properties like mean, standard deviation, and higher-order moments of the distribution of density in the region. These features (19 in all) are easily shown to be rotation-invariant, they differ from region to region, and yet they show general similarities for regions with similar patterns of density.

Once regions of density have been reduced to numeric feature vectors, linear discriminant analysis (LDA) can be used to learn how to recognize the difference between members of distinct classes. The goal of the approach is to capture the disulfide/non-disulfide distinction, *i.e.* to be able to distinguish all disulfides as a group from all other components of protein structure, even in spite of the existence of left- and right-handed conformational sub-classes. Therefore, a representative set of disulfides and non-disulfides is needed (discussed below in §3), in order to train the linear discriminant model. Let these two sets of example regions be called X and Y, respectively (X for disulfides and Y for non-disulfides), with p and q instances each: $|X| = p$ and $|Y| = q$. Let the feature vectors calculated for each instance in both sets be represented as: $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$, where $\mathbf{x}_i = \langle x_{i,1}, \ldots, x_{i,n} \rangle$, and similarly for the non-disulfides. Note that all feature vectors in both sets have the same dimensionality, n.

Given two sets of numeric feature vectors, a linear discriminant can be developed to classify feature vectors of future examples into the most appropriate class. A linear discriminant is a set of coefficients for a linear equation such that, when applied to a given feature vector, it predicts it is a member of class 1 when the value of the linear combination exceeds a decision threshold (typically 0), or class 2 otherwise. The discriminant can be represented by the following equation:

$$\sum_{i=1}^{n} a_i x_i \geq \theta,$$

where the coefficients a_i are the parameters of the linear discriminant model, the x_i are the individual feature values, and θ is the decision threshold.

Linear combinations are simple to compute. However, it is critical to determine a set of coefficients that gives the optimal discrimination between the two classes. One might think that a reasonable strategy would be to choose the a_i to represent the vector separating the means of the two classes, $\mathbf{a} = (\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y)$, effectively determining to which centroid the region represented by the vector is closest by projecting each point onto this line. However, this approach does not necessarily produce optimal classification. In 1936, R. A. Fisher introduced an alternative discriminant model that is optimal. By minimizing the risk of classification error, Fisher found that the optimal direction in space on which to project instances, maximizing class separation, is given by: $[\Sigma^{-1} \cdot (\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y)]^T$, where Σ is the covariance matrix, $\Sigma_{ij} = (1/p)\sum_{k=1}^{p}(x_{k,i} - \mu_i) \cdot (x_{k,j} - \mu_j)$. Hence the corresponding discriminant equation becomes

$$\left[\mathbf{x} - \tfrac{1}{2}(\boldsymbol{\mu}_X + \boldsymbol{\mu}_Y) \right] \cdot \left[\Sigma^{-1} \cdot (\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y) \right]^T \geq 0,$$

with the center of all the data (averaged over both classes) subtracted from the query feature vector $\mathbf{x}$, and using 0 as the decision threshold. The covariance matrix can easily be computed (and inverted) off-line, and the product with the difference of the means forms a simple set of coefficients, $\mathbf{a} = [\Sigma^{-1} \cdot (\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y)]^T$. Given a new test vector, it is a simple matter to subtract the overall mean, $\boldsymbol{\mu} = 1/2(\boldsymbol{\mu}_X + \boldsymbol{\mu}_Y)$, and multiply by the coefficients to yield $d = \mathbf{a} \cdot (\mathbf{x} - \boldsymbol{\mu})$. If $d \geq 0$, then that instance is predicted to be a member of class 1 (*e.g.* disulfide); otherwise class 2 (non-disulfide).

3. Methods

3.1. Training set

To train the linear discriminant, representative sets of disulfide and non-disulfide regions are needed. For the disulfide set, a subset of the PDBSelect was searched for instances where two S_γ atoms were within 2.2 Å of each other (since the mean interatomic distance in disulfide bridges is 2.05 Å). The PDBSelect contains a curated list of high-quality non-redundant (<25% pairwise homology) structures in the PDB (Hobohm *et al.*, 1992); only proteins with at least 75 amino acids that had been determined at a resolution of 3 Å or better were used. A total of 98 instances of disulfides were identified
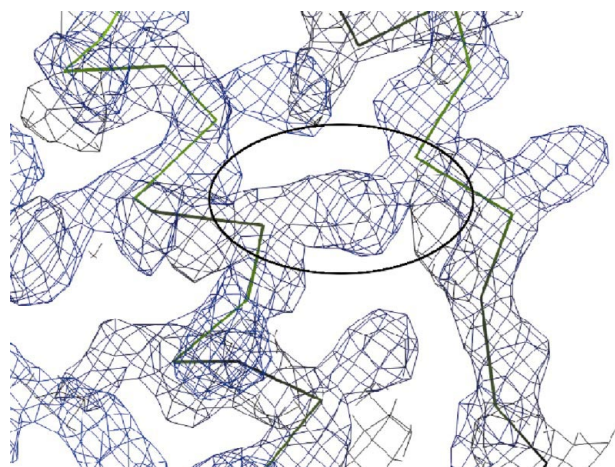
Figure 1
A disulfide bridge connecting a helix and a strand in a 2.8 Å map. The backbone trace among C_α atoms is shown in green. The density shown represents a contour level of approximately 1σ. The disulfide bridge, circled, can be seen to have a slight bulge and twist.

among 47 such proteins, forming the set X (class 1) for training the classifier. The specific coordinates taken to be the location of the disulfide were the midpoint of the two S_γ atoms, which is typically near (within 1 Å of) but not always right on the center of mass of the two side chains, nor the axis between the C_α atoms, due to the spiral or twist. For each of the 47 proteins, back-transformed electron density maps were generated at 2.8 Å using calculated structure factors, with the map borders adjusted to cover the monomer. This resolution was chosen because many maps, *e.g.* those derived from MAD data at synchrotrons, come out in this medium-resolution range.

For the non-disulfide set (Y, class 2), points were chosen throughout a representative protein known not to contain any disulfide bridges: 1FDI (formate hydrogenase), a large α/β protein with 715 residues. A back-transformed map was also calculated for 1FDI at 2.8 Å. Then this map was traced using
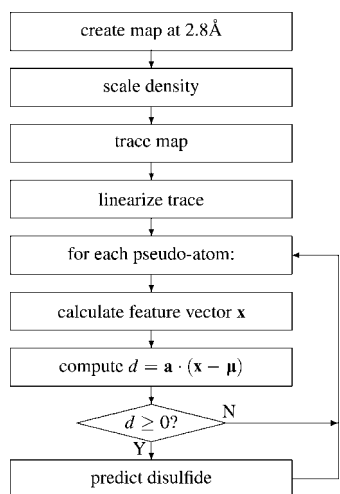

Figure 2
Procedure for determining whether a given region of density represents a disulfide bridge using a linear discriminant.

the skeletonization routine in *TEXTAL* (Ioerger & Sacchettini, 2002), producing a set of 7152 pseudo-atoms spaced roughly 0.5 Å apart along the medial axis of the density contours throughout the protein (including backbone and side chains). Each of these coordinates was taken to be an example of a non-disulfide.

3.2. Feature extraction and calculation of coefficients

For each of the 98 examples of disulfides and 7152 examples of non-disulfides, 76 features were calculated to form feature vectors. The features were those described in (Holton *et al.*, 2000), used for pattern matching in the *TEXTAL* model-building program. The features consist of 19 distinct values reflecting local statistical or geometric properties of the density, evaluated at four different radii, 3, 4, 5 and 6 Å, since some features capture different information in spheres of different size.

Once these feature vectors were calculated for the two sets, the 76×76 covariance matrix Σ was computed and inverted (using a standard matrix-inversion algorithm written in C). Then it was multiplied by the vector difference between the centroids of the two class, $\mu_X - \mu_Y$, to give a vector of coefficients, $\mathbf{a} = [\Sigma^{-1} \cdot (\mu_X - \mu_Y)]^T$. Finally, the global mean was calculated: $\mu = 1/2(\mu_X + \mu_Y)$. The vectors $\mathbf{a}$ and μ constitute the parameters of the linear discriminant model.

3.3. Procedure for recognizing disulfides in new maps

To apply the linear discriminant to a new electron density map, the following steps are taken. First, the map should be reduced to 2.8 Å by limiting structure factors, if higher resolution data are available. (If resolution is slightly worse than 2.8 Å, say up to 3 Å, the procedure will still work, though its accuracy drops off as density patterns begin to look more diffuse than those on which the discriminant was trained.) Then the map is traced using the skeletonization routine in *TEXTAL* (also requires scaling the density using the *caprascale* routine in *TEXTAL*). This produces a PDB file containing coordinates of pseudo-atoms along the medial axis of the contours; if a disulfide occurs, it is presumed to be centered on (or close to) one of these points. However, since disulfide bridges are known not to occur in terminal side chains, the branches of the trace may be stripped off using a linearization routine also in *TEXTAL* (which leaves long, contiguous skeletons mostly along the backbone of the protein, but also connected through disulfide bridges and other close contacts). For each coordinate along the linearized trace, the 76-dimensional feature vector is calculated describing the local pattern of density in the surrounding sphere. This vector is input into the linear equation with the model parameters to compute a value $d = \mathbf{a} \cdot (\mathbf{x} - \mu)$. If $d \geq 0$, then it is predicted to be a disulfide, or more specifically, in the vicinity of the midpoint of a disulfide bridge. The steps are summarized in Fig. 2.

4. Results

When the linear discriminant model is applied to examples in the original training set, the classification was found to be highly accurate at recognizing disulfide bridges. Among the 98 examples of true disulfides, 97/98 had scores ≥ 0. Among the 7152 non-disulfides, 7066/7152 (98.8%) had negative scores. The number of false positives can be reduced by 50% (from 86 to 35) by increasing the decision threshold θ from 0.0 to 2.0. This increases the accuracy of prediction on non-disulfides to 99.5%, while losing only one more true positive (96/98).

When the linear discriminant method is run on back-transformed electron density maps of proteins containing the example disulfides and predictions are made at the coordinates of each trace point, some interesting observations can be made about the locations of the points receiving positive scores. First, the 3–4 trace points closest to the midpoint of the $S_\gamma-S_\gamma$ bond tend to all have positive scores (see Fig. 3). It is important to note that no trace points necessarily fall exactly on the coordinates of the $S_\gamma-S_\gamma$ midpoint, although the discriminant model was specifically trained with such points, and only such points, as positive examples. (Recall that the trace is constructed from analysis of the density in an uninterpreted map and only follows contours, knowing nothing about atomic coordinates.) However, there is almost always a trace point within 0.5 Å of the true $S_\gamma-S_\gamma$ midpoint, given their 0.5 Å spacing and the high density across the bond.

In fact, it turns out that there is enough flexibility in the discriminant model that two trace points on either side (up to
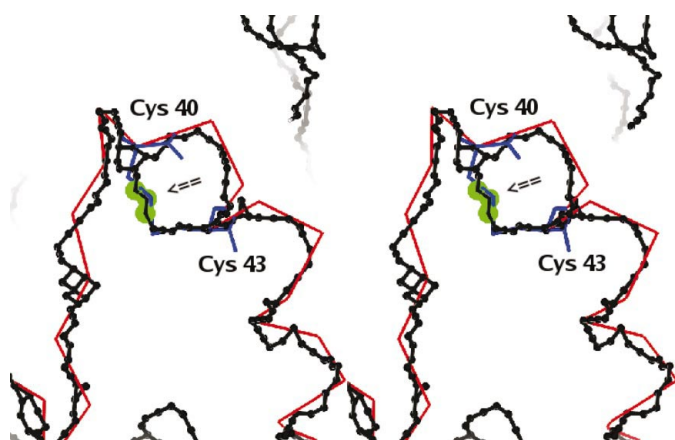

Figure 3
Recognition of a disulfide bridge between Cys40 and Cys43 in tryparedoxin (Alphey *et al.*, 1999), in a loop connecting an α-helix and a β-strand. The trace of a 2.8 Å map generated from experimental phases is shown by small black spheres, representing pseudo-atoms spaced at ~0.5 Å. Most of the side-chain branches have been trimmed off by the trace-linearization routine, except for those of the two cysteines, of which the connectivity through density is retained in the trace. The red lines show the C_α backbone trace of the true structure. The linear discriminant was calculated at each of the trace coordinates, and a larger green sphere is placed over those pseudo-atoms that are predicted to be near a disulfide ($d \geq 0$). The only such trace atoms with positive scores are the three indicated by the arrow, which are associated with the true disulfide bridge. For reference, the backbone and side-chain atoms of the cysteine residues in the true structure are shown in dark blue. The figure is shown in 'cross-eyed stereo'. All figures in this paper were created using *Spock* (Jon Christopher, 1998).

about 1 Å away) are recognized as disulfides. This means that the local density patterns and hence feature vectors are all sufficiently similar, despite the small shifts in the center of the analysis region. This is fortunate because it implies that: (i) the search for possible disulfides does not have to be so fine-grained that it lands precisely upon the $S_\gamma-S_\gamma$ midpoint, requiring increased sampling, and (ii) there will likely be multiple trace points in a neighborhood that registers a positive score, lending confirming evidence to the presence of a disulfide.

Another observation that can be made from running the discriminant method on back-transformed maps is that trace points near the termini of side chains occasionally also have positive scores, though they have nothing to do with disulfide bridges. The cause of this phenomenon is unclear, but it is of little consequence if the linearized trace is used, which strips off the branches into side chains. When the trace for 1FDI (the non-disulfide-containing protein used as the source of negative examples) is linearized, around 1/3 of the trace points are removed, leaving 4493. Among these, 4490/4493 have negative d scores, showing that many of the false positives were associated with side chains and are easily filtered out by this method; only three points (<0.1%) remain as false positives in the entire map.

To evaluate the accuracy of the linear discriminant on experimental (*i.e.* non-synthetic) maps, the method was applied to five real electron density maps containing one or more disulfide bridges (Table 1): α_{2u}-globulin (Chaudhuri *et al.*, 1999), penicillopepsin (James & Sielecki, 1983), granulocyte macrophage colony-stimulating factor (Rozwarski *et al.*, 1996), RNAse S (Sevcik *et al.*, 1993), and tryparedoxin (Alphey *et al.*, 1999). The maps were phased and density-modified using *CNS* (Brünger *et al.*, 1998). The original resolutions are shown in Table 1, though all structure factors were truncated to generate 2.8 Å maps.

The linear discriminant was able to pick out all but one of the disulfide bridges in these five maps (see Table 2). Some false positives (regions with positive scores but not near disulfide bridges) were observed, though usually with lower discriminant scores, and hence lower ranking; the false positives tended to lie along the backbone, with no particular pattern. Peaks are counted as pseudo-atoms from the linearized trace that have positive discriminant scores, $d > 0$. However, they are clustered, so local regions with several nearby atoms with positive scores are only counted once (for example, the group of three pseudo-atoms in Fig. 3 would be considered as a single peak). Furthermore, it is important to note that: (i) atoms within 6 Å of the border of the map were disregarded, since the features for complete spheres of density could not be computed for them, and (ii) only peaks falling on a single copy of the molecule (or tetramer in the case of α_{2u}-globulin) were counted; peaks falling on symmetry-related fragments appearing in the map were disregarded. 'Ranks' gives the ranking of the peaks corresponding to the true disulfides, among the false positive peaks.

α_{2u}-Globulin had four peaks, corresponding to one disulfide in each monomer of the tetramer. No false positives (coordi-

Table 1
Real (experimental) maps used in this study.

The map for α_{2u}-globulin was made to cover a tetramer. The other maps were made around monomers.

Protein	Resolution (Å)	No. of residues	No. of molecules in map	No. of disulfides
α_{2u}-Globulin	2.5	158	4	4†
Penicillopepsin	2.8	323	1	1
Granulocyte macr.	2.35	123	1	2
RNAse S	2.5	97	1	1
Tryparedoxin	2.0	145	1	1

† One per monomer of the tetramer.

Table 2
Results of the linear discriminant.

'Peaks' gives the number of clusters of pseudo-atoms with discriminant score > 0. 'Maximum score' is the score of the top peak. 'True positive' gives the number of peaks located at true disulfide bridges in the structure. 'False positive' gives the number of peaks located in other parts of the structure. 'Ranks' gives the ranks of the true positives among all the peaks with positive scores.

Protein	Peaks	Maximum score	True positive	False positive	Ranks
α_{2u}-Globulin	4	2.91	4/4	0	1,2,3,4
Penicillopepsin	5	4.99	1/1	4	1
Granulocyte macr.	4	8.68	1/2	3	1
RNAse S	3	2.26	1/1	2	1
Tryparedoxin	3	10.51	1/1	2	2

nates with $d > 0$ at non-disulfides) were observed. For penicillopepsin, four false positives were observed, but the true disulfide was recognized as the highest-ranked peak. Granulocyte macrophage CSF was the only map in which a true disulfide was missed, which was due to a lack of trace points across the bridge caused by a break in the density at the default contour threshold used. But its other disulfide was recognized as the highest ranked peak, as was the disulfide in RNAse S. Finally, there were three peaks on the monomer for tryparedoxin, the middle-ranked of which corresponded to the true disulfide bridge.

5. Conclusion

Disulfide bridges can be automatically recognized in electron density maps *via* pattern recognition. The method presented primarily relies on two computational techniques: feature extraction and linear discriminant analysis. The extracted features capture various aspects of the pattern of density in a region in the form of numeric values that are rotation-invariant (independent of orientation). The linear discriminant makes evaluation of these feature values as efficient as computing a simple linear combination, where the coefficients are optimized for discrimination between disulfides and non-disulfides in the training set. Though the experiments were carried out at 2.8 Å, the coefficients could easily be re-optimized on maps at other resolutions. It is important, however,

that the resolution of maps being searched matches that of the training set, to maximize recognition of the patterns. Thus higher-resolution structure factors might have to be truncated, for example.

Automated recognition of disulfide bridges in uninterpreted electron density maps could provide a number of benefits to both protein crystallographers, as well as model-building programs. A protein crystallographer would benefit from knowing the locations of disulfide bridges, which helps in sequence identification by localizing fragments to regions containing cysteine in the sequence. In addition, automated model-building programs could exploit this information to help resolve ambiguities and avoid mistakes in backbone tracing. We are exploring ways of incorporating this disulfide-recognition technique into *CAPRA*, the backbone-tracing routine in *TEXTAL*, in order to make better decisions about the connectivity of C_α chains (Ioerger & Sacchettini, 2002).

References

Alphey, M., Leonard, G., Gourley, D., Tetaud, E., Fairlamb, A. & Hunter, W. (1999). *J. Biol. Chem.* **274**, 25613–25622.
Brünger, A., Adams, P., Clore, G., Gros, P., Grosse-Kunstleve, R., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N., Read, R., Rice, L., Simonson, T. & Warren, G. (1998). *Acta Cryst.* D**54**, 905–921.
Chaudhuri, B., Kleywegt, G., Bjorkman, J., Lehman-McKeeman, L., Oliver, J. & Jones, T. (1999). *Acta Cryst.* D**55**, 753–762.
Christopher, J. (1998). *Spock*, http://quorum.tamu.edu/.
Creighton, T. (1988). *BioEssays*, **8**, 57–63.
Fisher, R. (1936). *Annals Eugenics*, **7**, 179–188.
Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). *Protein Sci.* **1**, 409–417.
Holton, T., Christopher, J., Ioerger, T. & Sacchettini, J. (2000). *Acta Cryst.* D**56**, 722–734.
Ioerger, T. & Sacchettini, J. (2002). *Acta Cryst.* D**58**, 2043–2054.
Ioerger, T. & Sacchettini, J. (2003). *Methods in Enzymology*, edited by C. Carter & R. Sweet, Vol. 374, pp. 244–270. New York: Academic Press.
James, M. & Sielecki, A. (1983). *J. Mol. Biol.* **163**, 299–361.
Jones, T., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Levitt, D. (2001). *Acta Cryst.* D**57**, 1013–1019.
Oldfield, T. (1996). *Crystallographic Computing 7, Proceedings from the Macromolecular Crystallography Computing School*, edited by P. Bourne & K. Watenpaugh. Oxford University Press.
Perrakis, A., Morris, R. & Lamzin, V. (1999). *Nature Struct. Biol.* **6**, 458–463.
Richarson, J. (1981). *Adv. Protein Chem.* **34**, 167–339.
Rozwarski, D., Diederichs, K., Hecht, R., Boone, T. & Karplus, P. (1996). *Proteins*, **26**, 304–313.
Sevcik, J., Hill, C., Dauter, Z. & Wilson, K. (1993). *Acta Cryst.* D**49**, 257–271.
Terwilliger, T. (2000). *Acta Cryst.* D**56**, 965–972.
Thornton, J. (1981). *J. Mol. Biol.* **151**, 261–187.