Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Eigensystem analysis of the refinement of a small metalloprotein

Kevin Cowtan[a] and Lynn F. Ten Eyck[b]*

[a]University of York, Heslington, York YO10 5DD, England, and [b]San Diego Supercomputer Center, University of California at San Diego, La Jolla, CA 92093-0505, USA

Correspondence e-mail: lteneyck@sdsc.edu

Received 1 May 1999
Accepted 30 March 2000

The eigenvalues and eigenvectors of the least-squares normal matrix for the full-matrix refinement problem contain a great deal of information about the quality of a model; in particular the precision of the model parameters and correlations between those parameters. They also allow the isolation of those parameters or combinations of parameters which are not determined by the available data. Since a protein refinement is usually under-determined without the application of geometric restraints, such indicators of the reliability of a model offer an important contribution to structural knowledge. Eigensystem analysis is applied to the normal matrices for the refinement of a small metalloprotein using two data sets and models determined at different resolutions. The eigenvalue spectra reveal considerable information about the conditioning of the problem as the resolution varies. In the case of a restrained refinement, it also provides information about the impact of various restraints on the refinement. Initial results support conclusions drawn from the free R factor. Examination of the eigenvectors provides information about which regions of the model are poorly determined. In the case of a restrained refinement, it is also possible to isolate places where X-ray and geometric restraints are in disagreement, usually indicating a problem in the model.

1. Introduction

Full-matrix least squares has been widely used for the refinement of small molecules for many years. The approach has been adopted for two reasons. The refinement converges in very few steps and as a byproduct produces estimates for the standard deviations of all the refinement parameters (Stout & Jensen, 1989). A notable review of the theory of least squares for crystallographic applications is given by Diamond (1981).

Full-matrix refinement has not been widely applied to protein structures. This is partly because the size of the refinement normal matrix increases as the square of the number of parameters, rendering the calculation impractical until recently. Also, since the data resolution is typically lower, the data-to-parameter ratio is typically much smaller than for small molecules, requiring strong geometrical restraints to be placed upon the model to ensure convergence to a reasonable solution. These restraints are treated as additional data. The restraints typically dominate the estimated standard deviations for the parameters. If there are any inappropriate restraints or if the scaling of restraints and diffraction data is not correct (a common problem), the estimated standard deviations will be unreliable (Kleywegt & Jones, 1995).

Eigensystem analysis of the least-squares normal matrix has been suggested (Ten Eyck, 1996, 2000) as a possible means to obtain accuracy information for protein models at low resolution. Similar calculations have been performed in the past for small molecules (mentioned by Watkin, 1988). Preliminary results from the eigensystem analysis of a protein refinement are presented here and give some insight into the working of the refinement calculation. The theory is set out in detail in Ten Eyck (2000). A short geometrical description of the approach is given here.

Least-squares refinement involves the minimization of some refinement residual by bringing the model into agreement with the X-ray data and geometrical restraints. The refinement residual is a function in the high-dimensional space whose axes are the model parameters; these parameters are varied to locate a minimum in the residual. The full-matrix least-squares approach determines new values for the parameters by constructing a second-order approximation to the refinement residual and then moving to the minimum of the approximation.

The minimum will typically have different curvatures along different directions in parameter space. The second-order approximation is a multi-dimensional quadratic function, such that a contour of constant refinement residual will form a hyper-ellipse around the minimum. The principal axes of this hyper-ellipse need not be aligned with the parameter axes. The curvatures of the refinement residual along each of these axis are given by the eigenvalues of the normal matrix and their directions are given by the eigenvectors.

The curvatures and directions provide information about which parameter combinations are well determined or badly determined and about the overall conditioning of the problem. Examination of the curvatures and hence of the shape of the function in the region of the minimum also provides insight into the behavior of various refinement methods. The approximation to the minimum is equivalently described by either the normal matrix, by its inverse or by the eigenvalues and eigenvectors. Depending on what information is required one representation may be more useful than the others.

1.1. Definitions

E.s.d.: estimated standard deviation. Calculated for individual parameters or their combinations (*e.g.* bond lengths) from the variance–covariance matrix.

Eigenparameter: the eigenvectors represent a rotated set of axes in the parameter space. The 'eigenparameters' are the ordinates in the parameter space in terms of the rotated axes.

(Eigen-)parameter contribution: the 'contribution' of a parameter to an eigenvector (or *vice versa*) is the square of the cosine of the angle between the eigenvector and the parameter axis. The eigenvectors are orthonormal, so that the sum of the 'contributions' from every parameter to any eigenvector is 1.

Goodness-of-fit: the ratio of the actual fit of the model to the observations that are the objective of the fit, GooF = $\left(\{\sum_i [R_i(\mathbf{x}) - R_i^o]^2/\sigma_i^2\}/(n_r - n)\right)^{1/2}$, where $R_i(\mathbf{x})$ is the calculated value of the ith objective function, R_i^o is its desired value, σ_i is the expected e.s.d., n_r is the number of objectives and n is the number of parameters.

Normal equations: the least-squares refinement equations may be written $\mathbf{J}^T\mathbf{J}\delta x = \mathbf{J}^T\mathbf{r}$.

RHS: right-hand side (of the normal equations).

J: the matrix of derivatives of the residuals with respect to the parameters, $J_{ij} = \partial r_i/\partial x_j$.

N: the refinement normal matrix. $\mathbf{N} = \mathbf{J}^T\mathbf{W}\mathbf{J}$ and is real, symmetric and non-negative definite. **W** is the variance-covariance matrix for the *observations*. For statistically independent observations, **W** is a diagonal matrix.

$\mathbf{N}^{-1}$: the inverse of the normal matrix (when it exists). If the restraints are statistically independent and have unit variance, this is the variance–covariance matrix of the parameters at convergence. The correlation matrix may be obtained from the covariance matrix by dividing each row and column by the square root of the diagonal element (the standard deviation of the corresponding parameter).

R factor and $R_{\rm free}$: the R factor is defined as $R = \sum \big||F^o| - |F^c|\big| / \sum |F^o|$, where $|F^o|$ is the observed structure-factor amplitude and $|F^c|$ is the calculated structure-factor amplitude. $R_{\rm free}$ is an R factor computed using reflections that were not included in the refinement (Brünger, 1992).

r: the vector of weighted restraint residuals, $r_i = [R_i(\mathbf{x}) - R_i^o]/\sigma_i$. For X-ray terms, $R_i(\mathbf{x})$ will be the scaled calculated intensity $s|F_c(h)|^2$ and R_i^o will be the observed intensity $|F_o(h)|^2$. σ_i is the estimated standard deviation of the observed intensity for observation i.

x: the vector of parameters.

$\delta\mathbf{x}$: the vector of shifts required to minimize the residuals, subject to the assumptions of least squares.

2. Implementation and testing

The eigenvalue analysis described above was tested using two data sets at different resolutions for the protein *Azotobacter vinelandii* 7-Fe ferredoxin (Stout *et al.*, 1998). The protein crystallized in space group $P4_12_12$, with unit-cell parameters $a = b = 54.8$, $c = 92.6$ Å for the frozen crystals. The structure contains 106 residues and two Fe–S clusters. The structure was originally solved by Stout (1993) with room-temperature data to 1.9 Å (9586 reflections). The model was refined in *X-PLOR*, with a final R factor of 21.5%. 30 water molecules were modeled.

The structure was later re-refined (Stout *et al.*, 1998) using data from frozen crystals to a resolution of 1.3 Å (30 880 reflections). The model was refined with anisotropic thermal parameters and 162 water molecules (nine parameters per atom = 9211 parameters) using *SHELXL* (Sheldrick, 1997) to an R factor of 15%. The refinement was monitored by tracking the free R factor, but the final model was computed using all of the data. Therefore, the free R factor was estimated for this model by perturbing the coordinates and forcing the thermal parameters to be isotropic. Anisotropic refinement was then

repeated with a free set of 5% of the reflections (chosen by selecting every 20th reflection), giving a value of 19%. (Since most of the results described in this paper involve examination of the conditioning of the refinement problem rather than actual refinement of the model, the free reflections have been included for the remaining calculations.)

To calculate the least-squares normal matrix, the refinement program *SHELXL* (Sheldrick, 1997) was modified to write out the upper triangle of the normal matrix as a binary file, along with the RHS vector of the normal equations. This data was then read into a separate program which performed the eigenvalue and eigenvector calculation using the *LAPACK* routine *ssyev* (Anderson *et al.*, 1999). Tests revealed that the matrix diagonalization must be performed at 64-bit precision for numerical stability. Calculation at 32-bit precision may be possible with preconditioning (§3), but this depends on the propagation of errors in the diagonalization algorithm and requires careful investigation.

Both calculations are extremely memory intensive. The *SHELXL* calculation requires the upper triangle of the normal matrix and the eigenvector calculation requires a full matrix (which holds the eigenvectors at the end of the calculation) to be held in real memory. At 64-bit precision, the *SHELXL* calculation for the high-resolution case of 9200 parameters requires roughly 400 Mb of memory and the diagonalization requires roughly 700 Mb of memory. The binary files were of similar sizes. These sizes vary as the square of the number of parameters.

The calculations were performed using a Cray T90 vector computer, which calculates and stores all floating-point numbers to 64-bit precision. For the high-resolution problem, the *SHELXL* calculation of a single cycle of least-squares refinement including writing the normal matrix took about 50 CPU min and the diagonalization took about 120 CPU min. The normal matrix calculation was also performed on a DEC Alphaserver 4100 (533 MHz 21164) and took 3.5 h at 32-bit precision. These calculations are therefore within the reach of a dedicated fast workstation with sufficient memory. CPU time varies at up to the third power of the number of parameters.

Normal matrices were calculated for a range of different refinement problems, including different resolutions of X-ray data and different types and numbers of stereochemical restraints. All the calculations were performed using one of two starting models. In §§5 and 7 this model is the isotropic *X-PLOR* model after restrained full-matrix refinement against the 1.9 Å data; in §6 this is the anisotropic *SHELXL* model after restrained full-matrix refinement against the 1.3 Å data. Thus, the results for unrestrained or lower resolution calculations do not represent the conditioning of the problem at the minimum, but rather the curvature of the refinement residual at some point near the minimum. (The second-order expansion is equivalent to the assumption that the curvature matrix is slowly varying.) This allows more direct comparisons of the curvatures and also avoids the problem that many of the unrestrained refinements described would be unstable over multiple cycles.

Diagonalization of the normal matrix produces the eigenvalues and corresponding eigenvectors in order of increasing eigenvalue. This convenient numbering has been adopted in referring to particular eigenvalues and eigenvectors.

A suite of small programs has been written to analyze the resulting large data files, including conversion of eigenvectors to a form which may be visualized using the *XTALVIEW* molecular-graphics software (McRee, 1992).

3. Scaling the refinement parameters

The elements of the normal matrix and, therefore, its eigenvalues and eigenvectors are totally dependent on the units chosen for each model parameter. Thus, if a scale factor is applied to some subset of the parameters, the eigenvalue spectrum and distribution of parameters amongst eigenvectors will also change.

Different scalings will lead to the emphasis of different information in the eigensystem analysis. It is necessary to understand the effects of the choice of parameter scales on the refinement problem, so that the results of eigensystem analysis will be correctly interpreted and so that the scaling of the model parameters may be chosen to highlight the desired features of the refinement problem.

3.1. Scaling of similar parameters

An arbitrary set of units may be chosen for any parameter in the refinement calculation. For example, in the refinement of positional parameters alone in an orthorhombic crystal form, changing all the parameters from fractional to ångstrom coordinates involves only a rescaling of each parameter. Since the normal matrix is the self-product of the matrix of derivatives of the residuals with respect to the parameters, scaling the parameters will apply the inverse scale to both rows and columns of the normal matrix, *i.e.*

$$x'_i = s_i x_i \tag{1}$$

implies

$$N'_{ij} = (1/s_i s_j) N_{ij}, \tag{2}$$

where $\mathbf{x}'$ and $\mathbf{N}'$ are the vector of scaled parameters and the corresponding normal matrix.

Let $\mathbf{S}$ be a diagonal matrix, with diagonal elements s_i. Then

$$\mathbf{x}' = \mathbf{S}\mathbf{x} \tag{3}$$

$$\mathbf{N}' = \mathbf{S}^{-1}\mathbf{N}\mathbf{S}^{-1}. \tag{4}$$

The right-hand side of the normal equations and the vector of shifts will also be scaled as

$$\mathbf{r}' = \mathbf{S}^{-1}\mathbf{r} \tag{5}$$

$$\delta\mathbf{x}' = \mathbf{S}\,\delta\mathbf{x}. \tag{6}$$

When the normal equations are assembled the scaling matrices cancel. The resulting changes to the model will be unaffected by the change in parameterization.

Similarly, the variance–covariance matrix changes under scaling of the parameters. This changes the e.s.d.s of the rescaled parameters,

$$\mathbf{N}'^{-1} = \mathbf{S}\mathbf{N}^{-1}\mathbf{S}. \quad (7)$$

However, once the results are shifted back onto an absolute ångstrom scale the e.s.d.s of the model parameters are unchanged. The correlation matrix is also unaffected by the choice of parameter scales, since dividing the elements of the variance–covariance matrix by the standard deviation of each parameter exactly cancels the effect of the scaling.

Consider a refinement calculation for an orthorhombic system using fractional coordinates. Normally, both the X-ray data and the geometrical restraints will operate isotropically when expressed in ångstrom coordinates (assuming a sphere of data is collected and the overall anisotropy is low). The mean variances of the atomic positions in ångstroms should be similar along different directions. If we shift to fractional coordinates, the mean variance along *longer* axes will be *smaller*. There will, therefore, be some separation of coordinate types in the eigenvalue spectrum. Positional parameters along long axes will tend to contribute to eigenvectors with larger eigenvalues (and smaller variances) and positional parameters along short axes will tend to contribute to eigenvectors with smaller eigenvalues (and larger variances).

Separation of parameters by crystal direction can be prevented by applying scale factors which bring the position parameters onto an orthogonal ångstrom scale (or any other distance unit). This analysis can be extended to arbitrary crystal forms by allowing **S** to be a block-diagonal matrix whose 3×3 diagonal blocks correspond to the orthogonalization matrix. In this case the correlation matrix will be slightly altered.

Positional parameters will still be separated in the eigenvalue spectrum according to atomic type, since the contribution of an atom to the diffraction pattern varies with its number of electrons. Furthermore, different numbers and types of stereochemical restraints are applied to different types of atoms. When using X-ray terms alone, it may be informative to further scale the orthogonal ångstrom coordinates by the number of electrons in the particular atom or even a combination of number of atoms and U value. The former case may indicate well and poorly determined domains; the latter could indicate if parts of the structure may be well determined despite having high thermal motion.

This approach has been used by Tronrud (1992) in the *TNT* refinement package, in which the parameters are scaled by the curvature of the atomic density, combining information from the size, shape and thermal motion of each atom. This approach significantly accelerates the convergence of the refinement.

3.2. Scaling of diverse parameters

The problem of scaling together diverse parameters, including positional parameters along different directions, isotropic Us, diagonal and off-diagonal terms of anisotropic Us and special parameters such as the overall scale factor or partial occupancies is more difficult. Clearly, the mean curvatures of the positional and thermal parameters can be changed arbitrarily by choice of units (*e.g.* representing thermal parameters as Bs or Us). Parameters of different types may not therefore be compared on the basis of curvature alone. In some cases (as later in this paper) it may be useful to choose units which give rise to very different mean curvatures for positional and thermal parameters, so the variation amongst parameters of each type may be examined separately.

For numerical purposes (*i.e.* to minimize the impact of limited numerical precision) it is often convenient to chose a scaling which minimizes the range of curvatures. This is referred to as 'preconditioning' the normal matrix (see, for example, Trefethen & Bau, 1997). The simple system

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad (8)$$

can be conditioned by pre-multiplying the equations by a matrix $\mathbf{M}^{-1}$, where **M** is some approximation to **A** which is also easily invertible. This gives

$$\mathbf{M}^{-1}\mathbf{A}\mathbf{x} = \mathbf{M}^{-1}\mathbf{b}. \quad (9)$$

Since **M** is an approximation to **A**, the matrix $\mathbf{M}^{-1}\mathbf{A}$ is a (generally very poor) approximation to the identity matrix. The resulting system of equations may be rapidly solved by iterative methods or solved by direct methods to a greater precision. In the following analysis we will apply this approach in the form

$$[(\mathbf{M}^{-1})^T\mathbf{A}(\mathbf{M}^{-1})][\mathbf{M}\mathbf{x}] = (\mathbf{M}^{-1})^T\mathbf{b}, \quad (10)$$

where $\mathbf{M}^T\mathbf{M}$ is some approximation to **A**. When **M** is symmetric (commonly the case) the transpositions can be omitted.

This type of approach could be applied to the scaling of different refinement-parameter types as follows. Solve the normal equations and calculate e.s.d.s for all the parameters and then scale every parameter by the inverse of its e.s.d. This is the equivalent of setting the scaling matrix **S** to the reciprocal of the square root of the diagonal elements of the variance–covariance matrix (the inverse normal matrix). Thus, $\mathbf{S}^{-1}\mathbf{S}^{-1}$ is a diagonal approximation to $\mathbf{N}^{-1}$ and $\mathbf{S}^{-1}\mathbf{N}\mathbf{S}^{-1}$ is an approximation to the identity matrix.

Since all the eigenvalues of the identity matrix are unity, preconditioning the normal matrix will tend to reduce the dynamic range of the eigenvalues, in the extreme case to the point where they are all equal. Clearly, this conceals any information which was present in the eigenvalue spectrum. However, it may be useful to apply a block scaling to all the thermal parameters so that the mean variance of the positional parameters and the mean variance of the thermal parameters are equal. Results of this approach will be published in a future paper.

3.3. Parameter scales in *SHELXL*

For the purposes of these initial studies, no additional scaling was applied. The parameter units were determined by the internal representation in *SHELXL*. Positional para-

meters are represented in fractional coordinates along the cell axes. The test structure is orthorhombic, so the coordinate axes are orthogonal. Some separation of positional parameters is expected by axis length, but in practice this is small in comparison with the range of curvatures arising from other causes.

Thermal parameters are represented as *U*s. In the case of anisotropic thermal parameters, the matrix elements are positional variances and covariances along the cell directions (which are orthogonal in this case) in Å^2 units. Thus, no separation is expected between different thermal components.

The relative scales of the fractional positional coordinates and *U* values leads to a dramatic difference in curvature between positional and thermal parameters. This was fortunate in this case because it separated the positional and thermal parameters in the eigenvalue spectra, simplifying their interpretation. It may be useful to adjust the parameter scales in other cases to achieve this same condition.

4. Weighting the refinement restraints

The variance–covariance matrix is only obtained by inversion of the normal matrix when the restraints are statistically independent and have unit variance. If this is not the case, the variance–covariance matrix must be calculated by inverting ($\mathbf{J}^T\mathbf{W}\mathbf{J}$), where $\mathbf{W}$ is the variance–covariance matrix of the restraints. Restraints are conventionally normalized to unit variance, but two special cases must be considered. The experimental e.s.d.s for the X-ray data may be unreliable and the restraints may be correlated.

In macromolecular refinement it is common to weight the X-ray restraints to correct the e.s.d.s, which are often poorly estimated by data-processing packages. This weight is an additional refinement parameter, adjusted to match the goodness-of-fit between X-ray and stereochemical restraints. *SHELXL* does not refine this weight, but rather allows the user to specify the weight if the e.s.d.s are known to be poorly estimated. The e.s.d.s of the test data gave reasonable statistics using the default *SHELXL* settings. This will have to be addressed if good estimates of model precision are to be obtained for typical structures in the PDB. The e.s.d.s for geometrical restraints are based on many well refined independently determined high-resolution structures and may therefore be considered both reliable and independent.

The estimation of parameter variances and covariances by the methods described here will be invalid when the errors in the experimental data are correlated, for example in the case of systematic errors localized in reciprocal space. Note that correlation of errors means that the error estimate for one observation is a function of the error estimates of other observations. The much more common case in which the errors are poorly estimated but depend on overall properties of the data processing rather than on the individual values of other error estimates does not lead to systematic error in the e.s.d.s. Uniformly poor estimates of the e.s.d.s of the observations merely increase the overall uncertainty in the parameters of the model.

The normal matrix and its inverse provide valuable information concerning the behavior of the refinement calculation, even if there are problems with determining the exact variance–covariance matrix. Analysis of the effects of different refinement protocols in terms of the effect on the structure of the normal matrix and its inverse can identify problems and provide quantitative measures of improvements achievable. Comparison of different parameterizations (for example, different models for restraining atomic displacement parameters) can demonstrate quantitatively the level of detail that can be supported by particular data sets. This type of analysis automatically takes account of the resolution, range and accuracy of the data, the solvent content of the crystal and the presence of non-crystallographic symmetry, all of which complicate simple resolution-based rules of thumb. Finally, even though estimates of individual parameter e.s.d.s may be in error, they are still better error estimates than anything else presently available.

5. Case study: isotropic model at 1.9 Å

5.1. The eigenvalue spectrum

Restrained and unrestrained refinement calculations were set up using the room-temperature model and 1.9 Å data. The parameter shifts in *SHELXL* were set to zero, so that no refinement was actually performed, but the normal matrix was produced in each case. For the restrained calculation, geometrical restraints were placed upon bond lengths (DFIX keyword), angles (DANG), chiral volumes (CHIV), flatness of residues and rings (FLAT), 'bumps' (BUMP) and difference in *U* value between bonded atoms (DELU), using the default restraint values and dictionary of the *SHELX* software (Sheldrick, 1997). The model had 3547 parameters; there were 11 404 data and a total of 3444 geometrical restraints were generated for the restrained calculation.

The eigenvalue spectra for the unrestrained and restrained normal matrices were calculated. The resulting spectra are shown in Fig. 1. The eigenvalues cover a dynamic range of

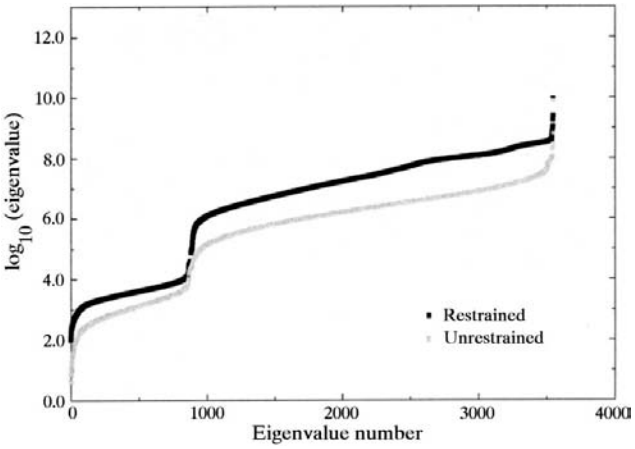

Figure 1
Eigenvalue spectra of restrained and unrestrained refinement normal matrices with 1.9 Å data.

Table 1
Number of restraints, type and desired e.s.d. for the 1.9 Å restrained refinement.

Type of restraint	*SHELXL* keyword	*SHELXL* restraint e.s.d.	Number of restraints
Bond length	DFIX	0.02 Å	862
Bond angle	DANG	0.04 rad	1177
Chiral vol.	CHIV	0.10 Å^3	149
Flat res./ring	FLAT	0.45 Å^3	262
Anti-bumping	BUMP	0.02 Å	15
Bonded atom *U*s	SIMU	0.14 Å^2	851

greater than 10^9 in each case. The *y* axis shows $\log_{10}$ (eigenvalue). This conclusively demonstrates the need for 64-bit arithmetic for this combination of parameter scales and data.

The eigenvalue spectra for both the restrained and unrestrained calculations show two distinct regions: a region of about 3000 larger eigenvalues and a region of about 1000 smaller eigenvalues. Since this model contains 959 atoms, the obvious interpretation is that the large eigenvalues correspond to combinations of positional parameters and the smaller eigenvalues correspond to combinations of thermal parameters. The separation is not total, *i.e.* every parameter has a non-zero contribution to every eigenvector, and so positional and thermal parameters may still be significantly correlated. The separation observed here is purely a result of the choice of units for the positional and thermal parameters. This may be confirmed by examining the correlation matrix, which reveals significant correlation between positional and thermal parameters.

The eigenvalues are higher for the restrained calculation than for the unrestrained case, indicating that all the parameters are better determined in the restrained case. The restrained spectrum also shows additional features. The 900 largest eigenvalues have increased more than the rest, resulting in a ridge at the top of the spectrum. This must be an effect of introducing the geometric restraints, which again suggests an interpretation. There are approximately 862 strong bond-length restraints (Table 1), resulting in a similar number of the positional parameter combinations becoming significantly better determined.

To obtain a better idea of which parameters contribute to different eigenvectors, the most significant parameters contributing to 24 eigenvectors and their contributions are listed in Table 2 for the unrestrained refinement. The eigenvectors shown are the eigenvectors with the eight largest eigenvalues, eight eigenvalues from the middle of the list and the eight smallest eigenvalues. From the table it is clear that the eigenvectors corresponding to the smallest eigenvalues are made up principally of thermal parameters of side-chain atoms in residues 18, 83, 92 and 98. These side chains are also ill-defined in the electron-density map.

The eigenvectors corresponding to the largest eigenvalues are made up principally of *z* coordinates of atoms in the Fe–S clusters. The *z* coordinates are better defined than the *x* and *y* coordinates because the *z* axis is longer and the refinement parameters are expressed as fractional coordinates. Both the X-ray and geometrical restraints are isotropic in strength (given an isotropic resolution limit to the data) when expressed in orthogonal coordinates. Since the *z* axis is longer, an isotropic perturbation in orthogonal coordinates leads to smaller changes in the fractional *z* coordinate than in the others. This effect could be isolated by conditioning the matrix (§3).

Eigenvectors 3546 and 3547 (corresponding to the two largest eigenvalues) are made up exclusively from two special parameters, WAT1 and WAT2, which are the parameters of the bulk-solvent correction. The separation of these two parameters is a result of the choice of units for those parameters. There is one further special parameter, the overall scale factor OSF. This parameter contributes mainly to eigenvectors in the region between the thermal and positional parameters – around eigenvector 880. The overall scale factor is positively correlated with the *U* values in the model (mean correlation 5.6%), but much more weakly correlated with the positional parameters, and the correlation takes either sign.

The intermediate eigenvectors contain an even mix of many atomic parameters. Unlike the extreme eigenvectors, no one parameter contributes strongly to each eigenvector. This distribution can be seen more clearly in Fig. 2, where the number of parameters making significant contribution to each eigenvector is plotted against eigenvalue number. A 'significant contribution' is defined as a contribution of more than $1/n$, where n is the number of parameters. Equal eigenvalues imply degenerate eigenvectors. Thus, in the flat regions of the eigenvalue spectrum where there are many eigenvalues of similar magnitude, the eigenvectors combine contributions from many parameters. However, where the spectrum is steep and the eigenvalues have significantly different magnitudes, the eigenvectors only combine contributions from a few parameters. This highlights the fact that the eigenvalue calculation separates parameter combinations according to curvature of the residual: when there are many parameters whose variation leads to a similar effect of the residual, the eigenvectors form sets of orthogonal combinations of those

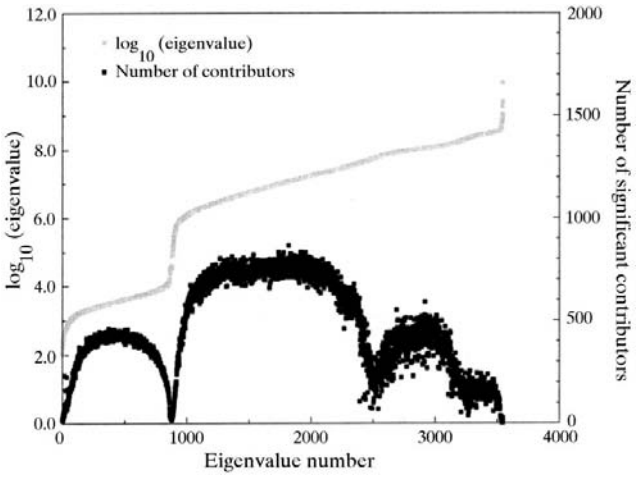

Figure 2
Number of significant parameters contributing to each eigenvector, unrestrained refinement with 1.9 Å data.

Table 2
Strongest contributions from parameters to eigenvectors for (*a*) the eight smallest eigenvalues, (*b*) eight median eigenvalues amd (*c*) the eight largest eigenvalues, 1.9 Å unrestrained calculation.

The eigenvalue numbers are given in the first column. Each entry gives the contribution (%), the parameter type and the atom name, respectively.

(*a*) Eigenvectors for eight smallest eigenvalues.

1	78 *U*	18 *U*	3 *U*	1 *U*	0 *y*	0 *U*	0 *U*	0 *U*
	OE1 92	CD 92	OE2 92	CG 92	CD 92	OE1 18	OE2 18	CD 18
2	66 *U*	25 *U*	8 *U*	1 *U*	0 *U*	0 *U*	0 *U*	0 *y*
	CD 92	OE2 92	OE1 92	CG 92	CE 98	OE2 18	CD 18	CD 92
3	58 *U*	27 *U*	12 *U*	2 *U*	1 *U*	0 *U*	0 *U*	0 *U*
	CE 98	CD 98	NZ 98	CG 98	CD 83	OE2 18	CD 92	OE1 18
4	55 *U*	32 *U*	5 *U*	4 *U*	1 *U*	1 *U*	0 *U*	0 *U*
	NZ 98	CD 98	CG 98	OE1 18	CE 98	OE2 18	CD 18	OE2 83
5	63 *U*	24 *U*	6 *U*	3 *U*	2 *U*	0 *U*	0 *U*	0 *U*
	OE1 18	CD 18	OE2 18	NZ 98	CD 98	CB 18	CG 98	OE2 92
6	45 *U*	39 *U*	6 *U*	5 *U*	2 *U*	1 *U*	0 *U*	0 *U*
	CD 18	OE2 18	CG 92	OE1 18	OE2 92	CD 92	CG 18	CD 98
7	50 *U*	34 *U*	5 *U*	4 *U*	3 *U*	2 *U*	0 *U*	0 *U*
	CG 92	OE2 92	CD 18	CD 92	OE2 18	OE1 92	NZ 98	CD 98
8	35 *U*	26 *U*	26 *U*	6 *U*	4 *U*	1 *U*	1 *U*	0 *U*
	CE 98	CD 98	NZ 98	CD 83	CG 98	OE2 83	CG 92	CD 92

(*b*) Eigenvectors for eight median eigenvalues.

1770	1 *y*	1 *x*	1 *y*	1 *x*	1 *x*	1 *y*	1 *y*	1 *z*
	ND2 80	CB 11	C 5	C 67	O 127	C 11	CB 15	CE 85
1771	1 *y*	1 *z*	1 *x*	1 *x*	1 *y*	1 *z*	1 *x*	1 *y*
	CD2 101	OD1 58	CG1 17	NE2 35	O 77	CG 52	CB 61	O 119
1772	1 *x*	1 *y*	1 *x*	1 *y*	1 *z*	1 *z*	1 *x*	1 *y*
	O 129	CG 71	N 83	O 114	CE1 35	CD2 35	CD1 26	C 19
1773	1 *x*	1 *z*	1 *y*	1 *x*	1 *x*	1 *x*	1 *y*	1 *x*
	NZ 84	CB 77	CB 72	O 119	CG2 82	O 77	CA 68	NE2 52
1774	1 *x*	1 *x*	1 *y*	1 *z*	1 *z*	1 *x*	1 *y*	1 *x*
	CB 56	CG2 14	N 58	CE- 10	OE2 83	CD 65	CE 12	CD2 32
1775	1 *y*	1 *y*	1 *x*	1 *y*	1 *y*	0 *y*	0 *y*	0 *y*
	O 114	CD2 26	O 119	O 84	CA 37	N 20	CA 65	CB 102
1776	1 *y*	1 *x*	1 *x*	1 *y*	1 *y*	1 *x*	1 *x*	1 *x*
	CD 21	CD1 2	CZ 25	CG2 54	OD2 15	CG2 14	CD2 31	N 94
1777	1 *y*	1 *x*	1 *x*	1 *z*	1 *y*	1 *z*	1 *y*	1 *x*
	O 135	O 126	CB 106	CE1 35	OE1 57	O 120	CG 35	CD1 67

(*c*) Eigenvectors for eight largest eigenvalues.

3540	30 *z*	29 *z*	25 *z*	6 *z*	2 *z*	2 *z*	1 *z*	1 *z*
	Fe4 107	Fe2 107	Fe3 107	Fe1 108	Fe2 108	S1 107	Fe1 107	S2 107
3541	47 *z*	15 *z*	13 *z*	12 *z*	7 *z*	1 *z*	1 *z*	0 *z*
	Fe1 108	Fe2 108	Fe3 107	Fe1 107	Fe2 107	S2 108	S1 108	Fe4 107
3542	63 *z*	20 *z*	9 *z*	2 *z*	1 *z*	1 *z*	1 *z*	0 *z*
	Fe2 108	Fe1 108	Fe2 107	Fe3 108	Fe1 107	Fe3 107	S4 108	S3 108
3543	38 *z*	21 *z*	13 *z*	12 *z*	9% *z*	2 *z*	1 *z*	1 *z*
	Fe2 107	Fe1 107	Fe4 107	Fe1 108	Fe2 108	Fe3 108	Fe3 107	SG 39
3544	55 *z*	13 *z*	12 *z*	9 *z*	5 *z*	2 *z*	1 *z*	0 *z*
	Fe1 107	Fe3 108	Fe4 107	Fe2 107	Fe3 107	Fe1 108	S3 107	SG 39
3545	79 *z*	5 *z*	4 *z*	4 *z*	2 *z*	1 *z*	1 *z*	1 *z*
	Fe3 108	Fe1 107	Fe2 107	Fe2 108	Fe4 107	S2 108	S4 108	Fe3 107
3546	83 *	17 *	0 *z*	0 *z*	0 *z*	0 *y*	0 *z*	0 *z*
	WAT2	WAT1	Fe1 107	Fe1 108	S1 107	Fe2 107	S3 108	Fe2 108
3547	83 *	17 *	0 *z*	0 *z*	0 *z*	0 *	0 *y*	0 *x*
	WAT1	WAT2	Fe1 107	S1 107	S3 108	OSF	Fe2 107	Fe3 107

similar parameters. When the curvature of a parameter has a unique value and is not correlated with another parameter, it will contribute to only a single eigenvector.

5.2. Classification of eigenvectors in terms of parameters

Further information concerning features of the model which are well or ill-determined may be obtained by dividing the refinement parameters into a small number of classes and then examining which classes of parameter contribute to each eigenvector. The parameters were divided into the following classes: (i) special parameters (overall scale factor, bulk-solvent parameters), (ii) parameters of atoms of the Fe–S clusters, (iii) parameters of solvent atoms, (iv) parameters of protein main-chain atoms, (v) parameters of protein side-chain atoms.

Each class of atomic parameters was further divided into positional and thermal parameters, giving a total of nine classes. The contribution from each class of parameter to every eigenvector is shown as a stacked bar chart. Colours indicate the parameter classes, with thermal parameters distinguished by lighter shades of the positional parameters.

Fig. 3 shows the class plot for the unrestrained calculation. The plot shows a clear separation between positional parameters (contributing to the eigenparameters of higher curvature, on the right of the plot) and thermal parameters (contributing to the eigenparameters of lower curvature, on the left).

Most of the Fe–S positional parameters are very well determined. There are also some very well determined side-chain positional parameters. Examination of the eigenparameters of highest curvature reveals these side-chain parameters to be coordinates of S atoms, which contribute more to scattering than the bulk of main-chain and side-chain atoms.

The worst defined positional parameters (around eigenvalue 950) include mainly contributions from side-chain positional parameters. These are generally parameters of atoms in long or apparently disordered side chains.

The distribution of parameter classes on the left of the plot is very similar to that on the right of the plot, with thermal parameters replacing positional parameters. The solvent

atoms appear to have been cautiously chosen, since neither their positional or thermal parameters are at the bottom of the respective regions of the spectrum.

Fig. 4 shows the corresponding plot for the restrained calculation. The main features are similar. However, the distribution of classes contributing to about 1000 of the best defined eigenparameters has been perturbed. This is consistent with the perturbation of the corresponding region of the eigenvalue spectrum. The strong bond-length restraints mean that many combinations of protein-atom positional parameters are now stronger than some of the Fe–S cluster parameters.

5.3. Effect of restraints

To examine the effect of various restraints on the conditioning of the problem, the normal matrices were calculated adding successive restraints to the unrestrained calculation in order to obtain some indication of the effect of various restraints. The eigenvalue spectra for all of the calculations are shown in Fig. 5. The types and numbers of different geometrical restraints are listed in Table 1.

Addition of bond-length restraints to the unrestrained calculation makes a significant difference to the shape of the spectrum. All of the eigenvalues increase, but in particular the 900 largest eigenvalues increase significantly more than the others, justifying the previous conclusions about this group of eigenparameters.

Addition of the bond-angle restraints significantly increases the remaining eigenvalues not affected by the bond lengths. This is consistent with the bond-angle restraints acting mainly perpendicular to the bond-length restraints. The few remaining positional restraints (chiral volumes, flat rings, anti-bumping) give a slight further increase in the remaining eigenvalues.

Restraining the U values of neighboring atoms increases the eigenvalues in the thermal region of the spectrum, including the very smallest eigenvalues.

The plot also shows that the positional restraints (particularly bond lengths and angles) do affect the smaller eigenvalues, which are mainly from thermal parameters. This supports the observation that the positional and thermal parameters are not separable.

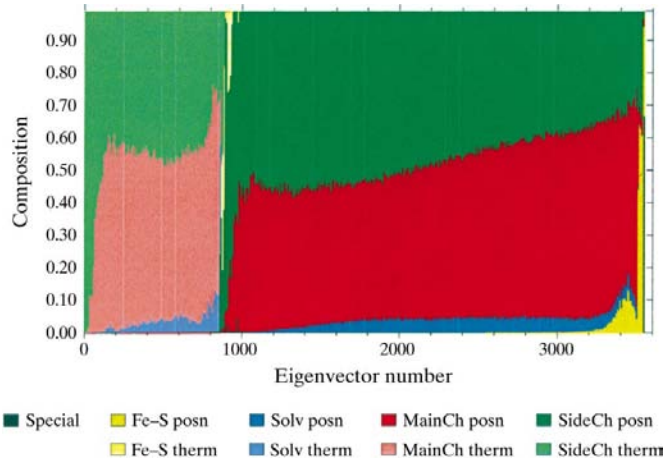

Figure 3
Class plot of eigenvector composition in terms of parameter classes against eigenvector number for unrestrained refinement with 1.9 Å data.

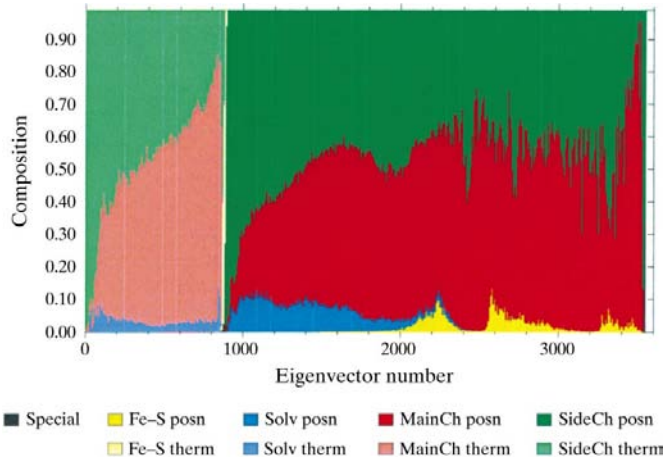

Figure 4
Class plot of eigenvector composition in terms of parameter classes against eigenvector number for restrained refinement with 1.9 Å data.

5.4. Visualization of eigenvectors

The shape of the eigenvalue spectrum and the effects of different restraints may be further understood by plotting some of the eigenvectors in three dimensions. Each eigenvector describes a unit vector in the parameter space. This can be visualized using a three-dimensional model of the molecule and attaching a vector to each atom based on the projection of the eigenvector onto the positional parameters of that atom. (At present, projections of the eigenvector onto U values are ignored.)

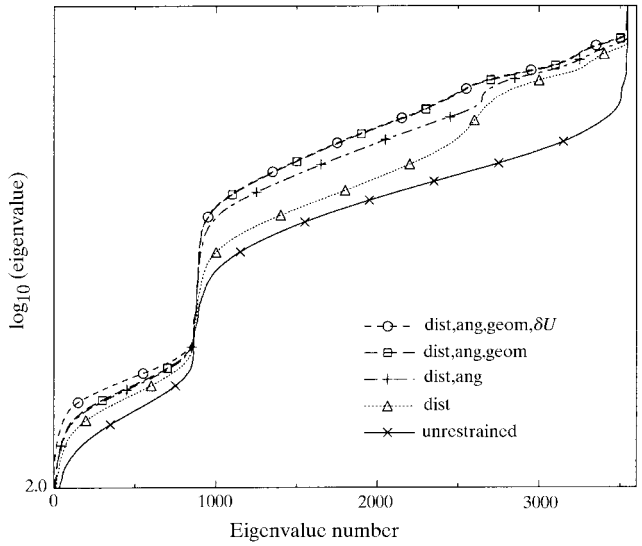

Figure 5
Effect of various geometrical restraints on the eigenvalue spectrum with 1.9 Å data. Note: line styles are obscured where lines overlap.

The resulting three-dimensional object indicates how the atoms of the model will move if a small change is made to that eigenparameter. The change can be positive or negative, with the graphical representation showing the positive direction. The vectors have been scaled so that an eigenvector which contains contributions from only a single atom will be 5 Å in length.

Fig. 6 shows the region of the model around residue 77 with eigenvector 1066 superimposed. This corresponds to an eigenvalue at the low end of the positional region of the spectrum. The eigenvector is involved in twisting the valine side chain about its bond, a motion which is not affected by any of the geometrical restraints. The same eigenvector also contributes to similar effects elsewhere in the structure.

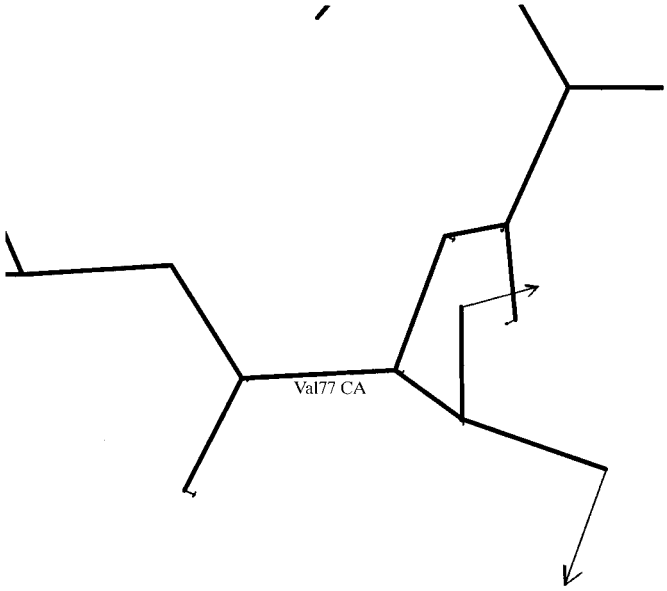

Figure 6
Eigenvector 1066: this poorly determined eigenparameter is strongly involved in the rotation of the valine side chain.

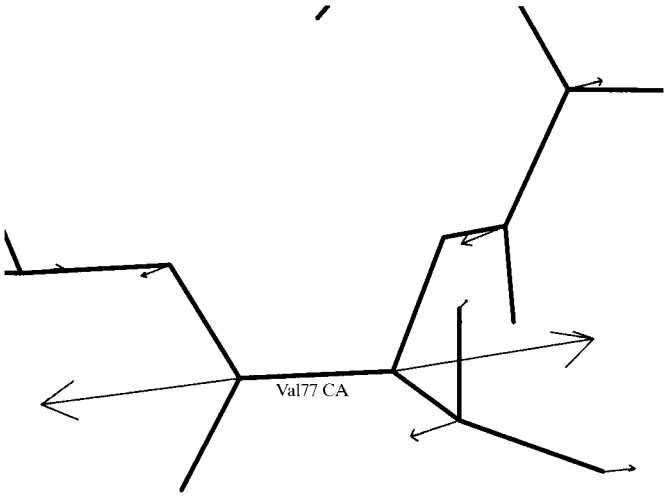

Figure 7
Eigenvector 3497: this well determined eigenparameter is strongly involved in the length of a main-chain bond.

Fig. 7 shows the region of the model around residue 77 with eigenvector 3497 superimposed. This corresponds to an eigenvalue at the high end of the positional region of the spectrum, on the ridge which has been attributed to bond length e.s.d.s. It is principally involved in the stretching of a main-chain bond (with lesser contributions to some of the other nearby bonds), a motion which is strongly affected by the geometrical restraints.

Fig. 8 shows the region of the model around residue 78 with eigenvector 3546 superimposed. This corresponds to the largest eigenvalue in the spectrum, excluding the eigenvectors representing the bulk-solvent correction. The strongest eigenvectors in the restrained calculation all correspond to distortions of tryptophan side chains (contrast this with the unrestrained calculation in Table 2, where the best determined parameters were those of the Fe–S clusters). The flatness restraint is clearly having a very strong effect on the few atoms to which it is applied.

5.5. The RHS of the least-squares equations

The RHS of the least-squares equations (the RHS vector) is obtained by premultiplying the residual vector of restraint disagreements by the matrix of derivatives with respect to the parameters. In theory, at the end of a refinement the RHS vector of the least-squares equations should be zero, otherwise technically the refinement has not converged. However in practice this is never the case, for a number of reasons.

(i) The eigenvalues have a large dynamic range. Small shifts to well determined parameters will perturb the normal matrix sufficiently that estimates for the ill-determined parameters will be subject to large errors. Ill-determined parameters may therefore not refine until all the well determined parameters have been fully refined.

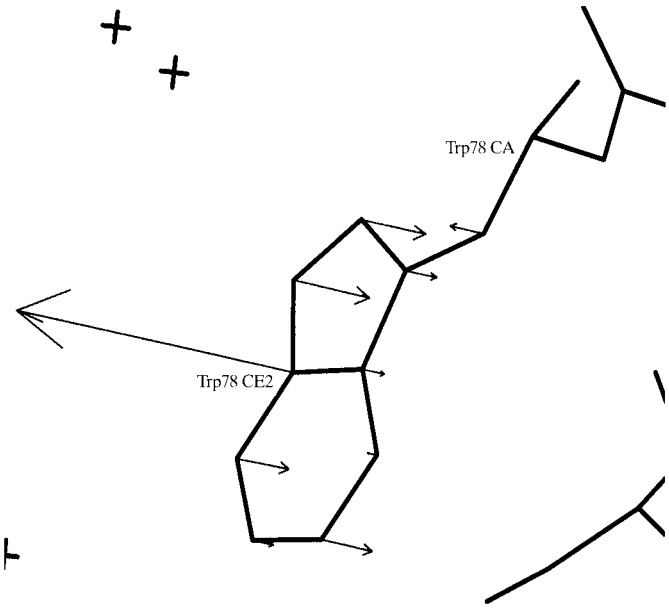

Figure 8
Eigenvector 3546: this very well determined eigenparameter is involved in the flatness of a tryptophan.

(ii) The normal matrix may not be slowly varying. If the normal matrix varies rapidly with some parameters, those parameters will refine very slowly or not at all. This is because in such cases the second-order approximation to the sum of squares of residuals is poor.

(iii) Limits of machine precision will introduce noise at all stages.

The remaining RHS vector of the normal equations can be projected onto the eigenvector axes (formed by taking the dot product of the RHS vector with each eigenvector in turn). The square of this projection is plotted as a function of eigenvector number in Fig. 9, along with the eigenvalue spectrum. The graph shows the curvature along each eigenvalue direction and the squared contribution to the RHS vector along that direction.

Both curves have a common shape, from which it is clear that although there is considerable variation from parameter to parameter, the maximum residual contributions from any parameter is closely related to the curvature of the refinement residual. Since the refinement calculation will be most strongly influenced by the largest remaining residual, it tends to partition the remaining residual equally amongst the parameters, thus matching the shape of the eigenvalue spectrum.

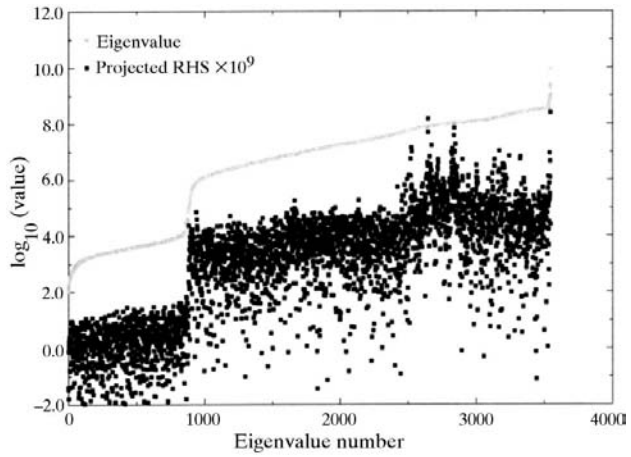

Figure 9
Square of the projection of the RHS onto the eigenparameters as a function of eigenparameter number, in comparison with the eigenvalue spectrum. The curves show a broadly similar shape.

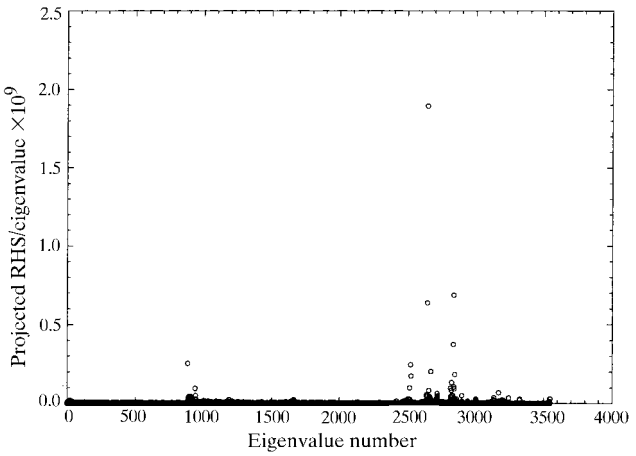

Figure 10
Ratio of the square of the projection of the RHS onto the eigenparameters to eigenvector as a function of eigenparameter number. The ratio clearly shows sharp peaks where the restraints are strongly dissatisfied.

However, sharp features are apparent in the residual spectrum around eigenvalues 2647 and 2840. These features become clearer if the ratio of the RHS spectrum to the eigenvalue spectrum is calculated, shown on a linear axis in Fig. 10. Taking this ratio normalizes the RHS contributions, since the eigenvalues are the inverse variances of the eigenparameters. It should be possible to apply a significance test to identify informative outliers in this plot.

Eigenvector 2647, corresponding to the highest peak in Fig. 10, is shown in Fig. 11. The eigenparameter includes strong contributions from the positional parameters for the side chain of residue 98. Residue 98 is a surface lysine for which the density is particularly poor. Variation of the eigenparameter corresponds to stretching and compressing bonds along the side chain. The restraint dissatisfaction has become concentrated in this eigenparameter because the X-ray terms (reflected in the map density) and the geometric restraints are irreconcilable for this side chain.

Comparison of this refinement with the high-resolution refinement (§6) suggests that this side chain is sufficiently disordered that modeling a second conformation would not help. The ratio plot has therefore revealed a genuine problem area in the structure. It is unlikely that a small number of fixed conformations adequately model the physical reality of this portion of the structure.

Note that peaks only appear in the RHS/eigenvalue ratio plot after full-matrix refinement has been applied to near-convergence. Models which have been refined using the conjugate-gradient method or which have not been refined to

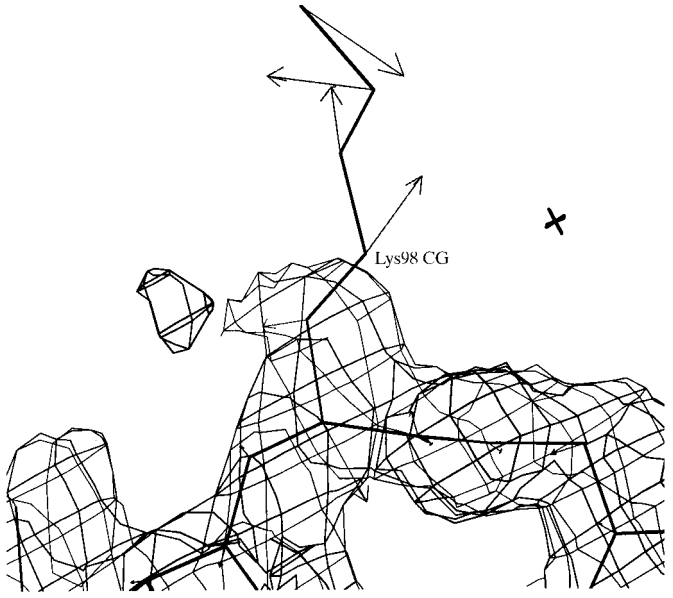

Figure 11
Eigenvector 2647: this eigenvector corresponding to a peak in the RHS spectrum. The side chain is one of the worst defined in the structure.

near convergence tend not to show sharp peaks. With an unrefined model, the constraint dissatisfaction is spread amongst all the parameters. As the refinement approaches convergence, the dissatisfaction is concentrated into those parameters for which the quadratic approximation is a poor description. In this case, the dissatisfaction is concentrated in those parameter combinations for which the model does not describe the X-ray data well and therefore are among the last to refine.

The RHS contribution/eigenvalue ratio has a particular statistical significance (see, for example, Kendell *et al.*, 1991). The inverse eigenvalues are the variances of the eigenparameters and the refinement shifts to the eigenparameters are given by the RHS vector in the eigenparameter space divided by the eigenvalues. Therefore, the RHS contribution/eigenvalue ratio represents the χ^2 normalized eigenparameter shift. This function may also be calculated in parameter space for a direct indication of the significance of the refinement shift applied to each parameter.

6. Anisotropic model at 1.3 Å

To test the usefulness of these methods at a higher resolution, a second set of test data were used. X-ray data is available from frozen crystals of the same ferredoxin protein. This data extends to 1.3 Å. The structure was re-refined by Stout *et al.* (1998) using *SHELXL* to produce a model with anisotropic thermal parameters. Additional solvent atoms are also present in this model, bringing the total number of atoms to 1018.

6.1. The eigenvalue spectra

Restrained and unrestrained refinement calculations were set up using the anisotropic model and the low-temperature data to 1.3 Å. *SHELXL* was again used to calculate normal matrices in each case. In addition to the geometrical restraints used for the isotropic model, additional restraints were placed on the anisotropic thermal parameters requiring equal thermal motion along the bond axis of bonded atoms (SIMU keyword) and restraining water atoms to be roughly isotropic (ISOR). The model had 9165 parameters, there were 30 880 X-ray data and 10 718 geometrical restraints were generated for the restrained calculation.

The eigenvalue spectra for the unrestrained and restrained normal matrices are shown in Fig. 12. The eigenvalue spectra each show two regions, one of about 3000 large eigenvalues and another of about 6000 smaller eigenvalues. This is again consistent with three positional parameters per atom and six thermal parameters.

In contrast to the 1.9 Å calculation, the eigenvalue spectra for the restrained calculation does not show any additional feature owing to the bond-length restraints and the increase in eigenvalues after application of the restraints is less pronounced. At this resolution, the X-ray terms are dominant in determining the positional parameters and thus the impact of the geometrical restraints is reduced.

6.2. Classification of eigenvectors in terms of parameters

The class plot for the 1.3 Å unrestrained refinement is shown in Fig. 13. The pattern of parameter contributions to the eigenparameters is similar to the lower resolution case, although there are now far more thermal parameters. The additional solvent atoms in this model cluster at the bottom of both the positional and thermal regions of the spectrum, indicating that these atoms are comparatively less well determined than the solvent atoms from the room temperature model.

The class plot for the restrained refinement is shown in Fig. 14. There is some perturbation to the distribution of parameters amongst the eigenvectors of largest eigenvalue in comparison to the unrestrained case, but the perturbation is less than for the room-temperature refinement. This confirms that the impact of the geometrical restraints is smaller in this case. In contrast to the room-temperature case, there is now some perturbation of the thermal parameter contributions from the unrestrained case. Again around 1000 eigenparameters have been perturbed, suggesting that this perturbation

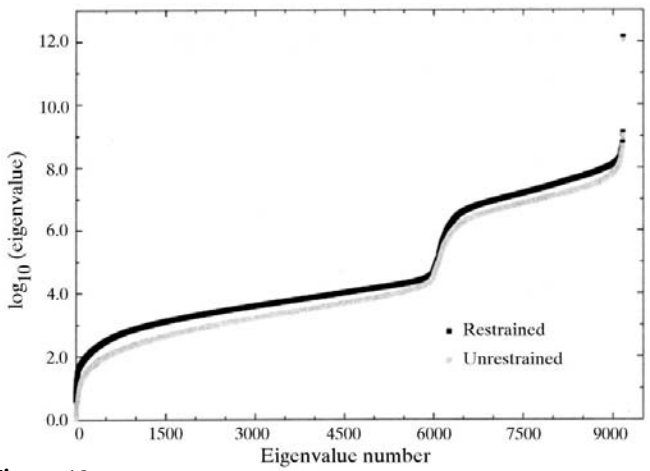

Figure 12
Eigenvalue spectra of restrained and unrestrained refinement normal matrices with 1.3 Å data.

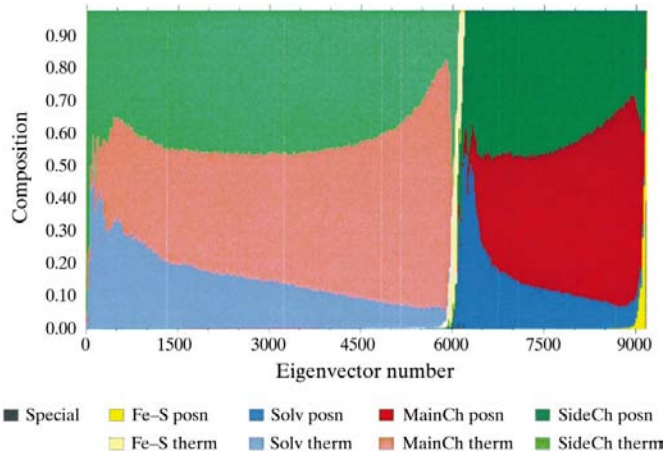

Figure 13
Class plot of eigenvector composition in terms of parameter classes against eigenvector number for unrestrained refinement with 1.3 Å data.

Table 3
Number of parameters, e.s.d. of a Fe-atom position and magnitude/intensity *R* factors as thermal parameters are grouped for unrestrained refinement against the 3.0 Å data (3230 data).

The model is seriously over-parameterized in all cases.

Grouping	No. of params	No. undetermined	Fe1–107 position e.s.d. (Å)	Goodness of fit	*R* factors R_1/wR_2
None	3547	317	∞	∞	0.196/0.543
CA—C—N	3335	105	∞	∞	0.198/0.553
CA—C—N—O	3230	0†	2.1×10^7	∞	0.201/0.566
CA—C—N—O, CB—CG	3162	0	3.3	45.5	0.203/0.571
All main, all side	3020	0	0.8	26.8	0.207/0.586

† $n_r = n$, thus the matrix is non-singular but the goodness-of-fit is infinite.

arises from the rigid-bond restraints on thermal parameters (the *SHELXL* SIMU restraint).

6.3. The RHS spectrum for the restrained calculation

The ratio of RHS contributions to eigenvalues for the restrained calculation at 1.3 Å is plotted in Fig. 15(*a*). This graph again shows some sharp peaks, but in contrast to the room-temperature case the largest feature is at the top end of the thermal region of the spectrum.

The eigenvectors corresponding to the highest peaks in this plot are listed in Table 4 in terms of their largest contributors. The eigenvectors in the large peak all include contributions from S-atom thermal parameters, which might be expected to be well determined. The common and distinctive feature of these eigenvectors is contributions from thermal parameters of atoms in the side chain of residue 18. The lesser peak comprises contributions from the long flexible side chains of residues 83 and 92 (both glutamates).

The model for residue 18 is shown in Fig. 16, with the electron density. The electron density is poor and the thermal ellipsoids for these atoms are extremely anisotropic. The thermal parameters appear to be trying to fit absent density and are in disagreement with the geometrical restraints. There is density to suggest an alternate, possibly better, conformation.

The side chain was moved into the alternative density and subjected to local refinement and the whole model was then refined using full-matrix least-squares refinement. The new side chain displayed good density and the *R* factor dropped, but only after refinement of the whole model (Table 5). There were no major motions during this refinement, but most parameters displayed small shifts, confirming that the whole model had been biased by the incorrect side chain. The RHS/eigenvalue ratio plot for the refined model (Fig. 15*b*) shows that the

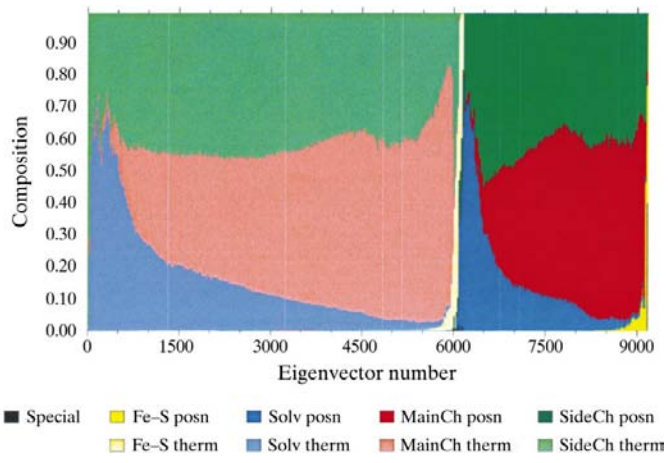

Figure 14
Class plot of eigenvector composition in terms of parameter classes against eigenvector number for restrained refinement with 1.3 Å data.

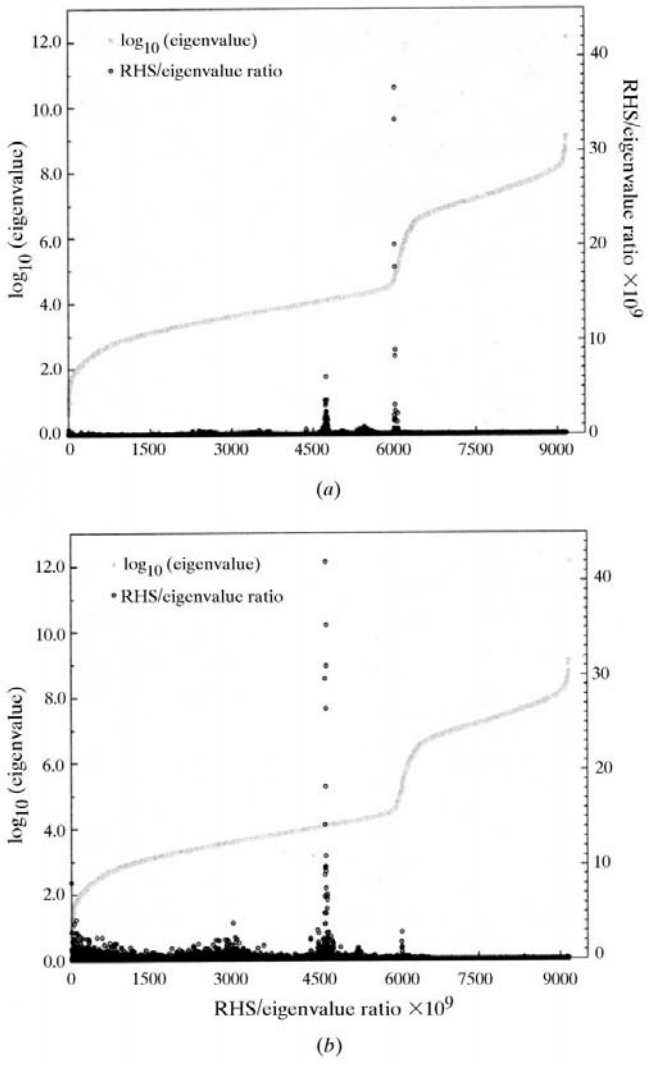

Figure 15
RHS/eigenvector ratio for the 1.3 Å data and (*a*) initial refined 1.3 Å model, (*b*) after modification of the side chain and refinement of the model.

Table 4
Strongest contributions from parameters to eigenvectors for peaks in the RHS/eigenvalue ratio for the initial refined 1.3 Å model.

Eigenvalue No.	RHS ev$\times10^{-9}$	Contribution (%), parameter type, atom name, respectively							
4737	4.1	2 U_{13} OE2 92	1 U_{13} CD 92	1 U_{11} OE2 92	1 U_{11} CD 92	1 U_{33} CG2 82	1 U_{22} CG 62	1 U_{23} CB 82	1 U_{33} OE2 92
4754	2.5	2 U_{23} OE2 83	1 U_{23} CD 83	1 U_{13} OE2 92	1 U_{22} OE2 83	1 U_{13} O 104	1 U_{13} O 62	1 U_{22} CD 83	1 U_{13} CD 92
6005	13.8	31 U_{13} SG 45	11 U_{12} SD 64	8 U_{23} S1 108	8 U_{13} OE2 18	6 U_{12} S1 108	4 U_{33} OE2 18	2 U_{13} SG 49	2 U_{13} SG 20
6006	25.2	25 U_{13} SG 45	13 U_{13} OE2 18	11 U_{12} SD 64	8 U_{13} SG 49	6 U_{33} OE2 18	5 U_{23} SG 24	2 U_{11} OE2 18	2 U_{23} OE2 18
6007	12.1	16 U_{12} SD 64	12 U_{23} S1 108	9 U_{12} S1 108	6 U_{13} OE2 18	5 U_{23} SG 11	4 U_{13} SG 16	4 U_{13} SG 49	4 U_{13} S1 108
6009	22.9	26 U_{13} SG 49	12 U_{13} S2 107	12 U_{13} OE2 18	5 U_{33} OE2 18	4 U_{13} SG 20	3 U_{22} SG 49	2 U_{23} SG 11	2 U_{11} OE2 18

original peak has disappeared, but the secondary peak arising from the Glu side chains has increased.

Residue 18 was re-examined in the room-temperature model. There was no clear density for the side chain, even with the low-temperature model as a guide. The residue does not contribute to any clear peaks in the RHS/eigenvalue ratio plot at low resolution. It is possible that this side chain is disordered at room temperature and ordered at low temperature.

7. Lower resolutions (2.0–3.0 Å)

An examination of the behavior of the refinement at 2.0 Å and worse resolution was carried out using the 1.9 Å resolution *X-PLOR* model with the room-temperature data truncated to the desired resolution. The model is therefore superior to that which could be obtained from the truncated data alone, but use of a fixed model allows a more direct comparison of the effect of reducing the number of X-ray terms in the calculation of the normal matrix.

Restrained and unrestrained calculations were set up at 2.0 (10 009 data), 2.5 (5427 data) and 3.0 Å (3230 data) resolution. The normal matrix was calculated for each case. The eigenvalue spectra for the restrained refinements are shown in Fig. 17. The model and geometric restraints in each case are identical; the only difference is the number of X-ray terms. The 1000 largest eigenvalues are similar in all three cases, suggesting that these combinations are determined principally by the geometrical restraints: this includes the parameter combinations attributed to bond-length restraints.

As the resolution worsens, the remaining eigenparameters become less well determined, with the thermal region of the spectrum dropping most. The reduction in the number of X-ray data increases the difference in scale between the positional and thermal parameters. This does *not* prove that the thermal parameters are more sensitive to data resolution than the positional parameters. Indeed, it seems reasonable to expect that as the thermal parameter of a poorly defined atom becomes undefined, the positional parameters will also become undefined.

The eigenvalue spectra for the unrestrained refinements are shown in Fig. 18. At 2.0 Å resolution, the problem is still well conditioned (over a single step; however, this does not mean that multiple steps of refinement will converge). At 2.5 Å resolution, all the eigenparameters are still defined, but the matrix is nearly singular. At 3.0 Å, there are 316 eigenvalues which are zero or approximately zero.

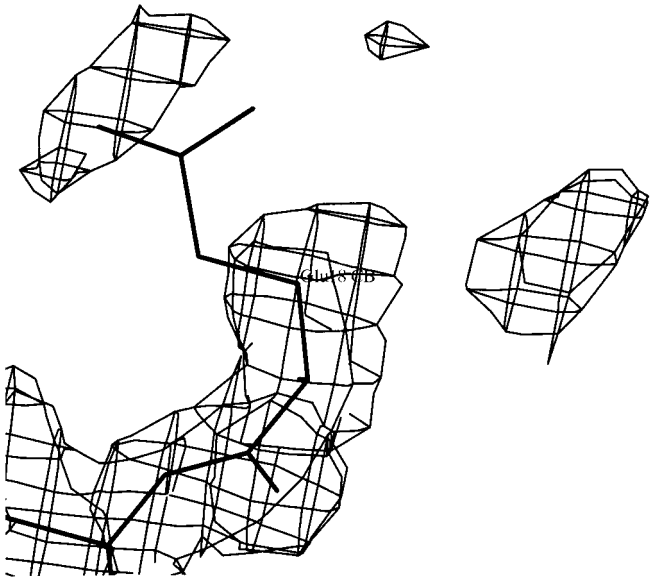

Figure 16
Residue 18 and density from the initial 1.3 Å model.

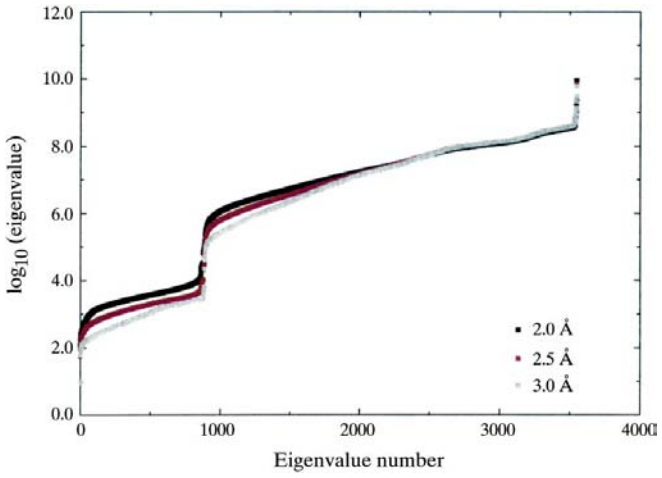

Figure 17
Eigenvalue spectra for restrained refinement of 1.9 Å model and truncated data at 2.0, 2.5 and 3.0 Å.

7.1. Estimation of e.s.d.s

Let $\mathbf{v}$ be a vector describing a combination of parameters for which the e.s.d. is required. The e.s.d. of $\mathbf{v}$ is given by the square root of the variance, given by $\mathbf{v}^T\mathbf{N}^{-1}\mathbf{v}$. (This formulation includes all the appropriate covariance terms.)

Let $\mathbf{Q}$ be the matrix whose columns are the eigenvectors and $\mathbf{\Lambda}$ be the diagonal matrix whose diagonal elements are the eigenvalues. Then

$$\mathbf{N} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T \tag{11}$$

$$\mathbf{N}^{-1} = \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^T. \tag{12}$$

The variance may also be calculated using the projection of v into the eigenparameter space,

$$\begin{aligned}\sigma_v^2 &= \mathbf{v}^T\mathbf{N}^{-1}\mathbf{v} \\ &= \mathbf{v}^T\mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^T\mathbf{v} \\ &= (\mathbf{Q}^T\mathbf{v})^T\mathbf{\Lambda}^{-1}(\mathbf{Q}^T\mathbf{v}).\end{aligned} \tag{13}$$

Since $\mathbf{\Lambda}^{-1}$ is the diagonal matrix of inverse eigenvalues, this simplifies to

$$\begin{aligned}\mathbf{v}' &= \mathbf{Q}^T\mathbf{v} \\ \sigma_v^2 &= \sum_{i=1,n} v_i'^2/\lambda_i.\end{aligned} \tag{14}$$

When the normal matrix becomes singular, some of the λ_i become zero and so in general it becomes impossible to calculate e.s.d.s. However, even in this case there are two possible approaches which may allow estimation of e.s.d.s for some parameters or parameter combinations.

(i) Isolate a subspace of the parameter space containing all the parameters which contribute to the undetermined subspace. e.s.d.s of combinations of the remaining parameters may be estimated normally. An e.s.d. can be calculated for any parameter combination where $v_i' = 0$ for every i for which $\lambda_i = 0$ in (14).

(ii) Identify the parameters which contribute significantly to the undetermined eigenparameters and reduce the parameterization of the model in a reasonable way to remove these parameters.

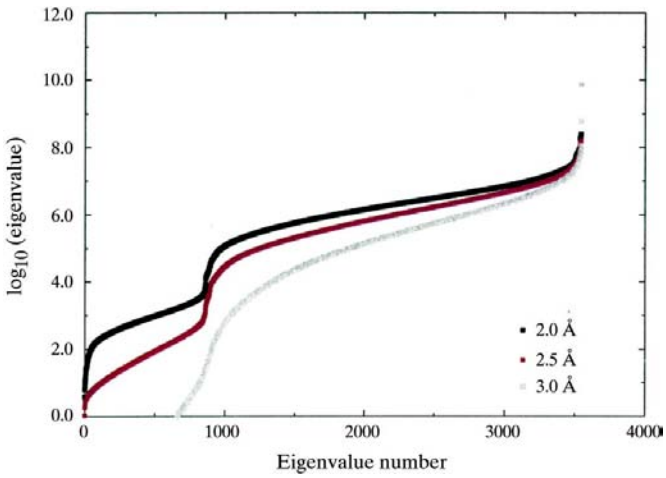

Figure 18
Eigenvalue spectra for unrestrained refinement of 1.9 Å model and truncated data at 2.0, 2.5 and 3.0 Å (un-normalized).

Table 5
Change in magnitude and intensity R factor during modeling of residue 18 side chain.

Refinement stage	Conventional R factor (R_1)	Intensity R factor (wR_2)
Initially refined model	0.1501	0.3585
Side chain 18 moved	0.1514	0.3715
Local refinement (5 residues)	0.1505	0.3673
Refinement of whole model	0.1481	0.3552

The first option initially seemed reasonable, but the ratio of the eigenvalues between the positional and thermal regions ranges from 10^2 to 10^6, so even a tiny contribution from a positional parameter to a poorly determined eigenvector can have a significant impact on the e.s.d.s. Conditioning the normal matrix would reduce the disparity of scale, but would also mix the positional and thermal parameters.

Examining the contributions to the undetermined eigenparameters at 3.0 Å, it appears that not even the positional parameters of the Fe atoms can be determined since they have significant contributions to undetermined eigenparameters. To confirm this result, the contributions to the e.s.d.s of the Fe parameters were examined for the non-singular matrix at 2.0 Å, but even in this case the most ill-defined eigenparameters contribute significantly to the Fe-atom positional e.s.d.s. The problem must therefore be reparameterized.

Several reparameterizations of the model were tested, replacing individual atomic Us with grouped Us within each residue. The model was not been re-refined, but individual Us were replaced by their average values within the group. The type of grouping, number of parameters, number of undetermined eigenparameters and R factor are listed in Table 3. This table also gives the positional e.s.d. of Fe1–107, in the cases where the matrix is non-singular. [These values are not normalized by goodness-of-fit and give at best a relative indication of e.s.d., for reasons given by Schwarzenbach *et al.* (1989). The goodness-of-fit is listed for reference.] As the number of parameters is reduced, the R factor increases. However, at the same time the problem becomes non-singular and the parameter e.s.d.s improve. (Note that the model is still severely over-parameterized at this resolution, as shown by the goodness-of-fit.)

The reduced-parameter model does not fit the data so well, but its parameters may be determined more precisely. This is in accordance with general experience from the free R factor (Kleywegt & Jones, 1995), although in this case the result has been obtained without performing any further refinement on the reparameterized model.

8. Conclusions

The calculation of the eigenvalue spectrum of the refinement normal matrix provides detailed information about the conditioning of the refinement problem. The spectrum itself

provides information about the range of variances for model parameters of a particular type and about the numerical precision required to perform the refinement with a particular choice of parameter units.

The comparison of different classes of parameters provides information about the reliability of modeling different classes of features within the structure. This information might be used to determine when a reasonable number of solvent atoms have been included or whether restraints may be added or removed for some features. Comparison of the eigenvalue spectra obtained with different combinations of geometrical restraints gives insight into the impact of the different types of restraint on the model.

The RHS spectrum gives valuable insight into the failure of refinement calculations to fully converge. If a part of a model is wrong and the geometry is not fully restrained, then refinement will often produce an unreasonable geometric configuration in the problem region. Adding geometrical restraints stabilizes the refinement, at a cost of concealing some of these problem regions. The projected RHS spectrum reveals regions where the disagreement between X-ray and other restraints is preventing convergence, thus allowing the location of problem regions which might otherwise have been missed owing to the effectiveness of the geometric restraints.

These techniques, while still in the early stages of development, offer considerable scope for improving the understanding of structure refinement. The insights presented here are on the whole in agreement with protocols developed through the practical application of refinement over many years (see, for example, Kleywegt & Jones, 1997); however, it is hoped that the tools presented here will provide a simpler and more objective means for the development and testing new refinement protocols. Further work is already in progress to examine the effect of conditioning the normal matrix and how the resulting information may be applied to stabilize an underdetermined refinement. Future objectives of this work will include the application of the information available from eigensystem analysis to identification of new parameterizations and geometric restraints and the comparison of the normal matrices obtained using least-squares and maximum-likelihood residuals.

The authors are grateful to the National Science Foundation for funding this work (grant DBI 9616115), the San Diego Supercomputer Center for use of facilities and Duncan McRee and Chris Stout for providing the test data. The authors would like to thank the referees for suggesting significant improvements to the paper.

References

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A. & Sorensen, D. (1999). *LAPACK User's Guide*, 3rd ed. Philadelphia: Society for Industrial and Applied Mathematics.

Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.

Diamond, R. (1981). *Structural Aspects of Biomolecules*, edited by R. Srinivasan & V. Pattabhi, pp. 82–122. Delhi: Macmillan India Ltd.

Kendell, M. G., Ord, J. K. & Stuart, A. (1991). *Advanced Theory of Statistics*, 5th ed., Vol. 2. London: Edward Arnold.

Kleywegt, G. J. & Jones, T. A. (1995). *Structure*, **3**, 535–540.

Kleywegt, G. J. & Jones, T. A. (1997). *Methods Enzymol.* **277**, 208–230.

McRee, D. E. (1992). *J. Mol. Graph.* **10**, 44–46.

Schwarzenbach, D., Abrahams, S. C., Flack, H. D., Gonschorek, W., Hahn, Th., Huml, K., Marsh, R. E., Prince, E., Robertson, B. E., Rollett, J. S. & Wilson, A. J. C. (1989). *Acta Cryst.* A**45**, 63–75.

Sheldrick, G. M. (1997). *SHELXL97. A Program for the Refinement of Crystal Structures from Diffraction Data.* Institüt für Anorganische Chemie, Göttingen, Germany.

Stout, C. D. (1993). *J. Biol. Chem.* **268**, 25920–25927.

Stout, C. D. & Jensen, L. H. (1989). *X-ray Structure Determination: A Practical Guide.* New York: John Wiley & Sons.

Stout, C. D., Stura, E. A. & McRee, D. E. (1998). *J. Mol. Biol.* **278**, 629–639.

Ten Eyck, L. F. (1996). *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey. Warrington: Daresbury Laboratory.

Ten Eyck, L. F. (2000). In *Crystallographic Computing 7*, edited by P. E. Bourne & K. Watenpaugh. Oxford University Press.

Trefethen, L. N. & Bau, D. (1997). *Numerical Linear Algebra.* New York: SIAM.

Tronrud, D. E. (1992). *Acta Cryst.* A**48**, 912–916.

Watkin, D. (1988). *Crystallographic Computing 4*, edited by N. W. Isaacs & M. R. Taylor, pp. 111–125. Oxford University Press.