Conjugate Direction Minimization: An Improved Method for the Refinement of Macromolecules.

D. E. Tronrud
Howard Hughes Medical Institute
and
Institute of Molecular Biology
University of Oregon
Eugene, OR 97403

Submitted August 21, 1991
Revised April 23, 1992

Abstract

A novel method of function minimization which combines the power of the diagonal approximation to the normal matrix with conjugate directions is described. This method approaches closer to the local minimum than the methods which are commonly used in macromolecular refinement. The weaknesses of the current methods are analyzed to explain the advantage of the conjugate direction method.

1 Introduction

A persistent problem with macromolecular refinement is that the R-factors of the final models are higher than those obtained in small molecule structures. Over the last ten years, even though the same basic type of model is used to represent the molecule, average R-factors have decreased from about 20 percent to 16 percent. The difference is the sophistication of the refinement methods used. It seems likely that further improvement could be achieved if more powerful techniques were available.

In pursuit of this goal a modification of the conjugate gradient method of function minimization (Fletcher, 1964) has been developed which uses more information about the function being minimized than any method currently used. In particular, it uses explicit knowledge of the diagonal elements of the normal matrix together with implicit knowledge of the off-diagonal terms learned from the history of the refinement to determine better directions in parameter space to search. This method can determine a set of parameters which agree better with the observations in a shorter amount of computer time than the methods described previously.

2 Overview of Function Minimization

The theoretical underpinnings of all refinement methods used in the latter stages of high resolution refinement are the same. The analysis begins by making a Taylor series expansion of the function being minimized about the current guess for the values of the parameters of the model ($\mathbf{x}_0$). The Taylor series expansion is

$$f(\mathbf{x}) = f(\mathbf{x}_0) \; + \; \mathbf{g}^t(\mathbf{x}_0)\,\mathbf{d} \; + \; \frac{1}{2}\mathbf{d}^t\mathbf{N}(\mathbf{x}_0)\,\mathbf{d} \; + \; \cdots \; , \tag{1}$$

where $\mathbf{g}(\mathbf{x})$ is the gradient of the function, $\mathbf{N}(\mathbf{x})$ is the second derivative or normal matrix, and $\mathbf{d}$ is the shift vector which takes $\mathbf{x}_0$ to $\mathbf{x}$. The higher order terms are always assumed equal to zero.

To find the value of $\mathbf{x}$ where $f(\mathbf{x})$ is minimal we take the derivative of Equation (1) with respect to $\mathbf{x}$ and solve for $\mathbf{d}$ when $\mathbf{g}(\mathbf{x})$ is $\mathbf{0}$. The result is

$$\mathbf{d} = -\mathbf{N}^{-1}(\mathbf{x}_0)\,\mathbf{g}(\mathbf{x}_0), \text{ and} \tag{2}$$

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{d}. \tag{3}$$

$\mathbf{x}$ defines the minimum in all cases where the higher order terms are, in fact, zero, and when $\mathbf{N}$ is positive definite, which is always the case in this application.

3 The Macromolecular Problem

Equations (2) and (3) require that the normal matrix be calculated and inverted. This matrix is of size $n \times n$ where n is the number of parameters in the model. For many macromolecular structures this number is of the order of 10,000. The calculation and inversion of this matrix is still impractical. To refine large models a method must be chosen which avoids these steps.

The authors of the several refinement packages in common use have chosen different ways to avoid this problem. The program X-PLOR (Brunger, 1987) uses the method of simulated annealing in the early stages of refinement. TNT (Tronrud, 1987) and X-PLOR (in later stages) both use the conjugate gradient method. SFRF (Agarwal, 1978) and EREF (Jack, 1978) both use a diagonal approximation to the normal matrix, while CORELS (Sussman, 1977) and PROLSQ (Hendrickson, 1980) use a sparse matrix approximation.

4 Review of the Conjugate Gradient Method

Without complete knowledge of the normal matrix, minimization of a quadratic function would require repeated cycles. In each cycle a shift vector is chosen ($\mathbf{d}_k$ for cycle k) and the minimum along that direction is found

with a line search. The minimization of a function along a direction reduces the problem to one parameter, named α. The new values for the full set of parameters (for cycle $k+1$) are

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_{k+1}\mathbf{d}_{k+1}. \tag{4}$$

The value of α_{k+1} is set to that value which minimizes $f(\mathbf{x}_k + \alpha_{k+1}\mathbf{d}_{k+1})$. It defines the minimum along the shift vector $\mathbf{d}_{k+1}$.

The particular set of directions searched determines the rate of convergence. For example, if one chooses to search along the axes of parameter space, first varying the x parameter of the first atom, then the y parameter and so on, the minimum can only be found after a number of cycles many times n.

Many cycles are required when one parameter at a time is varied because the parameters (and therefore the shift directions) are interdependent; thus, in subsequent cycles previously searched directions must be searched again. The number of cycles could be reduced if a series of directions could be identified which were independent. This independence is described mathematically as the direction vectors being conjugate to the normal matrix (Luenberger, 1973), which is defined explicitly as $\mathbf{d}_l^t\mathbf{N}\mathbf{d}_m = 0$ when $l \neq m$.

A conjugate direction method is one in which a series of directions are

devised which are conjugate with respect to the normal matrix but do not require the normal matrix in order for them to be determined. In the particular conjugate direction method called conjugate gradient, the direction vector for cycle $k+1$ is determined from

$$\mathbf{d}_{k+1} = -\mathbf{g}_k + \beta_{k+1}\mathbf{d}_k, \text{ and} \tag{5}$$

$$\beta_{k+1} = \frac{\mathbf{g}_k^t \mathbf{g}_k}{\mathbf{g}_{k-1}^t \mathbf{g}_{k-1}}, \tag{6}$$

where β_{k+1} is chosen to ensure that $\mathbf{d}_{k+1}$ is conjugate to all previous directions. $\mathbf{d}_0$ and β_1 are defined to be $\mathbf{0}$ and 0, respectively, which results in the first cycle being a steepest descent cycle ($\mathbf{d}_1 = -\mathbf{g}_0$).

5 Limitations of Conjugate Gradient

The fundamental limitation of the conjugate gradient method is that it requires, in general, n cycles to reach the minimum. We need a procedure which will perform most of the function minimization in the first few cycles.

The eigenvalues of the normal matrix (Leunberger, 1973) provide information about how a method will refine parameters in the early cycles. The normal matrix describes the shape of the minimum of the function, and its eigenvalues define how oblong the neighborhood of the minimum is. Because the normal matrices for the functions usually minimized in macromolecular

refinement are nearly diagonal there is a close correspondence between the eigenvectors and the parameters of the model. For a perfectly diagonal normal matrix, the eigenvectors are the axes of parameter space and the diagonal elements, or curvatures, are the eigenvalues.

The method of steepest descent works best when all the eigenvalues or diagonal elements are equal. If they are not equal the parameters with the largest curvatures dominate. The conjugate gradient method must infer the differences in curvature from the history of the search but this takes more cycles than we give the method in practise.

This problem is especially serious when positional parameters are compared to thermal parameters. The curvatures for positional parameters are much larger than those for thermal parameters; therefore, refinement of thermal parameters is blocked by the influence of the positional parameters. This effect is usually avoided by refining thermal parameters with the positional parameters held constant and vise versa.

A more intractable problem arises because the curvatures associated with numerically large thermal parameters are much smaller than those of smaller thermal parameters. In all models produced by refinement using the conjugate gradient method and methods which simplistically incorporate curva-

tures, the large thermal factors are poorly refined and probably should have even larger values than those obtained during the refinement process. In addition atom types with many electrons, such as sulfur and iron, have large curvatures. The thermal factor shifts of these atoms will be overestimated, resulting in an oscillation about the correct value.

6 Improvements in the Conjugate Gradient Method

The conjugate gradient method uses the steepest descent method to produce its first shift direction, or "seed" direction. The rate of convergence of early cycles can be improved if a seed that incorporates as much information as practical about the function is used. We would like a direction which will include compensation for the differences in the eigenvalues of the normal matrix. Because in X-ray crystallography the diagonal terms of the normal matrix dominate, a diagonal approximation to the normal matrix provides a powerful and quick alternative to the steepest descent method of generating shift directions. In this procedure the search direction is calculated by

$$\mathbf{d}_{k+1} = -\mathbf{N}_{d,k}^{-1}\,\mathbf{g}_k, \tag{7}$$

where $\mathbf{N}_{d,k}$ is the diagonal approximation to the normal matrix for the parameters of cycle k. For the fastest rate of convergence, this shift vector should be used as a seed for conjugate direction searches. It is not clear, however, how one should calculate β of Equation (6).

The refinement problems that we address are the wide range of magnitude of the eigenvalues of the normal matrix and the existence of off-diagonal terms. If we could choose a different set of parameters, for which the normal matrix was simpler, the rate of convergence would improve. Ideally one would choose a system of parameters such that all the eigenvalues were equal and all the off-diagonal elements were zero; then one cycle of steepest descent minimization would suffice.

Let us assume that we have determined a matrix ($\mathbf{M}$) which will transform the usual crystallographic parameters into such a set of parameters ($\mathbf{x}'$). The transformations between the familiar parameters and the new ones will be

$$\begin{array}{cc} \mathbf{x}' = \mathbf{M}\mathbf{x} & \mathbf{x} = \mathbf{M}^{-1}\mathbf{x}' \\ \mathbf{g}' = \mathbf{M}^{-1t}\mathbf{g} & \mathbf{g} = \mathbf{M}^{t}\mathbf{g}' \\ \mathbf{N}' = \mathbf{M}^{-1t}\mathbf{N}\mathbf{M}^{-1} & \mathbf{N} = \mathbf{M}^{t}\mathbf{N}'\mathbf{M} \end{array} \tag{8}$$

We can perform Fletcher–Reeves conjugate gradient minimization on the function using this new parameter space. The equations will be as before, Equations (4) thru (6), but with primes added:

$$\mathbf{x}'_{k+1} = \mathbf{x}'_k + \alpha_{k+1}\mathbf{d}'_{k+1}, \tag{9}$$

$$\mathbf{d}'_{k+1} = -\mathbf{g}'_k + \beta'_{k+1}\mathbf{d}'_k, \text{ and} \tag{10}$$

$$\beta'_{k+1} = \frac{\mathbf{g}'^t_k \mathbf{g}'_k}{\mathbf{g}'^t_{k-1}\mathbf{g}'_{k-1}}. \tag{11}$$

Instead of working with the $\mathbf{x}'$ parameters we can substitute back to the original $\mathbf{x}$ parameters. The resulting equations are

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_{k+1}\mathbf{M}^{-1}\mathbf{d}'_{k+1}, \tag{12}$$

$$\mathbf{M}^{-1}\mathbf{d}'_{k+1} = -\mathbf{M}^{-1}\mathbf{M}^{-1t}\mathbf{g}'_k + \beta'_{k+1}\mathbf{M}^{-1}\mathbf{d}'_k, \text{ and} \tag{13}$$

$$\beta'_{k+1} = \frac{\mathbf{g}^t_k \mathbf{M}^{-1}\mathbf{M}^{-1t}\mathbf{g}_k}{\mathbf{g}^t_{k-1}\mathbf{M}^{-1}\mathbf{M}^{-1t}\mathbf{g}_{k-1}}. \tag{14}$$

In these equations the shift vectors, $\mathbf{d}'$, are all premultiplied by $\mathbf{M}^{-1}$. It would be simpler to eliminate this complication by simply defining $\mathbf{d} = \mathbf{M}^{-1}\mathbf{d}'$. The final equations for conjugate direction refinement, derived from recombined parameters, but operating on the "native" parameters are

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_{k+1}\mathbf{d}_{k+1}, \tag{15}$$

$$\mathbf{d}_{k+1} = -\mathbf{M}^{-1}\mathbf{M}^{-1t}\mathbf{g}'_k + \beta'_{k+1}\mathbf{d}_k, \text{ and} \tag{16}$$

$$\beta'_{k+1} = \frac{\mathbf{g}^t_k \mathbf{M}^{-1}\mathbf{M}^{-1t}\mathbf{g}_k}{\mathbf{g}^t_{k-1}\mathbf{M}^{-1}\mathbf{M}^{-1t}\mathbf{g}_{k-1}}. \tag{17}$$

At this point the matrix $\mathbf{M}$ is undefined. The optimal choice for $\mathbf{M}$ would require that $\mathbf{M}^{-1t}\mathbf{N}\mathbf{M}^{-1}$, the normal matrix for the new parameters, be equal to the identity matrix. To calculate the optimal $\mathbf{M}$ we need both

the normal matrix and its inverse; thus, we made no gains in computational efficiency over the full matrix method. However, if we recognize that in crystallography $\mathbf{N}$ is almost diagonal we can set $\mathbf{M} = \mathbf{N}_d^{1/2}$. Then $\mathbf{M}^{-1}\mathbf{M}^{-1t}$ in the Equations (15), (16) and (17) will be replaced by $\mathbf{N}_d^{-1}$. Making this substitution we obtain

$$\mathbf{d}_{k+1} = -\mathbf{N}_d^{-1}\mathbf{g}_k + \beta'_{k+1}\mathbf{d}_k, \text{ and} \tag{18}$$

$$\beta'_{k+1} = \frac{\mathbf{g}_k^t\mathbf{N}_d^{-1}\mathbf{g}_k}{\mathbf{g}_{k-1}^t\mathbf{N}_d^{-1}\mathbf{g}_{k-1}}. \tag{19}$$

The seed direction ($\mathbf{d}_1$ when $\beta'_{k+1} = 0$ and $\mathbf{d}_0 = \mathbf{0}$) is now the shift calculated from the diagonal approximation to the normal matrix, as we desired. In addition, however, we have an equation for β.

In summary, we have a minimization method where the diagonal terms of the normal matrix are explicitly included and the off-diagonal elements are dealt with via a set of conjugate directions.

Agarwal (1978) suggested a similar method; however, his equation for β was incorrect. In the present nomenclature, his proposal for β was

$$\beta = \frac{\mathbf{d}_k^t\mathbf{d}_k}{\mathbf{g}_k'^t\mathbf{g}'_k}. \tag{20}$$

In conjugate gradient refinement β is equal to ratio of the length of the gradient at point k divided by that length at point $k-1$. Because k should

be closer to the minimum than point $k-1$, β should be less than unity. An estimate of Agarwal's β can be achieved by examining his β for cycle 2, which is

$$\beta_2 = \frac{\mathbf{g}_0'^t \mathbf{g}_0'}{\mathbf{g}_1'^t \mathbf{g}_1'}. \tag{21}$$

As before, the parameters after cycle 1 should be closer to the minimum than the starting parameters, resulting in $\beta_2 > 1$. This value results in the undesirable outcome that the previous cycle's direction is considered more important that the direction calculated from the current parameters. This now explains why Agarwal found it necessary to place an empirical upper limit of 0.4 on β. The value of β calculated with Equation (19) typically falls between 0.5 and 0.9. The empirical value of 0.4 falls closer to the typical value than either setting β to zero (and using the method of Equation (7)) or using the equation of Agarwal(1978).

7 Some Comparisons

Parallel refinement runs were performed to compare the convergence properties of the four types of function minimization described in the text. The methods are steepest descent (SD), conjugate gradient (CG), diagonal approximation to the normal matrix (also called "gradient over curvature"

or GC), and the new conjugate direction method (CD). The test structure was the thermolysin-phosphoramidon inhibitor complex (Weaver, 1977) using data collected between 20 and 2.3Å resolution (a total of 13,730 reflections). The starting model was the "native" coordinates of thermolysin (Holmes, 1982) with a crude phosphoramidon model appended and displaced solvent atoms removed. The starting model, which contained a total of 2637 atoms, was known to contain a number of errors. The initial R-factor was 21.7 percent.

Refinement was carried out using the TNT refinement package (Tronrud, 1987), modified to include the new conjugate direction method as an option. (The crystallographic portions of the diagonal elements of the normal matrix were calculated by the method of Agarwal, 1978). All four methods were run with the thermal parameters held constant because the refinement methods which do not use curvatures cannot simultaneously vary both positional and thermal parameters. Separate tests were made to compare GC and CD refinement in which both positional and thermal parameters were allowed to change simultaneously. The only differences between these test runs were the set of parameters varied and the method used. All other aspects, such as weights, were identical. The results of these tests are displayed in Figure 1.

The methods which use curvatures (GC and CD) are superior to the methods which do not (SD and CG). After 15 cycles of refinement the conjugate gradient run is similar to the "gradient over curvature" method because the diagonal elements of the normal matrix for the positional parameters are all approximately equal to each other and the conjugate gradient method can accommodate their differences relatively quickly. However this is not the case for all types of parameters; shifts in B-factors that are numerically small and numerically large have different effects on the value of the function.

The comparison between the run of gradient over curvature refinement and the run of conjugate direction refinement in which both XYZ's and B's were varied shows the clear superiority of the new method. The R-factor of the model produced by 20 cycles of conjugate direction refinement was 13.2 percent and still dropping, with good geometry (bond length rms error 0.027Å and bond angle rms error 3.5°).

The reason the new method produces a lower value for $f(\mathbf{x})$ is not because the other methods are stuck in higher local minima. For either conjugate gradient or conjugate direction to work they must be close enough to a minimum that the higher order terms of Taylor's series expansion are insignificant. Each method will proceed to the minimum of the expansion,

which is the local minimum. That minimum is the same for the two methods because the function itself is unchanged, only the set of directions to be searched has been altered by the new method. Eventually the conjugate gradient or steepest descent method will descend as low as the conjugate direction method; it will simply take many more cycles to get there.

Figure (1) shows that even after 20 cycles the new method has not reached a minimum either. Methods with even greater power of convergence should be able to produce parameter sets where $f(\mathbf{x})$ is even lower, using affordable amounts of computer time.

8 Acknowledgements

This work was supported in part by grants to B. W. Matthews from the NIH (GM20066) and the Lucille P. Markey Charitable Trust.

References

[1] Agarwal, R.C. Acta Cryst. A43, 791-809 (1978)

[2] Brunger, A.T., Kuriyan, K., Karplus, M., Science 235, 458-460, (1987)

[3] Fletcher, R., Reeves, C., Computer Journal 7, 149-154 (1964)

[4] Hendrickson, W.A., Konnert, J.H., In *Computing in Crystallography*, edited by Diamond, R. Ramaseshan, S., Venkatesan, K., 13.01-13.25 (1980), Bangalore: Indian Academy of Sciences.

[5] Holmes, M.A., Matthews, B.W., J. Mol. Biol. 160, 623-639 (1982)

[6] Jack, A., Levitt, M., Acta Cryst. A34, 931-935 (1978)

[7] Luenberger, D.G. *Introduction to Linear and Nonlinear Programming* (1973). Reading, MA. Addison-Wesley

[8] Sussman, J.L., Holbrook, S.R., Church, G.M., Kim, S., Acta Cryst. A33, 800-804 (1977)

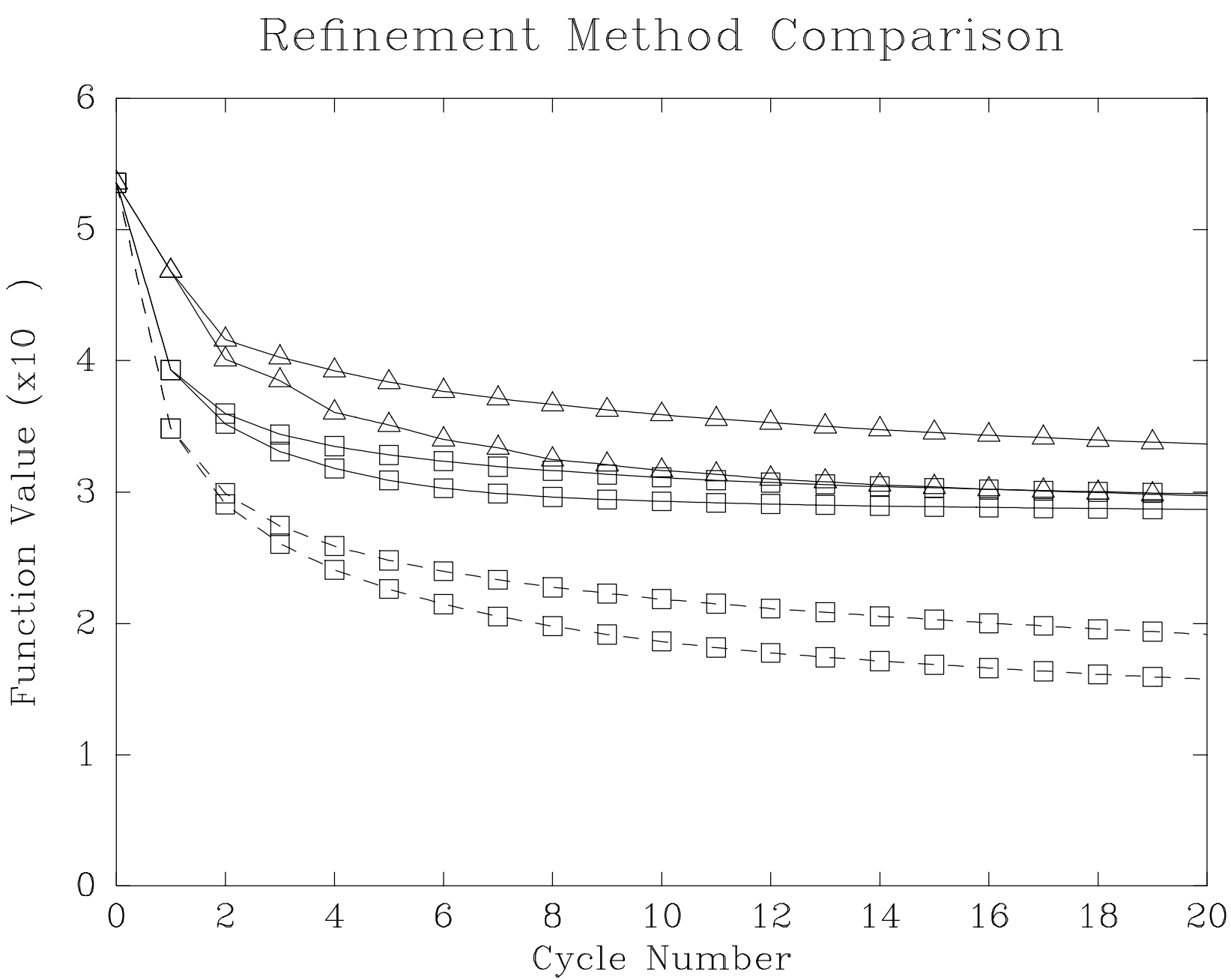
[9] Tronrud, D.E., Ten Eyck, L.F., Matthews, B.W., Acta Cryst. A43, 489-501 (1987)

[10] Weaver, L.H., Kester, W.R., Matthews, B.W., J. Mol. Biol. 114, 119-132 (1977)

9 Figure Captions

Figure 1: This graph shows the drop in the value of the function that is minimized in refinement over 20 cycles of refinement. Four different methods of minimization are compared. In some test runs (solid lines) only the positional parameters were varied while in the rest (broken lines) both the positional and thermal parameters were varied. The function is $\sum(F_o(hkl) - F_c(hkl))^2$ after the F_o's and F_c's have been scaled to each other, plus the sum of the geometry deviation terms. The methods represented with triangles required 18.5 minutes of CPU time per cycle on a VAX 3600 computer. The methods represented with squares required the additional calculation of curvatures and took 22 minutes per cycle.

This plot demonstrates that the conjugate direction method produces a lower function value for a given number of cycles of refinement.


Refinement Method Comparison
Function Value (x10)
Cycle Number
0
1
2
3
4
5
6
0
2
4
6
8
10
12
14
16
18
20