to be 0·112 Å shorter than the sum of the Pauling covalent radii for carbon and sulphur. This bond must therefore possess considerable double-bond character.

Cucka (1963) also puts forward an argument for the existence of a similar ring system in that his calculations support the evidence for hydrogen atoms attached to C(2), C(3), C(4) and N(2). Thus it appears, from these two X-ray investigations and Hedgley's observations, that there is now sufficient evidence to confirm the existence of the tautomeric form (I) of pyridaz-3-thione not only in solution but in the solid state as well.

We should like to acknowledge the very helpful suggestions that came from Professor Dame Kathleen Lonsdale and Dr Judith Milledge concerning the attempts made to analyse the thermal vibrations of this structure. Although the analysis did yield qualitative indications of the molecular vibrations to be found in structures of this type, they were not sufficiently accurate to merit publication. The calculations were carried out with Dame Kathleen's kind permission on the Pegasus computer in her department, and in this connection we are grateful to Dr C. J. Brown for the use of his programs.

We also acknowledge the use of programs written by Drs J. S. Rollett, Jean Dollimore and J. W. Jeffery, and one of us (M.B.H.) is grateful to the Commonwealth Scholarships Commission, for the tenure of an award while in this country.

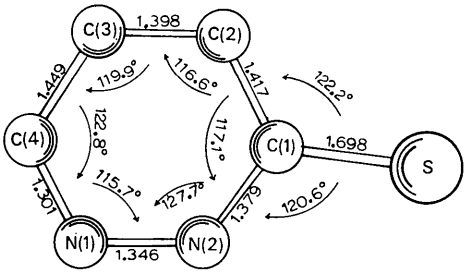

Fig. 4. A drawing of the pyridaz-3-thione molecule (minus H atoms) showing bond lengths and bond angles.

References

BERGHUIS, J., HAANAPPEL, IJ. M., POTTERS, M., LOOPSTRA, B. O., MACGILLAVRY, C. H. & VEENENDAAL, A. L. (1955). *Acta Cryst.* **8**, 478.
CUCKA, P. (1963). *Acta Cryst.* **16**, 318.
CUCKA, P. & SMALL, R. W. H. (1954). *Acta Cryst.* **7**, 199.
HEDGLEY, E. J. (1956). Ph. D. Thesis, Birmingham Univ., England.
HEDGLEY, E. J., HEIKEL, T. A., KNIGHT, B. C. & RIMMINGTON, C. (1959). *Lancet*, p. 1183.
International Tables for X-ray Crystallography (1962). Vol. III. Birmingham: Kynoch Press.
PENFOLD, B. (1953). *Acta Cryst.* **6**, 707.
PHILLIPS, D. C. (1956). *Acta Cryst.* **9**, 819.
WILSON, A. J. C. (1942). *Nature, Lond.* **150**, 152.

Acta Cryst. (1966). **21**, 253

A Mathematical Model-Building Procedure for Proteins

BY ROBERT DIAMOND

Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge, England

(*Received* 3 *November* 1965)

The procedure is intended primarily as an intermediate step between the interpretation of an electron density map of medium resolution and the refinement of the structure. The procedure builds a representation of a polypeptide using images of amino acids as determined in small structures, rendered suitably flexible by rotations about single bonds (and other lines if required) and uses the method of least squares to fold the resulting chain and side chains to approach the guide points (derived from the electron density map or otherwise) as closely as possible. Thus an idealized structure can be derived from a selection of rough coordinates and a knowledge of the sequence. The procedure is also capable (with limitations) of bridging uncertain regions.

Some novel mathematical techniques of general interest are described and employed. These include reversion and a sliding filter as means of combating non-linearity. The sliding filter is a means of suppressing large shifts by excluding from the least-squares process those eigenvectors of the normal matrix which have small eigenvalues. This is done in a manner depending on the residual.

A means of achieving accurate rotations in three dimensions without setting up a matrix is also given.

1. Introduction

1.1. *General*

It is a difficult matter, even at the highest resolution, to determine atomic coordinates in a protein with an accuracy better than 0·25 Å. It follows that if a flexible chain with idealized links can be threaded through such coordinates (here called *guide coordinates*) with comparable accuracy then such a chain is reasonable in its stereochemistry and equally consistent with the

X-ray observations. A method of refining the course of such a flexible chain against the X-ray observations has already been described (Diamond, 1965). This refinement procedure leaves bond lengths and (in most cases) inter-bond angles unaltered, and it is therefore desirable that the chain being refined should at the outset possess reasonable values for these quantities free from the substantial random errors which are inevitable if the coordinates of atoms are determined independently. The procedure described in this paper is designed to provide a chain structure as a starting point for the chain refinement process.

The advantages offered by this approach may be summarized as follows:

(i) The resulting stereochemistry is necessarily consistent with studies of small structures.
(ii) The parameters involved are the conformational parameters which are the centre of attention in contemporary studies of protein conformation. They number approximately $\frac{1}{2}N$ where N is the number of non-hydrogen atoms, compared with $3N$ for x, y and z.
(iii) The conservation of the integrity of the chain ensures that the final coordinates of each atom are influenced by many atoms acting cooperatively, thus providing, in effect, a better 'signal to noise ratio' than would be the case if all the x, y, z were independent.
(iv) By the use of substantial lengths of chain and matrix diagonalization techniques, those combinations of rotations which for the smallest shifts provide the biggest improvement may be determined and employed.

Against all this may be raised the one objection that by imposing predetermined characteristics on the structure one may be enforcing an error if the structure genuinely differs from expectation in an unalterable manner, but it seems unlikely that the X-ray evidence in any instance will be strong enough to establish a variation of a type which cannot be accommodated by the provision of appropriate flexibility. In such an event one is always free to release the atoms concerned from the chain constraints.

The procedures described here are regarded, therefore, as primarily a crystallographic tool, rather than a trial and error search process such as that of Némethy & Scheraga (1965), although they may already be used to some extent in that way, and have enormous potential for development in that direction.

In its present form the method can handle any chain for which:

(i) Side chains (unless rigid) join the main chain at a point not a ring, as in nucleic acids.
(ii) Side chains which are themselves forked have not more than one branch flexible.
(iii) Side chains, if flexible at all, are flexible in the innermost bond ($C_\alpha - C_\beta$, χ_1).
(iv) If a side chain is flexible, at least one of the main chain bonds adjacent to the point of union (C_α) must also be flexible. (Usually both such bonds are flexible, but if only one is, then it must be the same one throughout the chain, and the direction of progress will be dictated by this. Rigid side chains (*e.g.* proline) do not introduce such restrictions.)
(v) The link, *i.e.* that part of the main chain occurring between two side chains, must not itself be altered in shape by any of the rotations permitted to it. (See §2.2.4, however).
(vi) The link must contain 4 or 5 atoms.

Proteins naturally meet all these requirements and there may well be other polymers for which the procedure is useful.

The first four of these requirements arise because the topology of the molecule is implied entirely by the ordering of the entries for atoms, main chain parameters, and side chain parameters in the listing which is assembled. If these rules were departed from, the logic of the program would need reorganizing. The source of the fifth requirement will be clear in §2·1, and the sixth is simply a storage requirement.

Hitherto, hydrogen atoms have not been included.

1.2. *Outline procedure*

The program* first reads *standard groups* in the form of named sequences defining the various types of side chain and two copies of the main chain link (described more fully in §2.1). These sequences are introduced by cards carrying 3-letter names such as ARG for arginine, and contain coordinates of the various atoms each with an identifying name (such as CB for the β carbon atom) and with parameter cards inserted between pairs of atoms where it is intended that the line joining them should be a line about which rotation may take place. Inter-bond angles may be made variable by using a dummy atom to define an axis of rotation normal to the plane containing the bonds concerned. The orientation of these groups is such that the two copies of the main chain link, (called the *link* and *precursor*), together with any one of the side chain groupings, form a fragment of a polypeptide chain consisting of the side chain and main chain up to and including the neighbouring C_α atom on each side. The link and precursor each include a C_α at each end, the one which is common to both being set at the origin. Each side chain grouping is also referred to that position as origin, but does not normally contain an entry corresponding to the C_α position. The spatial relationship between the link and the precursor is used to define the relationship produced by the building process, and

* The program is written in FORTRAN II and requires a 32K storage machine of 36-bit words. It uses the IBM internal code as given on page 97 of IBM publication C28-6054-2 *Fortran Manual* for purposes of character recognition and in its present form is therefore specific to IBM machines.

which, if perpetuated without alteration, yields a helix, normally an α-helix.

Having read the standard groups the program then reads the amino acid sequence which it is required to build, in which each residue is introduced by a sequence card bearing the name of one of the standard groups. Each sequence card may be followed by as many or as few coordinate cards as desired, these being the rough coordinates used to guide the process. Each of these is identified by a name which must match the name of the corresponding atom in the link group or in the standard group nominated by the previous sequence card. Each sequence card may also be followed by parameter cards which, if present, cause the corresponding angles in the computed structure to have their *initial* values altered by the amounts given on the cards. The *final* values adopted by these angles, however, are determined by the guide coordinates, so that the initial and final values of such an angle are only equal (in general) if the selection of guide points used leaves the angle concerned indeterminate (*e.g.* in a side chain if only main chain guide coordinates are given). The standard groups must be provided in Cartesian coordinates in Å. The guide coordinates must be given in crystallographic fractional coordinates and the output listing is also in this form. All internal working is in Cartesian coordinates in Å.

The basic sequence of events is best described by the flow diagram of Fig. 1. Note that if sequence cards alone are given the building process perpetuates the relationship between the link and precursor, thereby building (usually) α-helix until some guide points are

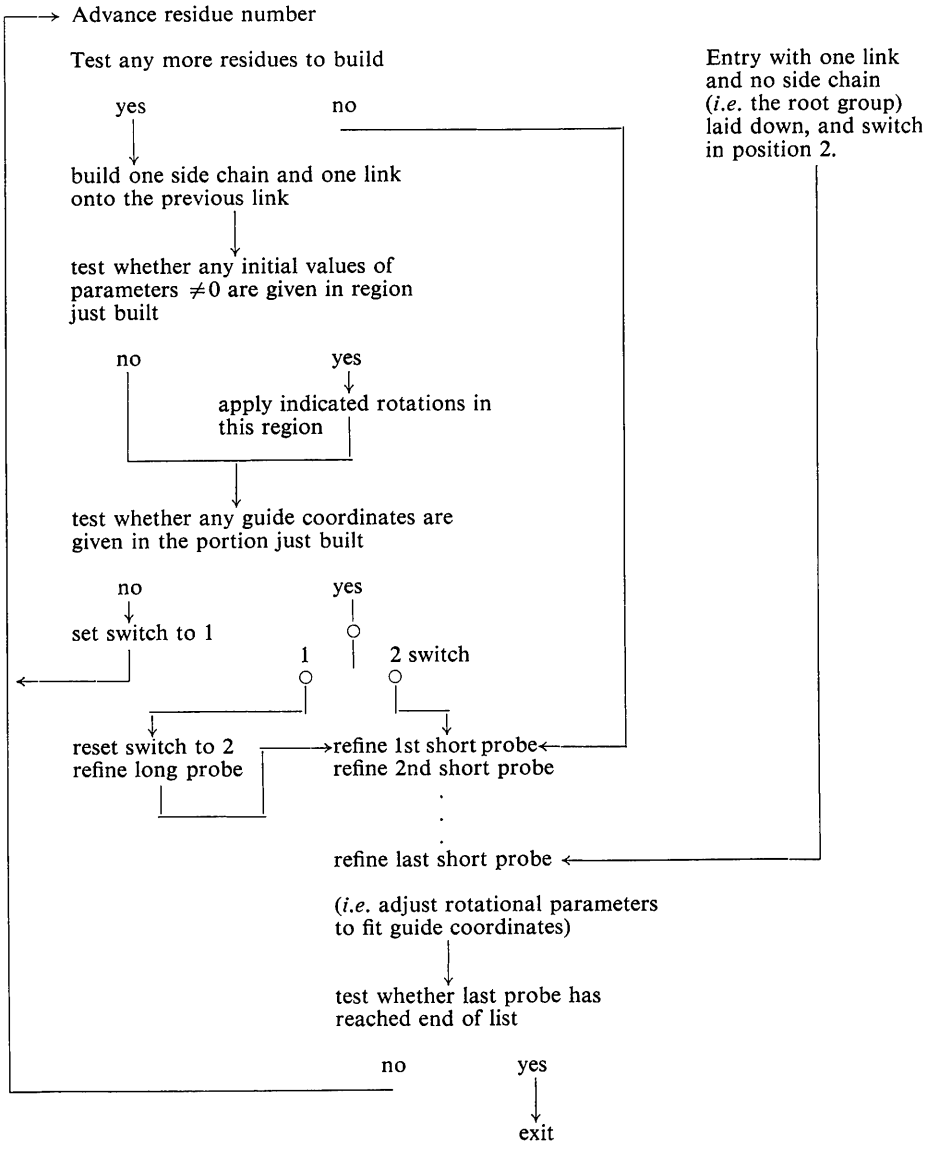

Fig. 1. Outline flow diagram.

encountered. When these are encountered a long probe refinement is done, *i.e.* the entire helix is adjusted at its root (*i.e.* fixed end) to bring its tip into coincidence with the guide points. This is done in a way which is formally analogous to the elastic bending of a beam clamped at one end, wherein the strain is distributed so as to minimize the elastic strain energy consistant with providing the required movement of the tip.

The program then goes on to refine several short probes (at least one, not more than four) which are lengths of most recently built chain, whose conformations are still under review. Each probe has its control quantities set by data cards, which specify, *inter alia*, its length in residues (except for the long probe whose length is governed by the length of helix which has been built without guidance). The first short probe is likely to contain the newly built link and side chain and half of the previous link. This includes φ and ψ (in the new proposed notation of Edsall, Flory, Kendrew, Liquori, Némethy, Ramachandran & Scheraga, 1966) of the new residue and therefore possesses sufficient freedom to allow the new residue to take up any required conformation, assuming that the previously built residue has been correctly placed. This first short probe is the one that does nearly all the work; in building a non-helical region, for example, the initial conformation presented to this probe is normally α-helical, and this probe may be required to produce rotations of the order of hundreds of degrees to negotiate a corner. All such angular rotations are determined with the use of linear least-squares methods, and special techniques have been developed which ensure rapid convergence from far outside the region of linear behaviour; these are described below. Despite this, it occasionally happens that the process finds a local minimum away from the correct conformation, and in such cases parameter cards may be used to modify the initial conformation before refinement so as to put it in the right convergent region. In practice, this only occurs in the main chain with conformations accessible only to glycine in the right hand half of the steric map of Ramachandran, Ramakrishnan & Sasisekharan (1963), *i.e.* if φ differs from the α-helical value by $\sim +100°$, and such conformations are fortunately uncommon. False minima associated with side chains have been found to occur only with side chains having large rigid groups when it is possible for the newly built link and the side chain group each to prevent the other from reaching the required position, but this too can be very easily avoided, either by using parameter cards as above or by using only a few guide points in the rigid group so that any conflict between the new link and the side chain is unbalanced and therefore quickly resolved.

The second and subsequent probes are normally longer than the first, and they permit the revision of conformations following further building. Thus, if a probe length of 9 residues is used (the maximum with present compilations) each residue gets refined nine times by this probe before it is left alone, and on the last occasion it has a fixed (finalized) residue on one side of it and eight already well positioned residues still under revision on the other side of it, so that each residue is eventually well bedded in with quite long range considerations taken into account.

The program terminates only when the root end of the last (longest) probe reaches the end of the molecule.

2. Mathematical aspects

2.1. *The building process*

Unlike most other model-building procedures which have been described in the literature, *e.g.* Némethy & Scheraga (1965), this procedure does not keep a record, in the form of a rotation matrix, of the orientation of any particular link of the chain, and the development of the initial coordinates for a link and side chain (*i.e.* prior to refinement) is a one-step matrix multiplication in which the prefactor is a 3×5 matrix containing the coordinates of the five atoms in the precursor, *i.e.* the last previous link in the model, and the postfactor is a constant matrix, and of these there is one for each side chain type and one for the link.

Fig. 2 shows schematically the arrangement and ordering of atoms in the link and precursor as input in the standard groups. We may define certain matrices, as follows, using capital letters for the standard group orientation and lower case for the corresponding quantities in the computed model.

$$\mathbf{L} = \begin{pmatrix} x_1 & x_2 & \cdots\cdots & x_5 \\ y_1 & y_2 & \cdots\cdots & y_5 \\ z_1 & z_2 & \cdots\cdots & z_5 \end{pmatrix}$$

defines the link, in which the last column vanishes. Similarly **P** defines the precursor, in which the first column vanishes, and for each side chain type we define a matrix **S** containing standard group coordinates in three rows and as many columns as the side chain

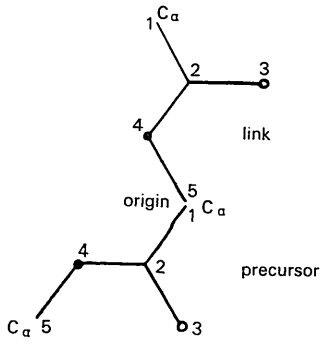

Fig. 2. Schematic diagram of the link and precursor as used for building proteins. This diagram is for identification only and does not imply that the link and precursor are normally coplanar. The arrangement shown implies that the input and final listings are to begin at the N-terminal end and that the C-terminal end is to be built first. This procedure could be reversed by interchanging the link and precursor and reversing the numbering. Solid circles: nitrogen; open circles: oxygen.

requires. Then if $\mathbf{p}$ is a matrix obtained from the computed model by inserting the coordinates of the last link to be computed and subtracting the first column from every column (*i.e.* taking the tip C_α as origin), the coordinates of the next link and side chain to be built are, relative to the same origin,

$$\mathbf{l}=\mathbf{AL}$$
$$\mathbf{s}=\mathbf{AS}$$

where $\mathbf{A}$ is an orthogonal transformation satisfying

$$\mathbf{p}=\mathbf{AP}\,.$$

Postmultiplying this equation by $\tilde{\mathbf{P}}(\mathbf{P}\tilde{\mathbf{P}})^{-1}$, where the tilde denotes a transpose, gives

$$\mathbf{p}\tilde{\mathbf{P}}(\mathbf{P}\tilde{\mathbf{P}})^{-1}=\mathbf{AP}\tilde{\mathbf{P}}(\mathbf{P}\tilde{\mathbf{P}})^{-1}=\mathbf{A}\,.$$

Thus

$$\mathbf{l}=\mathbf{AL}$$

becomes

$$\mathbf{l}=\mathbf{p}\tilde{\mathbf{P}}(\mathbf{P}\tilde{\mathbf{P}})^{-1}\mathbf{L}$$

and

$$\mathbf{s}=\mathbf{p}\tilde{\mathbf{P}}(\mathbf{P}\tilde{\mathbf{P}})^{-1}\mathbf{S}\,.$$

Thus matrices of the form

$$\mathbf{V}=\tilde{\mathbf{P}}(\mathbf{P}\tilde{\mathbf{P}})^{-1}\mathbf{L} \quad \text{and} \quad \tilde{\mathbf{P}}(\mathbf{P}\tilde{\mathbf{P}})^{-1}\mathbf{S}$$

may be calculated once and for all at the outset, after which the store originally containing the standard groups as input is over-written with a library of matrices $\mathbf{V}$, and the building operation consists in selecting the appropriate $\mathbf{V}$ for link and side chain and evaluating

$$\mathbf{l}=\mathbf{pV} \quad \text{and} \quad \mathbf{s}=\mathbf{pV}\,.$$

There are two conditions which must be met for this process to work. The first is that $\mathbf{P}\tilde{\mathbf{P}}$ must be non-singular so that it may be inverted. This means that at least three of the columns of $\mathbf{P}$ must be linearly independent, whereas it is an unfortunate fact that the precursor is a two-dimensional figure so that the rank of $\mathbf{P}\tilde{\mathbf{P}}$ is only two. Now, it is in any case convenient to maintain five columns in $\mathbf{p}$ and $\mathbf{l}$, so that the $\mathbf{p}$ for one calculation may be obtained from the $\mathbf{l}$ of the previous one by subtracting the first column from every column, but this does mean that, as defined, the first column in $\mathbf{p}$ is effectively a spare, and the singularity of $\mathbf{P}\tilde{\mathbf{P}}$ is dealt with by loading the spare column with the vector product of the second column with the third. and the matrices $\mathbf{p}$ and $\mathbf{V}$ are calculated accordingly, In this form the process can cope with a two-dimensional link, but would fail with a linear one.

The second requirement is one of accuracy. Suppose that $\mathbf{p}_0$ is the initial precursor group in the computed model, on which all else will be built, then

$$\mathbf{l}_1=\mathbf{p}_0\mathbf{V}=\mathbf{p}_0\tilde{\mathbf{P}}(\mathbf{P}\tilde{\mathbf{P}})^{-1}\mathbf{L}$$

and

$$\mathbf{p}_1=\mathbf{l}_1'\,.$$

i.e. $\mathbf{l}_1$ with the first column subtracted from every column, and

$$\mathbf{l}_2=\mathbf{p}_1\tilde{\mathbf{P}}(\mathbf{P}\tilde{\mathbf{P}})^{-1}\mathbf{L}$$

which contains a translation and a rotational part equal to

$$\mathbf{l}_1\tilde{\mathbf{P}}(\mathbf{P}\tilde{\mathbf{P}})^{-1}\mathbf{L}=\mathbf{p}_0[\tilde{\mathbf{P}}(\mathbf{P}\tilde{\mathbf{P}})^{-1}\mathbf{L}]^2\,.$$

i.e. it is clear that the coordinates of the qth link contain the qth power of the matrix $\tilde{\mathbf{P}}(\mathbf{P}\tilde{\mathbf{P}})^{-1}\mathbf{L}$.

Now, if $\mathbf{L}'$ is $\mathbf{L}$ with the first column subtracted from every column, and if the standard link and precursor are perfect copies of each other, then

$$\mathbf{L}'=\mathbf{BP}$$

where $\mathbf{B}$ is an orthogonal transformation, and it is easy to show that $[\tilde{\mathbf{P}}(\mathbf{P}\tilde{\mathbf{P}})^{-1}\mathbf{L}]^q$ contains $\mathbf{B}^q$. It is clear that if the process is to be applied many times without introducing systematic drift of the link coordinates, $\mathbf{B}$ must be a highly perfect orthogonal matrix, *i.e.* $\mathbf{L}$ and $\mathbf{P}$ must be perfect images of each other. Now, the program is designed so that this stringent requirement need not be met by the standard groups as read in, but is imposed by revision of $\mathbf{L}$ as part of the initialization, and before the $\mathbf{V}$ are calculated, so that the $\mathbf{V}$ for the link contains the revised figures.

If $\mathbf{B}_0$ is defined by*

$$\mathbf{L}'=\mathbf{B}_0\mathbf{P}$$

where $\mathbf{L}'$ and $\mathbf{P}$ are *as input*, then, as before

$$\mathbf{B}_0=\mathbf{L}'\tilde{\mathbf{P}}(\mathbf{P}\tilde{\mathbf{P}})^{-1}$$

and $\mathbf{B}_0$ is then approximately orthogonal if $\mathbf{L}'$ is only approximately an image of $\mathbf{P}$. We then write

$$\mathbf{B}_{n+1}=\tfrac{1}{2}(\mathbf{B}_n+\tilde{\mathbf{B}}_n^{-1})$$

and iterate until $\mathbf{B}_n$ is perfect, and then discard $\mathbf{L}'$ in favour of $\mathbf{B}_n\mathbf{P}$. This is a second order process, and it may be shown that if

$$\mathbf{B}_n\tilde{\mathbf{B}}_n=\mathbf{I}+\mathbf{m}_n$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{m}_n$ is a matrix of residuals, then

$$\mathbf{m}_n=4\left(\frac{\mathbf{m}_0}{4}\right)^{2^n}.$$

Thus for a computation of q links the matrix $q\mathbf{m}_n$ should have elements of the order 10^{-10}. With $q=250$ and $\mathbf{m}_0$ containing elements $\sim 4\times 10^{-3}$ it appears that $n=2$ should suffice, and that $n=3$ would certainly be enough. The program allows n to go up to 4, and this has enabled a helix of 20 residues to be built with the coordinates in the first and last link agreeing to six significant figures (at which point the output was trun-

* This equation implies that it is possible to represent $\mathbf{L}'$ as a linear (not necessarily orthogonal) transformation of $\mathbf{P}$. If $\mathbf{L}'$ and $\mathbf{P}$, as input, do not have this property then this equation represents equations of condition (with the residuals omitted) and the following equation for $\mathbf{B}_0$ provides a least-squares solution.

cated), even though the input link coordinates possessed only slide-rule accuracy. The α-helix produced had a rise per residue of 1·4985 Å.

A geometrical analogy to the refinement of **B** may be helpful. The columns of **B** represent three vectors in an arbitrary orientation, except that ideally they should be mutually perpendicular and of unit length. If these three vectors are taken to represent a unit cell, then the corresponding columns of $\tilde{\mathbf{B}}^{-1}$ are the reciprocal lattice vectors. If we continually replace the (crystallographic) vector **a** by $\frac{1}{2}(\mathbf{a}+\mathbf{a}^*)$ (likewise **b** and **c**), we are bound to approach the unit cube as unit cell. Clearly the final $\mathbf{B}_n$ splits the difference between the matrix required to relate the initial **L′** to **P** and that required to relate **P** to the initial **L′**.

The program must also be supplied with a representation of the first group $\mathbf{p}_0$, called the *root group*, regardless of whether guide points are also provided for it, and this group, too, is replaced by $\mathbf{C}_n\mathbf{P}$ before use, where $\mathbf{C}_n$ is derived and refined in just the same way as $\mathbf{B}_n$.

2.2. *Refinement procedures*

2.2.1. *General*

Each probe which is to be refined has a free end, most recently built, and a root end which is generally attached to a finalized and therefore fixed portion of the computed model. To begin with, however, when only the root group or a few residues have been built, the allowed length of a probe may exceed the then existing length of computed model, in which case the root end of the probe will have a termination group as its footing. This group must be provided among the standard groups and is automatically incorporated at the end of the listing. It normally includes only dummy atoms, parameter cards, and a free cast-off card sufficient to provide the root group with six degrees of freedom. Alternatively, it may contain just a fixed cast-off card in which case the root group will be fixed where it is planted, as may be useful if it is desired on one occasion to build further upon the results obtained on an earlier occasion.

Since the tip of each probe is always free (though it may be guided) equations of constraint, in the mathematical sense, are not required. The only consequence of the chain character of the problem is that the parameters are highly correlated; the fact that the computed coordinates of any one atom may be a function of many parameters is not a complication.

Let **D** be a column vector, partitionable in threes, each three elements comprising difference vectors of the form $\mathbf{d}=\mathbf{r}_{\text{guide}}-\mathbf{r}_{\text{model}}$ where the **r**'s are position vectors, then the *equations of condition* may be written

$$\mathbf{D}=\frac{\partial\mathbf{R}}{\partial\boldsymbol{\Theta}}\boldsymbol{\Theta}+\boldsymbol{\varepsilon}+\text{2nd and higher order terms}$$

where $\boldsymbol{\varepsilon}$ is an irreducible residue of the same form as **D**, **R** relates to the *model* coordinates, $\boldsymbol{\Theta}$ is a column vector containing the rotational shifts to be applied and $\frac{\partial\mathbf{R}}{\partial\boldsymbol{\Theta}}$ is a rectangular matrix of the form

$$\frac{\partial\mathbf{R}}{\partial\boldsymbol{\Theta}}=\begin{bmatrix}\frac{\partial x_1}{\partial\theta_1} & \frac{\partial x_1}{\partial\theta_2} & \cdots\\ \frac{\partial y_1}{\partial\theta_1} & & \\ \frac{\partial z_1}{\partial\theta_1} & & \\ \frac{\partial x_2}{\partial\theta_1} & & \\ \vdots & & \end{bmatrix}$$

the θ_i being elements of the vector $\boldsymbol{\Theta}$.

We wish to minimize $\tilde{\boldsymbol{\varepsilon}}\boldsymbol{\varepsilon}$, which is the sum of the squares of the distances between the guide points and the refined model coordinates. Linear least-squares theory then sets

$$\frac{\partial\tilde{\mathbf{R}}}{\partial\boldsymbol{\Theta}}\boldsymbol{\varepsilon}=\mathbf{0}$$

giving the *normal equations*

$$\frac{\partial\tilde{\mathbf{R}}}{\partial\boldsymbol{\Theta}}\mathbf{D}=\frac{\partial\tilde{\mathbf{R}}}{\partial\boldsymbol{\Theta}}\frac{\partial\mathbf{R}}{\partial\boldsymbol{\Theta}}\boldsymbol{\Theta}$$

to which the solution is

$$\boldsymbol{\Theta}=\left(\frac{\partial\tilde{\mathbf{R}}}{\partial\boldsymbol{\Theta}}\frac{\partial\mathbf{R}}{\partial\boldsymbol{\Theta}}\right)^{-1}\frac{\partial\tilde{\mathbf{R}}}{\partial\boldsymbol{\Theta}}\mathbf{D}$$

in which the quantities within the round brackets form the *normal matrix*. This theory ignores the presence of the second and higher order terms in the equations of condition. Including them leads to a quite intractable problem, but it is clear that they must be important here because the movements produced by large rotations cannot be expressed as a linear function of each θ. We resort to the usual technique of applying repeated cycles of refinement (one would suffice for a linear problem) and aim to make the convergence as rapid as possible.

We define the spindle of any parameter as a vector of unit length in the direction of the line joining the points (atoms) in the computed structure between which a parameter entry has been inserted. These entries originate in the standard groups from which the final listing is assembled. The origin of a spindle is always to be at the end nearer the root and simple rules have been laid down for locating the origin of a spindle at a fork in the chain where it cannot always be arranged that the origin is on an adjacent entry in the listing. If we denote a spindle vector as **n** then the components of $\mathbf{n}_i\times\mathbf{r}_{ij}$ form a 3×1 submatrix of $\partial\mathbf{R}/\partial\boldsymbol{\Theta}$, relating parameter i to atom j where $\mathbf{r}_{ij}$ is the position vector of the atom expressed relative to the origin of the spindle. All the quantities appearing in the normal equations are therefore readily accessible. Only

guided atoms on the free side of a parameter have derivatives with respect to that parameter, since we treat each parameter as if it contributes only to movements to the free side of it. (If, in any given instance, it is required that the free end of a probe should not move and that its root end should, then a combination of shifts exists and may be found which expresses this, provided that a suitable termination group is present.)

Elements of the matrix $(\partial\widetilde{\mathbf{R}}/\partial\boldsymbol{\Theta}\mathbf{D})$ (of which there is one for each parameter) are formally analogous to the total couple exerted on each spindle supposing the guided atoms to be tied to the guide points with elastic in which the tension is proportional to the corresponding $|\mathbf{d}|$. No weighting scheme has been employed in this work (mainly for storage reasons), but a weighting of the equations of condition would correspond to differing strengths of elastic. In any case, the response of the system (*i.e.* shifts produced in a least-squares cycle) is elastic with uniform torsion constants in the spindles. These too, could be weighted to represent the differing stiffnesses of inter-bond angles and dihedral angles (see §4.2); however, these elastic analogies are only valid if one supposes that for any least-squares cycle, the initial conformation is an equilibrium one.

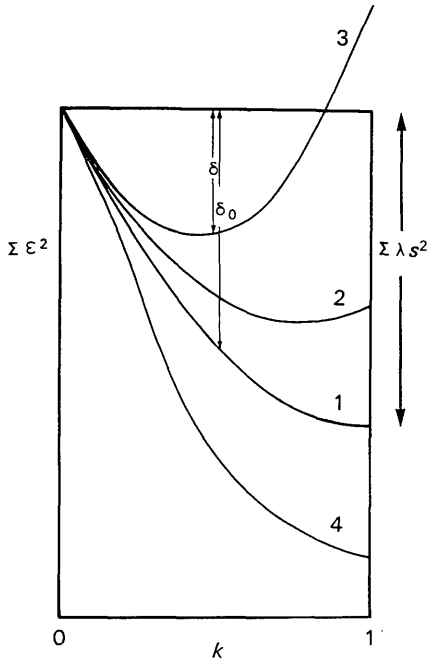

Fig. 3. Schematic representation of the possible behaviour of the sum of squares of errors in a least-squares problem. k is a scalar fraction such that the shifts actually applied are k times the calculated shifts. Curve 1 is parabolic and corresponds to a linear problem having no 2nd and higher order terms in the equations of condition. The minimum occurs for full shifts ($k=1$). Curve 2 corresponds to a non-linear problem where the non-linearity is hindering, but not enough to prevent convergence. Curve 3 corresponds to a case where the non-linear terms are so large that the error actually rises if full shifts are applied, giving rise to an unstable situation. Curve 4 is a case where the non-linearity is actually helping. All such curves have the same slope at $k=0$ and all can occur in practice. The ratio of the ordinates δ/δ_0 is $g(k)$ (see text).

2.2.2. *Reversion, or the use of fractional shifts*

Suppose that the solution $\boldsymbol{\Theta}$ to the normal equations has been found and that a scalar multiple, k, of these shifts is applied, then for a linear problem the sum of the squares of the errors descends parabolically as k runs from 0 to 1 and the overall drop is equal to $\Sigma\lambda s^2$ where λ are the eigenvalues of the normal matrix and s are the corresponding eigenshifts, the summation being over those eigenvectors which contribute to $\boldsymbol{\Theta}$. This will become clearer in the next section. For the present, the relevant point is that $\Sigma\lambda s^2$ is an available quantity, so that the parameters of the parabola are fully determined without reference to the actual behaviour of the sum of the squares of the errors.

We may then define a linearity index, $g(k)$, by

$$g(k)=\frac{\text{reduction in sum of squares of errors actually obtained}}{\text{reduction in sum of squares of errors that would be produced by the same shifts if the system was linear}}=\frac{\delta}{\delta_0}.$$

(Fig. 3).

$g(k)$ approaches unity if k is small or if the elements of $\boldsymbol{\Theta}$ are themselves small*. $g(k)$ exceeds unity if the non-linearity is helping, is less than unity if it is hindering and is negative in unstable situations (Fig. 3), and as such it serves to characterize the behaviour of the problem and provides a criterion for the application of fractional shifts. Normally, full shifts ($k=1$) are applied and if the linearity index is found to be below some minimum acceptable value g_0 (normally zero), then back shifts are applied to halve k, this halving being repeated until the criterion is met, as it must be, provided $g_0<1$. We call this process reversion and it is used only defensively, *i.e.* if the technique of filtering has failed to produce a satisfactory value of $g(1)$.

2.2.3. *Filtering*

Filtering provides a means of selecting those linear combinations of parameters which are most effective in improving the fit, and of excluding those combinations of parameters which disturb the structure grossly and to little advantage. A fixed filter has been employed previously (Diamond, 1958) in a different context with a linear problem. In this work we use a sliding filter, principally to combat non-linearity. A sliding filter selects these combinations of parameters by reference to the current value of the residual. The theory will now be developed, ignoring the non-linear terms until the end.

---

* Computationally, on a final cycle, when the elements of $\boldsymbol{\Theta}$ are very small, $g(k)$ becomes indeterminate.

We rewrite the normal equations as

$$\mathbf{T}\frac{\partial\widetilde{\mathbf{R}}}{\partial\boldsymbol{\Theta}}\mathbf{D}=\mathbf{T}\frac{\partial\widetilde{\mathbf{R}}}{\partial\boldsymbol{\Theta}}\frac{\partial\mathbf{R}}{\partial\boldsymbol{\Theta}}\tilde{\mathbf{T}}\mathbf{T}\boldsymbol{\Theta}$$

in which $\mathbf{T}$ is orthogonal and

$$\boldsymbol{\Lambda}=\mathbf{T}\frac{\partial\widetilde{\mathbf{R}}}{\partial\boldsymbol{\Theta}}\frac{\partial\mathbf{R}}{\partial\boldsymbol{\Theta}}\tilde{\mathbf{T}}$$

is a diagonal matrix containing the eigenvalues λ of the normal matrix and $\mathbf{T}$ contains its eigenvectors (rows).

We then define the two column vectors*

$$\mathbf{U}=\mathbf{T}\frac{\partial\widetilde{\mathbf{R}}}{\partial\boldsymbol{\Theta}}\mathbf{D}\quad\text{and}\quad\mathbf{S}=\mathbf{T}\boldsymbol{\Theta},$$

in which the elements s of $\mathbf{S}$ are the eigenshifts obtainable from

$$\mathbf{S}=\boldsymbol{\Lambda}^{-1}\mathbf{U}$$

in which the inversion is now trivial. We recall a number of important properties (*cf.* Diamond, 1958):

(i) The elements of $\mathbf{S}$ are independent of each other because $\boldsymbol{\Lambda}^{-1}$ is diagonal, *i.e.* they are uncorrelated, so that if any s_i has its value disturbed in any way, or is inaccurately determined, this has no repercussions on the values of the remaining s_j, unlike the elements of $\boldsymbol{\Theta}$, which are heavily correlated.

(ii) If we regard the vector $\boldsymbol{\Theta}$ as the position vector of the solution point in a polydimensional parameter space, then $\mathbf{S}$ is also a position vector for the solution point, expressed relative to a rotated set of axes.

(iii) If $\boldsymbol{\varepsilon}'=\boldsymbol{\varepsilon}+\delta\boldsymbol{\varepsilon}$ is the matrix of residuals at a point $\delta\boldsymbol{\Theta}$ away from the solution point, *i.e.* at $\boldsymbol{\Theta}+\delta\boldsymbol{\Theta}$, then the sum of the squares of the errors at this point is

$$\tilde{\boldsymbol{\varepsilon}}'\boldsymbol{\varepsilon}'=\tilde{\boldsymbol{\varepsilon}}\boldsymbol{\varepsilon}+\widetilde{\delta\boldsymbol{\Theta}}\frac{\partial\widetilde{\mathbf{R}}}{\partial\boldsymbol{\Theta}}\frac{\partial\mathbf{R}}{\partial\boldsymbol{\Theta}}\delta\boldsymbol{\Theta}$$

since

$$\delta\boldsymbol{\varepsilon}=-\frac{\partial\mathbf{R}}{\partial\boldsymbol{\Theta}}\delta\boldsymbol{\Theta}\quad\text{and}\quad\frac{\partial\widetilde{\mathbf{R}}}{\partial\boldsymbol{\Theta}}\boldsymbol{\varepsilon}=\mathbf{0}.$$

This may be rewritten

$$\tilde{\boldsymbol{\varepsilon}}'\boldsymbol{\varepsilon}'=\tilde{\boldsymbol{\varepsilon}}\boldsymbol{\varepsilon}+\widetilde{\delta\mathbf{S}}\boldsymbol{\Lambda}\delta\mathbf{S}\quad(\delta\mathbf{S}=\mathbf{T}\delta\boldsymbol{\Theta}),$$

which shows that a contour of constant error in the neighbourhood of the solution is an ellipsoid in parameter space, this ellipsoid being wholly determined by the normal matrix, the eigenvectors of which are the principal axes of the ellipsoid, the lengths of which are proportional to each $\lambda^{-\frac{1}{2}}$. Thus a small λ implies a large uncertainty in the corresponding direction in the position of the solution point, but by (i) this does not affect the precision with which the remaining coordinates (on the rotated axes) are determined.

(iv) We may replace $\mathbf{S}$ by

$$\mathbf{S}_f=\mathbf{Z}_f\boldsymbol{\Lambda}^{-1}\mathbf{U}$$

and obtain the adopted shifts

$$\boldsymbol{\Theta}_f=\tilde{\mathbf{T}}\mathbf{S}_f,$$

where $\mathbf{Z}_f$ is a filter matrix having 1 on the diagonal in the positions of the f largest λ and zeros elsewhere. By suitably choosing the value of f we find those coordinates, s, of the solution point which are most accurately determined and which, by the same token, play the largest part in reducing the sum of the squares of the errors. These are clearly the most important eigenshifts and they will be referred to as dominant. The zeros at the remaining diagonal positions of $\mathbf{Z}_f$ set the remaining elements of $\mathbf{S}$ to zero.

The justification for suppressing the remaining eigenshifts is threefold. Firstly, we note that $\boldsymbol{\Theta}$ is built up by adding together a number of mutually perpendicular component vectors, the eigenshifts being their magnitudes and the eigenvectors their directions; consequently every such addition, regardless of the sign of each s, increases the distance of the point $\boldsymbol{\Theta}$ from the origin and contributes an amount s^2 to $\tilde{\boldsymbol{\Theta}}\boldsymbol{\Theta}$. If each spindle (bond) has unit torsion constant associated with it then $\tilde{\boldsymbol{\Theta}}\boldsymbol{\Theta}$ is a measure of the elastic strain energy imposed on the structure in order to fit the guide points (assuming its initial conformation to be an equilibrium one).

Now

$$\tilde{\boldsymbol{\Theta}}_f\boldsymbol{\Theta}_f=\tilde{\mathbf{S}}_f\mathbf{T}\tilde{\mathbf{T}}\mathbf{S}_f=\tilde{\mathbf{S}}_f\mathbf{S}_f=\sum_1^f s^2$$

and each s_i is proportional to $\lambda_i^{-1/2}$ so that large strain energies are introduced by small eigenvalues.

Secondly [by (iii)] the reduction in $\tilde{\boldsymbol{\varepsilon}}'\boldsymbol{\varepsilon}'$ produced by moving to the solution point from the origin is $\tilde{\mathbf{S}}\boldsymbol{\Lambda}\mathbf{S}$, *i.e.* each eigenshift contributes λs^2 to the reduction of $\tilde{\boldsymbol{\varepsilon}}'\boldsymbol{\varepsilon}'$. Combining these two results we see that for each and every eigenshift which is included

$$\text{eigenvalue}=\frac{\text{decrement in }\tilde{\boldsymbol{\varepsilon}}\boldsymbol{\varepsilon}\text{ attributable to this eigenshift}}{\text{increment in }\tilde{\boldsymbol{\Theta}}\boldsymbol{\Theta}\text{ attributable to this eigenshift}}$$

or

$$\frac{\text{improvement in fit}}{\text{strain energy imposed}}.$$

Thus each eigenvalue forms an inverted price tag, the large ones cheaply producing a large improvement in the fitting, and the small ones introducing large deformations and conferring very little benefit, so that there may well be a minimum value of λ below which it is not worth going. It also follows from this that underdetermined problems where, in principle, many solutions exist (*i.e.* those with one or more vanishing eigenvalue) are provided by filtering with the most economical solution, *i.e.* of the many solutions that could be provided the one which is actually generated is the one with smallest $\tilde{\boldsymbol{\Theta}}\boldsymbol{\Theta}$.

* This matrix $\mathbf{S}$ is unrelated to the matrix $\mathbf{S}$ used in §2.1, which will not be referred to again.

The third reason for filtering is that the dominant eigenvectors have the greatest range of convergence. The equations of condition including second-order terms may be written

$$\mathbf{D}=\boldsymbol{\varepsilon}+\frac{\partial \mathbf{R}}{\partial \boldsymbol{\Theta}}\boldsymbol{\Theta}+\mathbf{E}$$

where the ith element of the column matrix $\mathbf{E}$ is of the form

$$\tfrac{1}{2}\sum_j\sum_k \frac{\partial^2 r_i}{\partial\theta_j\partial\theta_k}\theta_j\theta_k=\tilde{\boldsymbol{\Theta}}\mathbf{F}_i\boldsymbol{\Theta}\ ,$$

where each $\mathbf{F}_i$ is a square matrix containing the second derivatives of $\mathbf{r}_i$. On transforming from $\boldsymbol{\Theta}$ to $\mathbf{S}$ we obtain

$$\mathbf{D}=\boldsymbol{\varepsilon}+\frac{\partial \mathbf{R}}{\partial \boldsymbol{\Theta}}\tilde{\mathbf{T}}\mathbf{S}+\text{column vector,}$$

the ith element of which is $\tilde{\mathbf{S}}\mathbf{T}\mathbf{F}_i\tilde{\mathbf{T}}\mathbf{S}$.

Now the columns of $(\partial\mathbf{R}/\partial\boldsymbol{\Theta})\tilde{\mathbf{T}}$ are orthogonal and proportional to $\lambda^{\frac{1}{2}}$ in the sense that

$$\mathbf{T}\frac{\partial\tilde{\mathbf{R}}}{\partial\boldsymbol{\Theta}}\frac{\partial\mathbf{R}}{\partial\boldsymbol{\Theta}}\tilde{\mathbf{T}}=\boldsymbol{\Lambda}\ ,$$

so that in this form the linear parts of the equations of condition represent, for the dominant eigenshifts, equations having large coefficients with small values of the unknowns, s. The terms involving squares and products of the dominant eigenshifts are then negligible provided that the transformation $\mathbf{T}$ does not yield matrices $\mathbf{T}\mathbf{F}_i\tilde{\mathbf{T}}$ with systematically enlarged elements in positions corresponding to the dominant eigenshifts – and there is no reason why it should since each $\mathbf{F}_i$ is quite unrelated to the normal matrix which $\mathbf{T}$ diagonalizes.

This is the basis of the sliding filter which causes the number f of degrees of freedom allowed to $\boldsymbol{\Theta}$ by $\mathbf{Z}_f$ to be dependent on the current value of the r.m.s. error, which is the best available index of the movement in $\boldsymbol{\Theta}$ to be expected on any one cycle of refinement. For a linear problem the increment in $\tilde{\boldsymbol{\Theta}}\boldsymbol{\Theta}$ for each eigenshift is the decrement in $\tilde{\boldsymbol{\varepsilon}}\boldsymbol{\varepsilon}$ attributable to this eigenshift divided by λ. In this problem, where macro-rotations are involved, it is clear that the r.m.s. rotation produced in any one cycle of refinement should not exceed ~ 1 radian; otherwise the shifts actually produced will bear little relation to the required shifts. This immediately suggests a minimum value for λ such that large rotations do not arise. In the program as written, f is set so that three conditions are all satisfied, these are that

(i) the smallest eigenvalue included, $\lambda_{\min}\geq\tilde{\boldsymbol{\varepsilon}}\boldsymbol{\varepsilon}\ .\ C_1/n$ where n is the order of the normal matrix and C_1 is a constant.

(ii) $\lambda_{\min}\geq C_2\lambda_{\max}$

(iii) $f\leq f_{\max}$

where C_1, C_2 and $f_{\max}$ are all constants which are read in as data. Here $C_1^{-1/2}$ is the largest r.m.s. shift which one is prepared to encounter in any one cycle of refinement, and the function $\tilde{\boldsymbol{\varepsilon}}\boldsymbol{\varepsilon}C_1/n$ is recalculated at the beginning of each cycle so that the number of degrees of freedom allowed to the solution, f, increases as the final solution is approached. C_1 may be set generously, *e.g.* $C_1^{-1/2}\sim 3$ radians, relying on the technique of reversion in cases where this results in an over-stepping of the convergent region. (See example 3.1 below.) C_2 serves to ensure that the least accurate eigenshift included is not worse than $C_2^{-1/2}$ times less accurate than the most accurate one. The third condition may be useful if it is desired to stipulate explicitly the number of degrees of freedom to be allowed in any case.

2.2.4. *Delays*

This is a technique designed to provide a measure of control over the way in which the strain energy $\tilde{\boldsymbol{\Theta}}\boldsymbol{\Theta}$ is distributed amongst the various parameters (see also the discussion §4.2). Two types of parameter are recognized by the program and these are usually used so that spindles coinciding with covalent bonds are set up as type 1 parameters and spindles which are set up to vary inter-bond angles are type 2 parameters, though this usage may be varied. Two delay constants are then set for each probe. If delay 1 is 2, for example, the first two type 1 parameters (counting from the tip of the probe towards the root) are switched off so that the tip of the probe then forms a rigid group on the end of a flexible mounting. Delay 2 is the number of further type 1 parameters that are then counted before any type 2 parameters are switched on. By this means one may, for example, keep the inter-bond angles constant while most of the fitting is being done, but allow them freedom only near the root of a probe several residues long when the probe reaches far enough ahead to give a fairly firm indication of the need to increase or decrease a bond angle. Such a probe is usually the last and usually has C_2 set conservatively, *i.e.* to admit only a few eigenshifts, so that large strains do not arise. By this means variations of a few degrees may be admitted to the tetrahedral angle at α carbon atoms when these are really called for.

If it is desired to alter the internal configuration of a peptide grouping this must be done with type 2 parameters with delays appropriately set so that the configuration of a link is never disturbed until after it has been used as prefactor in the building process to generate the next residue.

2.2.5. *Movement*

When the rotational shifts $\boldsymbol{\Theta}$ which are to be applied to the structure have been evaluated, they must clearly be applied in a rigorously circular fashion to conserve shape. Conventional rotation matrices are not used for this purpose as the alternative is very much simpler and probably has an advantage in speed, except perhaps near the root of a fairly long probe where many atoms may require to be moved through the same angle.

When an atom is to be rotated through an angle θ about a particular spindle the quantities immediately available are θ, the spindle vector $\mathbf{n}$, and the position vectors of the atom and of the origin of the spindle. If we obtain

$$\boldsymbol{\theta} = \theta \mathbf{n}$$

and $\mathbf{r}_0$ the initial position vector of the atom relative to the origin of the spindle, then the final position of the atom, relative to the same origin, is

$$\sum_0^\infty \mathbf{r}_m$$

where

$$\mathbf{r}_m = \frac{1}{m} \boldsymbol{\theta} \times \mathbf{r}_{m-1} .$$

This is the three-dimensional counterpart of using the power series $e^{i\theta} = 1 + i\theta - \frac{\theta^2}{2} - \ldots$ to achieve a rotation θ, and the summation may be taken to any required precision.

These rotational shifts are applied sequentially starting at the free end of the probe. As each parameter is encountered, all the atoms between it and the free end of the probe (or side chain) are moved by the above process, successive parameters taking the atoms from the positions they have been left in by previous operations.

If this process is not carried out accurately, cumulative errors will arise in the building process as described earlier. Such deformations of the link involving changes of its length in excess of 2×10^{-4} Å have never been encountered.

3. Examples

3.1. *An unguided bridge*

In this example a chain of five links was built between two regions where guide coordinates were given, there being no guidance in the intervening span. This particular example is chosen to illustrate the range of convergence of the mathematical procedures just described, rather than as a demonstration of their ability to postulate chemically reasonable structures. There is, in fact, a threefold infinity of solutions to this problem as presented to the machine, and the solution actually found is determined by the initial conformation and the minimization of $\tilde{\boldsymbol{\Theta}}\boldsymbol{\Theta}$ at each step and is done without reference to the steric map of the resulting conformation. It is recognized that this additional criterion must be employed before any useful 'structure-guessing' can be done, but, as explained in the introduction, this was not the primary purpose of the program. In the discussion we indicate ways in which it is thought that van der Waals and other interactions may be introduced.

Fig. 4(*a*)–(*k*) shows the initial conformation (*a*), which is α-helical, and the conformation after each of ten cycles of the long probe. The end at the bottom of each figure is the root end (C terminal in this case) and there are two links here whose positions do not alter. The lines marked heavily are main chain bonds where rotation is allowed, these occur in pairs each side of C_α except in one case where the residue is proline. The sequence in this region is HIS, PRO, GLY, ASN, PHE, and coordinates were calculated for all these side chains on every cycle as shown in the first and last diagrams. These side shains have a total of six rotatable bonds, none of which are guided, so that the derivatives $\partial r/\partial\theta$ for these angles are all zero. This gives rise to six vanishing eigenvalues with eigenvectors involving these parameters only. The filtering process excludes these, so that no rotations arise in any of the side chains.

There are then ten main chain parameters so that the order of the normal matrix is 16; however, it has

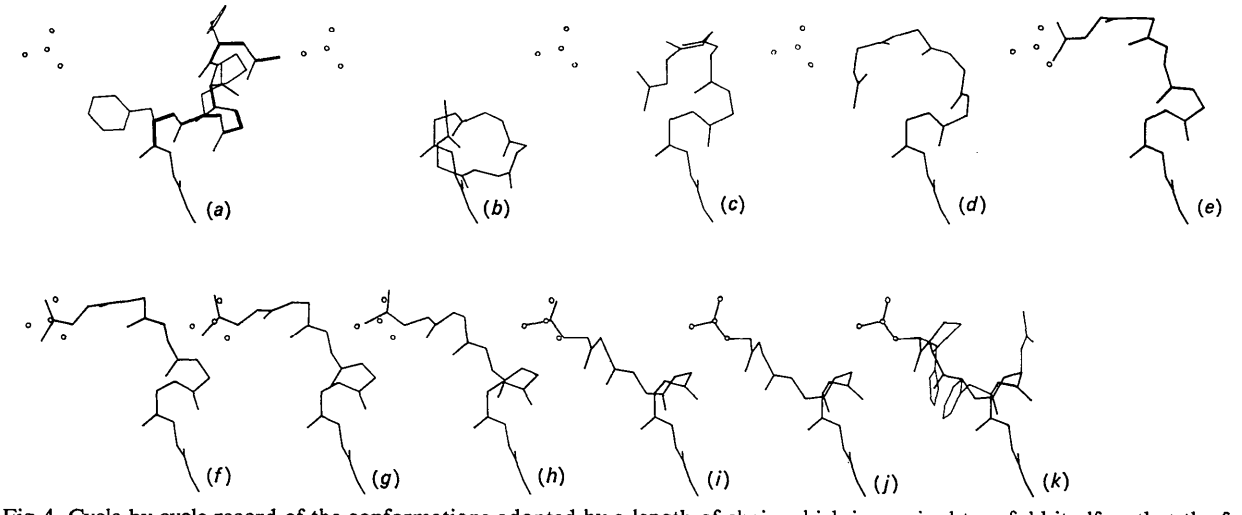

Fig. 4. Cycle by cycle record of the conformations adopted by a length of chain which is required to refold itself so that the free end of the chain comes into coincidence with the guide points shown as open circles on the left of each diagram. There are two peptides at the foot of each figure which do not move; they were guided into those positions and their positions were finalized just before this stage of the calculation was reached. The initial conformation shown in (*a*) is α-helical. The side chains were carried throughout the calculation, but for clarity are shown in only the first and last figures.

only six non-vanishing eigenvalues for the following reason. The four guided atoms form a rigid group and it requires six degrees of freedom to position and orient such a group; accordingly any parameters in excess of six are superfluous. This means that if there are n (>6) parameters contributing to the position and orientation of this rigid group then there are n–6 combinations of shifts in these parameters which leave the position and orientation of the group unchanged. Such combinations evidently contribute (if included) to $\tilde{\Theta}\Theta$ but do nothing to reduce $\tilde{\varepsilon}\varepsilon$, *i.e.* they have vanishing eigenvalues.*

The parameter nearest the tip of the probe is clearly responsible by itself for one vanishing eigenvalue because it only rotates one atom about a line through itself. (It could, in fact, have been completely excluded by setting delay 1 = 1.) The eigenvalue spectrum thus consists of six significant eigenvalues in the range 10^3 to 10^{-1}, then three of order 10^{-6}, which are computed zeros associated with the superfluous parameters in the main chain span, then one of order 10^{-14} ('round-off noise') associated with the parameter at the tip of the probe, and six identically zero associated with the side chains. The filtering process in this instance never admits more than the first six.

* These considerations imply that if an intermediate portion of any chain is ever to be modified without altering the two regions it connects, then at least seven parameters in the modified region must be varied, except in special cases, *e.g.* two distant but exactly collinear bonds.

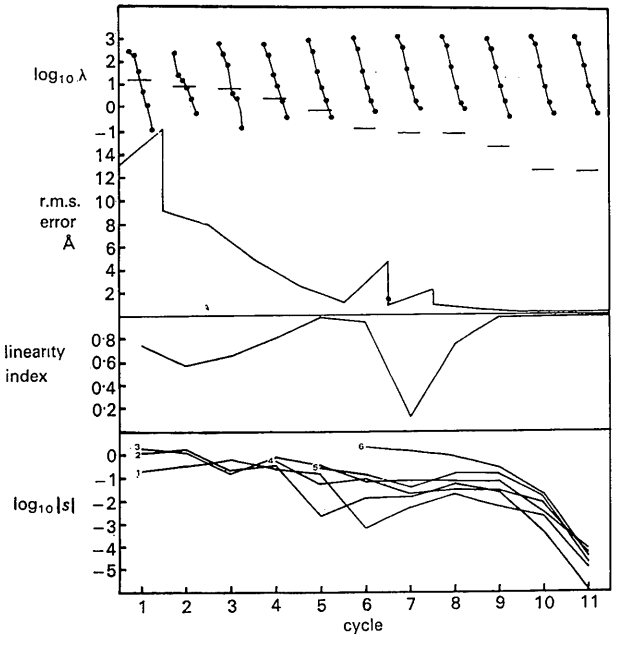

Fig. 5. Log of the calculations for Fig. 4. At the top of the diagram are shown the eigenvalue spectra given as $\log_{10}\lambda$, on which is superimposed a bar at the filter level, $\lambda_{\min}$, given by the r.m.s. error with $C_1=1$ as explained in §2.2.3. The constant C_2 was also set so that not more than four decades of the spectrum could pass the filter. Below this are given the r.m.s. error in Å between the four guide points and the tip of the probe, the linearity index (evaluated after reversion, if any) and the decimal logarithm of the modulus of each eigenshift (evaluated before reversion, if any). Note the tendency for reversions to be required if any $\log|s|$ exceeds zero, thereby contributing more than 1 radian² to $\tilde{\Theta}\Theta$, and how the sliding filter operates to prevent the premature inclusion of such eigenshifts. For fuller explanation see text.

Fig. 5 summarizes the sequence of events for eleven cycles of refinement (the last of which merely establishes that there is no further change to be made). The eigenvalue spectrum, the linearity index and the eigenshifts are shown for each cycle and the r.m.s. error in Å is shown between cycles. Each eigenvalue spectrum is shown on a logarithmic scale, only the first six being shown. The bar across each spectrum is the level for $\lambda_{\min}=\tilde{\varepsilon}\varepsilon C_1/n$ determined by the r.m.s. error at the end of the previous cycle and only eigenshifts with eigenvalues above the bar are admitted by the filter. Initially the tip of the probe is some 13 Å from its guide points and with C_1 set at unity only the first three eigenshifts are included. These yield shifts with an r.m.s. value (over the nine significant main chain bonds) of $\sim$0·75 radians, which oversteps the convergent region, and the r.m.s. error rises to 16 Å; however, one reversion (k=0·5) reduces the residual to 9 Å with a linearity index of 0·75. It has been estimated that if all six eigenshifts had been included on this cycle then the r.m.s. shift would have been $\sim$300 radians so that little more than one thousandth of the calculated shifts could have been applied, implying that some hundreds of cycles would have been required to achieve convergence.

Only after cycle three, when the distance is under 5 Å, does the orientation of the guiding group exert any influence. The sixth eigenshift is introduced for the first time on cycle six. Note that the conformation after cycle five is more or less correct except that the tip of the probe is required to rotate in its own plane. Such a movement requires large alterations to the span to achieve quite a small improvement, *i.e.* it is evident from the conformation that no further progress can be made unless small eigenvalues are admitted. The introduction of the sixth eigenshift again produces such large shifts that a double reversion (k= 0·25) is needed on cycle six and a single one on cycle seven (which might have benefited from a double one, as the single one yields a linearity index of only 0·12). From there on the refinement behaves smoothly, the final r.m.s. error between the guide points and the probe being 0·0044 Å.

3.2. *A short bridge*

This example is taken from hen egg-white lysozyme (Blake, Koenig, Mair, North, Phillips & Sarma, 1965). At the time of writing Phillips and co-workers are making a detailed study of their 2 Å Fourier map of this molecule obtaining provisional coordinates for the majority of atoms, and these coordinates are being used as guide coordinates to obtain a computed model.

For the most part, these computations are straightforward, but the example chosen here is of one point in the chain where it was thought to be worth while to use the computer method to look for alternative solutions. This is residue 16 which is glycine, so that the Fourier map shows no side chain to help in orienting the main chain links, only protuberances which are probably carbonyl oxygen atoms, but the tube of density between residues 15 and 17 seems to leave some room for manoeuvre.

This is a case where the number of parameters determining the conformation of the bridge is just equal to the number of degrees of freedom required by the tip of the probe, so that there is not a continuum of solutions as in the previous example but a finite number only (Fig. 6). The solution actually obtained therefore depends solely on the initial conformation, and this may be varied by the use of parameter cards. The first solution was obtained without using any parameter cards, so that the initial conformation of the links concerned was α-helical. The second solution was found using the first solution as initial conformation but modified by the inversion of the 16–17 link and the application of $-70°$ about φ_{15} [*i.e.* N(15)–C(15)$_\alpha$]. The third, fourth and fifth solutions were all obtained using the second solution as initial conformation but with one or other or both of the peptides 15–16 and 16–17 inverted and with the same treatment of φ_{15}. (The first also differs from the other four in that ψ_{17} [*i.e.* C(17)$_\alpha$–C(17)] was rigid in this case and free in all the others, *i.e.* N(17) was in most cases free to move on a circle.) This yielded a total of four final conformations, two of the initial conformations converging to the same result. All four solutions represent stable minima in the least-squares sense and three of them fit well at both ends of the bridge and permit the satisfactory continuation of building. The choice between them may then be made on the basis of the steric map (the necessary angles are listed, of course) and comparison with the Fourier map. Alternatively, the most promising one may be used as an initial conformation and additional guidance provided if the Fourier map indicates that it is not satisfactory as it stands.

3.3. *Other examples*

The technique is also being applied to myoglobin where the quality of the guide coordinates available is extremely good. Here the r.m.s. error between guide coordinates and computed model is typically 0·25 Å if all main chain atoms are used alone for guidance, or together with just sufficient side chain atoms to specify configuration, or about 0·35 Å if all atoms are used. These figures depend to some extent on the control quantities used.

Haemoglobin has reached a rather different stage. Perutz (1965) has built a tentative model of horse oxyhaemoglobin based on a knowledge of the sequence, a 6 Å Fourier map (Cullis, Muirhead, Perutz, Rossmann & North, 1962) and a knowledge of the detailed structure of myoglobin. This model is unlikely to be correct in every detail, but it is certainly correct in its main features and forms an excellent basis for further work. A computed replica of this model is being set up, and here the facilities of the long probe and of parameter cards are particularly useful. For example, if side chains are directed with the use only of parameter cards and no guide coordinates, or not at all, then a complete α-helix only needs guide coordinates at its ends, and the whole helix may form one long probe. A satisfactory representation of the *FG* corner and all the remaining structure down to the C-terminus of the β chain, *i.e.* about 425 atoms, has been produced with 60 guide atoms and 12 param-

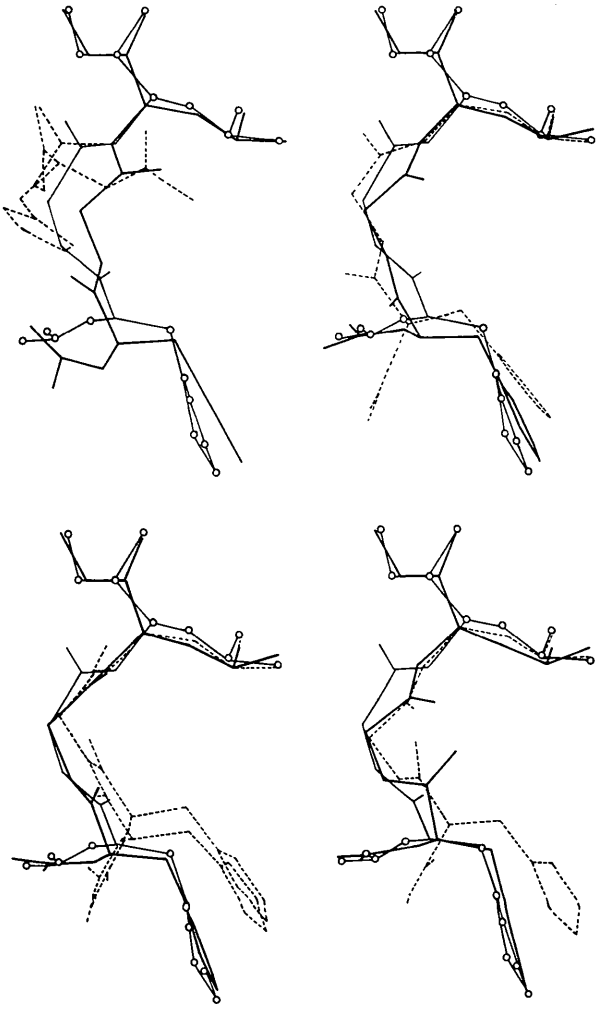

Fig. 6. Four short bridges built to the same guide coordinates using five different initial conformations (two of the final conformations being identical). The heavy lines in each case show the final conformation. The open circles are the guide coordinates joined by thin lines which, in the unguided portion, show the conformation obtained by direct reading of the electron density map. The broken lines indicate in each case the initial conformation, and the root end is at the top. The sequence hnre is 15 HIS, 16 GLY, 17 LEU.

eter cards* (Fig. 7). The time for such a calculation depends enormously on the probe lengths used, but using the long probe and four short probes of 1,2,3, and 6 residues this takes about 3 minutes on an IBM 7094 or 12 minutes on an IBM 7090.

* It happens that the structure of valine as provided by Parthasarathy & Ramachandran (1966), which is currently being used, is in a conformation which may not be incorporated in an α-helix, and needs a rotation of $\sim +110°$ in χ_1 (*i.e.* the $C_\alpha - C_\beta$ bond) to be imposed either by guidance or by parameter cards. This accounts for 7 of the 12 such cards, and a further 2 have no function except to position the others correctly.

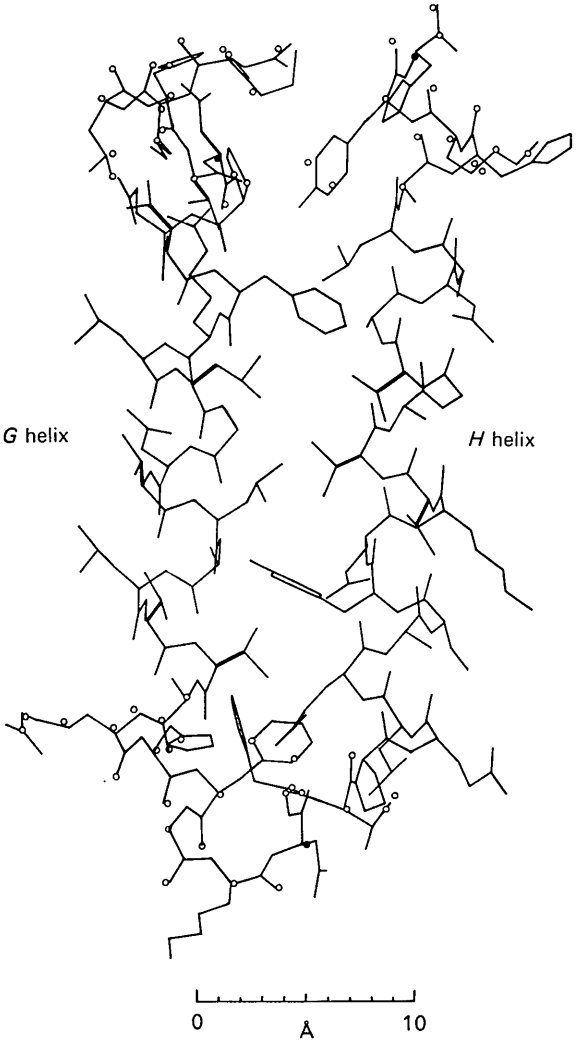

Fig. 7. A computed replica of the *FG* corner and the *G* and *H* helices of the β chain of horse oxyhaemoglobin. The guide points are taken from Perutz's model (Perutz, 1965) and (except for three) are shown as open circles close to the corresponding computed points. The three solid circles are guide points in the plane of the paper. Bonds which had rotations imposed on their initial conformation by parameter cards are shown darker; these include all valine side chains.

4. Discussion

4.1. *Interactions between atoms*

It is recognized that a weakness of the method in its present form is that it pays no attention to van der Waals interactions or any other influence that in reality affects the true conformation, and which certainly affects the acceptability or otherwise of any hypothetical conformation. It has hitherto been considered that it is for the originator of the guide coordinates to see to it that these do not imply impossible conformations, and that the user should scrutinize the results with a steric map. In most applications this is satisfactory, but for model postulating, as in the first example above, this is not sufficient.

The problem arises, therefore, of devising a procedure which simultaneously heeds two or more criteria which may even be in conflict, and which certainly will be in conflict, at least temporarily, if the guide coordinates require that a residue must cross a forbidden region of the steric map in order to reach another permitted region. Any technique which heeds the various criteria in turn, instead of simultaneously, is liable to oscillate. The proposals outlined below appear to offer at least some hope of success.

Suppose that a bridge is being erected, as in the first example, to reach a certain footing without anything being specified about the conformation in the span except that it must be sterically permissible, then one may suppose the atoms to be guided in up to two ways: (i) by guide coordinates, (ii) by supposing that where a close approach occurs each atom is (for 1 cycle anyway) guided towards a point which is half the overlap distance away in a direction corresponding to separating the atoms concerned. We may then set up two sets of normal equations at each cycle corresponding to (i) and (ii), of which the second will be null if it happens that no overlap occurs. Now diagonalize the first matrix by a transformation **T**, as before, and subject the second to the *same* transformation. (This will not diagonalize it.) The second is then effectively a set of simultaneous equations for the same unknowns, the eigenshifts, as is the first transformed matrix but we use the first in conjunction with a filter, so that only some of the eigenshifts are determined by matrix 1. Those which are left undetermined by matrix 1 are those which have comparatively little influence on the fitting anyway, and might as well have their values determined by the other criterion, *i.e.* by the corresponding partition of the second normal matrix, as transformed by **T**. This *partition* should also itself be diagonalized and filtered subsequently. This procedure has the merit that, in the linear region anyway, the sets of shifts provided by each criterion are orthogonal and therefore cannot interact or oscillate. There then remain questions concerning the appropriate positioning of the filter, which now becomes in effect an interface. Such questions are best answered from ex-

perience, and these ideas have not yet been implemented.

4.2. *Weighted parameters*

In discussing filtering (§2.2.3) we pointed out that one effect of filtering is to prevent the build-up of a large elastic strain energy as represented by the quantity $\tilde{\Theta}\Theta$. This analogy may be taken further so that the quantity minimized by filtering is not $\sum_i \theta_i^2$ but $\sum_i \omega_i \theta_i^2$ and the ω_i may be varied at will to represent the differing torsion constants of spindles of different kinds, especially between bond angles and dihedral angles. These considerations are only valid, of course, with a filtered solution. If the solution is not filtered it implies that we are prepared to pay any price in terms of deformation, in order to achieve a fit, in which case it would make no difference how the parameters are weighted. As outlined here, it is meaningful to speak of strain energy only if it is supposed that the initial conformation for each cycle is an equilibrium one; nevertheless, this line of development seems germane and may find application when more is known of conformational potential functions such as those of De Santis, Giglio, Liquori & Ripamonti (1965), Dunhill & Phillips (1966), Brant & Flory (1965), and others.

The theory is as follows. We revert to the normal equations and introduce a diagonal matrix $\mathbf{W}$, the elements of which are $\sqrt{\omega_i}$, as follows:

$$\frac{\partial \tilde{\mathbf{R}}}{\partial \Theta}\mathbf{D} = \frac{\partial \tilde{\mathbf{R}}}{\partial \Theta}\frac{\partial \mathbf{R}}{\partial \Theta}\Theta$$

$$\tilde{\mathbf{W}}^{-1}\frac{\partial \tilde{\mathbf{R}}}{\partial \Theta}\mathbf{D} = \tilde{\mathbf{W}}^{-1}\frac{\partial \tilde{\mathbf{R}}}{\partial \Theta}\frac{\partial \mathbf{R}}{\partial \Theta}\mathbf{W}^{-1}\mathbf{W}\Theta$$

$$\mathbf{T}\tilde{\mathbf{W}}^{-1}\frac{\partial \tilde{\mathbf{R}}}{\partial \Theta}\mathbf{D} = \mathbf{T}\tilde{\mathbf{W}}^{-1}\frac{\partial \tilde{\mathbf{R}}}{\partial \Theta}\frac{\partial \mathbf{R}}{\partial \Theta}\mathbf{W}^{-1}\tilde{\mathbf{T}}\mathbf{T}\mathbf{W}\Theta$$

where $\mathbf{T}$ diagonalizes $\tilde{\mathbf{W}}^{-1}\frac{\partial \mathbf{R}}{\partial \Theta}\frac{\partial \mathbf{R}}{\partial \Theta}\mathbf{W}^{-1}$, and we filter, as before, giving

$$\Theta_{\mathrm{f,w}} = \mathbf{W}^{-1}\tilde{\mathbf{T}}\mathbf{Z}_{\mathrm{f}}\left(\mathbf{T}\tilde{\mathbf{W}}^{-1}\frac{\partial \tilde{\mathbf{R}}}{\partial \Theta}\frac{\partial \mathbf{R}}{\partial \Theta}\mathbf{W}^{-1}\tilde{\mathbf{T}}\right)^{-1}\mathbf{T}\tilde{\mathbf{W}}^{-1}\frac{\partial \tilde{\mathbf{R}}}{\partial \Theta}\mathbf{D}.$$

Suppose that each angular shift θ_i, is replaced by φ_i such that

$$\varphi_i = \sqrt{\omega_i}\,.\,\theta_i;$$

then the second equation above represents the normal equations for the φ_i so that the subsequent filtering minimizes $\sum \varphi_i^2$.

This also offers other possibilities. With the present methods each probe is in this sense elastically uniform, so that it tends to bend from its root to fit at the tip (*cf.* an elastic beam clamped at one end). If each ω_i is set equal to the corresponding diagonal element of $\frac{\partial \tilde{\mathbf{R}}}{\partial \Theta}\frac{\partial \mathbf{R}}{\partial \Theta}$ so that $\tilde{\mathbf{W}}^{-1}\frac{\partial \tilde{\mathbf{R}}}{\partial \Theta}\frac{\partial \mathbf{R}}{\partial \Theta}\mathbf{W}^{-1}$ has 1's on the diagonal (and is the 'correlation matrix' for the variables θ) then the tip of the probe is made elastically softer than the root end so that the growing point is in a sense more sensitive than established regions. This might serve to reduce the number of probes required.

4.3. *Real space refinement*

A scheme for the refinement of a chain structure by direct reference to the X-ray observations has already been described (Diamond, 1965), and may be referred to as reciprocal space refinement. The present technique may also be developed to provide real space refinement in which the difference vectors $\mathbf{d}$ which make up $\mathbf{D}$ will be replaced by $\left(\frac{\partial^2 \varrho}{\partial \mathbf{r}\,\partial \mathbf{r}}\right)^{-1}\frac{\partial \varrho}{\partial \mathbf{r}}$ where ϱ is the calculated electron density and the round bracket contains the nine second derivatives of ϱ at the point in question, *i.e.* the required movement of each atom will be obtained from the gradient and curvature of the electron density.

It is envisaged that the technique in its present form would be employed first and that the output values of the rotations θ would be used to specify the initial conformation to the refining program so that the latter is not called upon to search the Fourier map for possible solutions.

It is also clear that the filter level would have to be dependent on the resolution of the electron density map, since in a low resolution map only a general indication of positions is available which is sufficient only to control the dominant eigenshifts.

I should like to acknowledge the use of a subroutine known as BCHOW obtained from the IBM Share Library and written by D. W. Matula which has been used (with some modification) to diagonalize the normal matrices.

I am also grateful to Dr D. C. Phillips for permission to publish examples taken from the structure of lysozyme.

References

Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1965). *Nature, Lond.* **206**, 757.

Brant, D. A. & Flory, P. J. (1965). *J. Amer. Chem. Soc.* **87**, 2791.

Cullis, A. F., Muirhead, H., Perutz, M. F., Rossmann, M. G. & North, A. C. T. (1962). *Proc. Roy. Soc.* A. **265**, 161.

De Santis, P., Giglio, E., Liquori, A. M. & Ripamonti, A. (1965). *Nature, Lond.* **206**, 456.

Diamond, R. (1958). *Acta Cryst.* **11**, 129.

Diamond, R. (1965). *Acta Cryst.* **19**, 774.

Dunhill, P. & Phillips, D. C. (1966). To be published.

Edsall, J. T., Flory, P. J., Kendrew, J. C., Liquori, A. M., Némethy, G., Ramachandran, G. N. & Scheraga, H. A. (1966). *J. Mol. Biol.* **15**, 399.

Némethy, G. & Scheraga, H.A. (1965). *Biopolymers*, **3**, 155.

Parthasarathy, R. & Ramachandran, G. N. (1966). To be published.

Perutz, M. F. (1965). *J. Mol. Biol.* **13**, 646.

Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). *J. Mol. Biol.* **7**, 95.