Acta Cryst. (1978). A**34**, 3–11

LALS: A Linked-Atom Least-Squares Reciprocal-Space Refinement System Incorporating Stereochemical Restraints to Supplement Sparse Diffraction Data

By P. J. Campbell Smith and Struther Arnott

Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907, *USA*

(*Received* 25 *January* 1977; *accepted* 8 *June* 1977)

LALS is a computer program for the refinement of molecular structures. It is primarily intended for helical macromolecules but has applicability in other fields. The refinement uses as data, X-ray structure factors, usually from fibre diffraction studies, and/or stereochemical information, including a comprehensive short-contact search. Various other geometrical constraints and restraints may be placed on the molecular conformation. As parameters of the refinement, *LALS* uses primarily dihedral angles about single bonds, assuming bond lengths and angles to be fixed and known. This greatly reduces the number of parameters from conventional atom-position refinements, as is necessary in systems where data are sparse. *LALS* has been used successfully to investigate the structures of a large number of more or less ordered polynucleotides, polysaccharides and other fibrous materials.

1. Introduction

Many macromolecules are not susceptible to conventional crystallographic structure determination because of their inability to form regular crystals. DNA, for example, has an irregular primary structure, is invariably associated with more or less disordered water and cations, and is a long, thin molecule which displays (in contrast to enzyme molecules) a very variable tertiary structure. All of these features contribute to the biological significance of DNA, and all of them are incompatible with the properties of a regularly formed crystal.

Yet, diffraction studies have indeed played a great part in the determination of the structures of DNA and of a wide range of other biopolymers, through the application of two types of technique. Firstly, physical manipulation often allows the thread-like molecules to be extended and 'combed' so that the thread axes are parallel, and in some cases the sample then locally orders its orientations about these axes to form very small regions of three-dimensional crystallinity. The consequent fibre or film X-ray diffraction data yield information about the ordering along the axes, and in favourable cases, about the local ordering in the planes perpendicular to these axes. However, even quite well-ordered samples rarely give sufficient data for conventional crystallographic analysis: the *A* form of DNA, for example, yields around 200 measurable data but contains more than 250 crystallographically non-equivalent atoms (Arnott & Hukins, 1972). Less well-ordered materials often give little more quantitative data than the unit-cell dimensions and molecular axial symmetry. The second technique therefore is that of analysing the diffraction data in conjunction with other data or assumptions, so that the role of the former becomes that of discriminating between otherwise acceptable models.

What additional information is available to supplement the diffraction data? A linear polymer with a regular secondary structure necessarily displays some helical symmetry and an important simplification of the problem results from the assumption that the conformation is indeed regular (within the limits of the determination) and that the asymmetric unit is therefore a single chemical repeat, rather than a complete turn of the helix. *A*-DNA, for example, displays an 11-fold screw axis relating the 11 residues in each helical turn. Since this symmetry is non-crystallographic, there is implicit the assumption that intramolecular rather than intermolecular forces determine the conformation: that is, that all 11 residues adopt the same conformation despite being in different intermolecular environments.

Another important simplification can result from the assumption that bond lengths and angles in polymers have the same (or very nearly the same) values as in the corresponding monomers. In many cases the monomers are susceptible to conventional crystallographic analysis, so that the atomic positions, and hence the bond lengths and angles, are known rather precisely. This reduces the solution of the polymer structure to that of determining dihedral angles about single bonds: one parameter per atomic position rather than three. Moreover, the conformations of some structural entities such as rigid rings (*e.g.* sugar rings, planar conjugated bases) may also be assumed to persist from monomer to polymer. It is therefore possible to prepare what we term the *linked-atom* description of the molecule: that is, one in which interatomic relations are described in terms of bond lengths, bond angles and dihedral angles. The linked-atom description of *A*-DNA, for example, contains only six variables: five dihedrals along the sugar–phosphate backbone and one about the base–sugar bond. In addition to reducing the number of parameters to be determined, we can improve the data-to-parameter ratio by increasing the number of data. In the case of the *A*-DNA example this is not necessary since we now have something like 200 X-ray intensity data to six parameters, but in many cases the number of these data which are quantifiable may be small or zero. Earlier modelling studies with linked-atom or similar approaches include those of Eyring (1932), Diamond (1965) and Arnott & Wonacott (1966).

One source of additional data is provided by the values of the variable parameters in either monomers or similar polymers whose structures are known. Thus, for example, to solve a new nucleic acid structure we could survey the values of the corresponding dihedral angles in *A*-DNA and other solved structures, and require that our new structure display parameters varying minimally (in a least-squares sense) from those in our survey. While this method has some uses, particularly as a preliminary model-building step, it has the serious disadvantage of ignoring any 'new' features that may exist in our 'new' nucleic acid (unless they be purely differences in symmetry), and hence biasing the determination towards the earlier structures.

A more useful source of stereochemical data is the requirement that the new model exhibit no over-short nonbonded interatomic distances. Whilst such a requirement could most accurately be embodied in a complete minimum-free-energy calculation, such methods are at present either too time-consuming, poor approximations, or both, and it is necessary in practice to compromise. Following Williams (1969) we have found a simple quadratic function to be satisfactory: that is, by varying our parameters we minimize C in:

$$C = \sum_i k_i(s_i - d_i)^2 \qquad (1)$$

where s_i is an interatomic distance in our model which is less than the desired minimum, d_i, and k_i is a weight. The summation is over all such distances, termed *contacts*, and the number of terms will, if all goes well, decrease as the refinement proceeds. An extension of this procedure, detailed later, can be used to incorporate hydrogen-bonding and coordination-bonding information.

2. The *LALS* refinement program

One important disadvantage of a refinement scheme such as is briefly outlined above, is the difficulty of writing a general computer program which has the capability of accommodating the wide range of bonding and packing patterns found in even quite chemically restricted fields of fibre structure analysis. Many such determinations, while interesting in their relations with other similar structures, are often not important enough to warrant the expense and trouble of reprogramming. Early programs incorporating some of this flexibility include those of Diamond (1965) and Arnott & Wonacott (1966), but neither of these is easily applicable to very branched structures.

The *LALS* program described here therefore had as one design objective the ability to cope with arbitrarily complicated systems without undue loss of efficiency. A summary of the logical flow in the program is shown in Fig. 1. The main refinement (shown by the solid arrows) is an iterative procedure because of the non-linear dependence of the data on the variables. Since these variables are most commonly angles, this non-linearity can be more pronounced than in conventional atomic position refinement and can sometimes lead to predicted shifts in parameters which are wrong in both sense and magnitude in the initial stages of refinement. For this reason the updating of parameters at the end of each cycle is subjected to some scrutiny and modification, of which further details are given later.

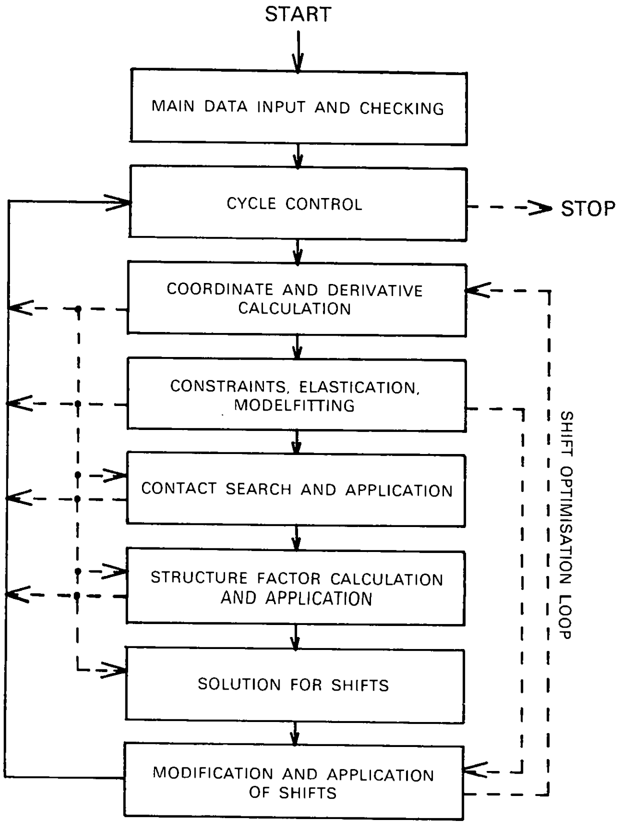
START
MAIN DATA INPUT AND CHECKING
CYCLE CONTROL
STOP
COORDINATE AND DERIVATIVE CALCULATION
CONSTRAINTS, ELASTICATION, MODELFITTING
CONTACT SEARCH AND APPLICATION
STRUCTURE FACTOR CALCULATION AND APPLICATION
SOLUTION FOR SHIFTS
MODIFICATION AND APPLICATION OF SHIFTS
SHIFT OPTIMISATION LOOP

Fig. 1. Summary of the logical flows in *LALS*. The main refinement cycle is shown by solid arrows; broken arrows represent optional pathways.

3. Constrained least-squares refinement

Although the principles of least-squares refinement are probably familiar to most conventional crystallographers, the use of Lagrange multipliers to impose constraints is not so widely known, nor are the implications of non-independent variables, and we present therefore the basis of this method here.

Given a system described by n variable parameters p_j, m observations d_i of this system, and knowledge of the system sufficient to predict values c_i corresponding to the d_i for a given set of p_j, the least-squares assertion is that the best estimate of the p_j which yields the observations is such that Ψ in (2) is a minimum.

$$\Psi \equiv \sum_{i=1}^{m} (c_i - d_i)^2. \qquad (2)$$

If the p_j are all independently variable, this implies:

$$\frac{\partial}{\partial p_j} \Psi = 0 \quad j = 1,n. \qquad (3)$$

To ease the solution of (3) we assume – and the implications of this assumption will be discussed later – that the functional dependence of the c_i on the p_j is known and linear:

$$c_i = \sum_j a_{ij}\, p_j + z_i \quad i = 1,m, \qquad (4)$$

whence $a_{ij} \equiv \partial c_i / \partial p_j$. Substituting (4) in (3), we obtain:

$$\sum_i a_{ij}(c_i - d_i) = 0 \quad j = 1,n. \qquad (5)$$

In practice, we generally make a guess at the values of the p_j and wish to refine these to the values satisfying (3). If our trial set consists of $p'_k(k = 1,n)$ with corresponding c'_i, we have from (4):

$$c_i - c'_i = \sum_k a_{ik}(p_k - p'_k) \quad i = 1,m. \qquad (6)$$

The bracketed terms in (6) are thus the required shifts in the parameters, s_k, so that the p_k which we are seeking are the sums of the p'_k and the s_k.

Solving (6) for the c_i and substituting in (5) we obtain:

$$\sum_i \sum_k a_{ij} a_{ik} s_k = \sum_i a_{ij}(d_i - c'_i) \quad j = 1,n. \qquad (7)$$

Equations (7) are n linear equations in the n unknowns s_k and may be solved by the usual methods.

Commonly, however, the p_j are not independently variable, but subject to certain constraints on the sets of values they may take. If these constraints, r in number, can be expressed (or approximated) by equalities such as the following, they can be incorporated into the least-squares process:

$$g_q \equiv \sum_j b_{qj}\, p_j + y_q \equiv 0 \quad q = 1,r, \qquad (8)$$

whence $b_{qj} \equiv \partial g_q / \partial p_j$.

The condition that Ψ in (2) be a minimum now no longer implies (3), but only the lesser condition:

$$\sum_j \frac{\partial \Psi}{\partial p_j} \mathrm{d}p_j = 0. \qquad (9)$$

In (9) the $\mathrm{d}p_j$ are not independent but must satisfy the constraint relations, derived by differentiating (8):

$$\sum_j b_{qj} \mathrm{d}p_j = 0 \quad q = 1,r. \qquad (10)$$

There are therefore only $n - r$ independent $\mathrm{d}p_j$. We can, however, multiply each equation (10) by a constant λ_q and add these to (9) to obtain:

$$\sum_j \left[\left(\frac{\partial \Psi}{\partial p_j} + \sum \lambda_q b_{qj} \right) \mathrm{d}p_j \right] = 0. \qquad (11)$$

We choose values for the r constraints λ_q, which are Lagrange undetermined multipliers, such that the coefficients of r of the $\mathrm{d}p_j$ are zero. When we do this, however, the remaining $\mathrm{d}p_j$, numbering $n - r$, are independent and therefore, analogously to (3), we may conclude that for all j:

$$\frac{\partial \Psi}{\partial p_j} + \sum_q \lambda_q b_{qj} = 0 \quad j = 1,n. \qquad (12)$$

Analogous to the unconstrained situation we thus have [*cf.* equations (7)]:

$$\sum_i \sum_k a_{ij} a_{ik} s_k + \sum_q \lambda_q b_{qj} = \sum_i a_{ij}(d_i - c'_i) \quad j = 1,n. \qquad (13)$$

Also, our original constraint expressions in (8) can be evaluated as g'_q at p'_j, yielding by subtraction from (8):

$$-\sum_k b_{qk} s_k = g'_q \quad q = 1,r. \qquad (14)$$

Together, (13) and (14) are $(n + r)$ linear equations in the $(n + r)$ variables $s_1 \ldots s_n$ and $\lambda_1 \ldots \lambda_r$ which are satisfied by a single set of values for the variables if the equations are all linearly independent. The solution of the equations then yields shifts s_k which when added to the p'_k give p_k, which [from (12)] are the values which minimize Ω:

$$\Omega \equiv \Psi + \sum_q \lambda_q g_q \equiv \sum_i (c_i - d_i)^2 + \sum_q \lambda_q g_q. \qquad (15)$$

To solve (13) and (14) we may restate them in matrix notation as:

$$\begin{bmatrix} \mathbf{L} & \mathbf{G} \\ \tilde{\mathbf{G}} & \mathbf{0} \end{bmatrix} . \begin{bmatrix} \mathbf{s} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{h} \\ \mathbf{t} \end{bmatrix} \qquad (16)$$

where the element L_{uv} is $\sum_i a_{iu} a_{iv}$ (n rows and columns), of s_v is s_k (n elements), of h_u is $\sum_i a_{iu}(d_i - c'_i)$ (n elements), of G_{uv} is $-b_{qk}$ (n rows, r columns), of λ_v is λ_q (r elements), of t_v is g'_q (r elements), and where $\tilde{\mathbf{G}}$ is the transpose of $\mathbf{G}$, and $\mathbf{0}$ is the zero matrix. Thus the formal solution involves the inversion of the left-hand-side matrix and the premultiplication of the right-hand-side vector by this inverse to give the vector of s and λ.

It is convenient, however, to permit the sets of equations (13) and (14) on occasion not to be linearly independent, that is, for there to be redundancies amongst either the variables or the constraints (of which instances will be presented later). In this case, the left-hand-side matrix will be singular, or, as invariably occurs in practice, nearly singular. To allow for this situation we first determine the eigenvalues and normalized orthogonal eigenvectors of the matrix. Restating (16), we have (17) and the eigenvalues and eigenvectors by definition satisfy (18) and (19):

$$\mathbf{A}\cdot\mathbf{p}=\mathbf{b} \tag{17}$$

$$\mathbf{A}\,.\,\mathbf{v}_i=\beta_i\mathbf{V}_i \tag{18}$$

whence

$$\mathbf{A}\,.\,\tilde{\mathbf{V}}\,.\,\mathbf{B}=\tilde{\mathbf{V}}$$

$$\mathbf{v}_i\,.\,\mathbf{v}_j=\delta_{ij} \tag{19}$$

whence

$$\mathbf{V}\,.\,\tilde{\mathbf{V}}=\mathbf{I}.$$

Here $\mathbf{v}_i$ are the $(n+r)$ eigenvectors of $\mathbf{A}$, β_i the eigenvalues, $\mathbf{V}$ is the matrix whose rows are the $\mathbf{v}_i$, $\mathbf{B}$ the diagonal matrix whose ith diagonal element is $1/\beta_i$, δ_{ij} is the Kronecker delta and $\mathbf{I}$ the unit matrix.

The singularities will now be manifest as very small eigenvalues (relative to the others) and corresponding eigenvectors which are ill-defined and prone to greatly magnified error components. However, provided our singularities arise from consistent redundancies, as they will in a physically meaningful system, we may validly set these eigenvalues to zero, since the eigenvectors correspond only to the differences between physically identical parameters. Using this modified set of eigenvalues and eigenvectors we then form the inverse of $\mathbf{A}$ using (20), which follows from (18) and (19) (see also Bickley & Thomson, 1964):

$$\mathbf{A}^{-1}=\tilde{\mathbf{V}}\,.\,\mathbf{B}\,.\,\mathbf{V}. \tag{20}$$

We may then solve for the parameter shifts and (incidentally) the constraint multipliers using:

$$\mathbf{p}=\mathbf{A}^{-1}\,.\,\mathbf{b}. \tag{21}$$

In the above discussion certain approximations have been made which are worthy of comment. Firstly, it was assumed that the c_i, the calculated values corresponding to the observations, are dependent linearly on the p_j. Generally this is not so. However, commonly the first derivatives of the c_i with respect to the p_j are of the same sign and the same order of magnitude over the range of parameter values between the starting and refined model, and almost invariably these derivatives approach their 'best' values as the parameters approach those of the 'best' model. Thus, whilst the calculated shifts will not be exactly correct, they will lead to a 'better' model which can then be iteratively refined until the shifts are deemed insignificant. Exactly the same argument applies to non-linear constraint expressions, and in practice the two converge to a stable solution together.

Lastly, it is often the case that the observations are not equally reliable and have quantifiable variances. In this case each term in the summation in (2) may be multiplied by a weight, w_i, proportional to the inverse of the variance d_i, and hence the elements of $\mathbf{L}$ and $\mathbf{r}$ in (16) will include w_i in their summations.

4. Generation of atomic coordinates

LALS is designed to be used both in a 'crystallographic' context, in which both intramolecular and lattice-packing relations are considered, and an 'isolated entity' situation where the packing details are either undetermined (as occurs in various disordered packing arrangements) or irrelevant (as, for example, in studies of local conformation with purely stereochemical data). For this reason there is some redundancy in the available parameters for positioning and orienting molecules, the appropriate parameters being chosen according to context. *LALS* also is written primarily for use with helical molecules, and although this is reflected in the nomenclature, the program has been used with success on systems with no helical symmetry.

Structures are described in the linked-atom method by relating each atom to those already defined by a distance, an angle, and a dihedral angle. The last two are refinable and normally can be chosen to correspond to actual bond angles and dihedrals about chemical bonds. Each linked-atom structure is then positioned relative to its local helix axis by another four refinable parameters.

In the case of a crystallographic system, where intermolecular relations are important, another four refinable parameters are provided for each independent structure in the unit cell.

In practice, this scheme has proved to be worth its seeming complexity, in that formal separation of molecular and lattice parameters allows the two to be refined either separately or jointly.

The computational details of these calculations are presented in the technical report available from the authors.

5. Constraints and restraints

In a least-squares refinement such as *LALS*, there commonly exist known relations between the varied parameters which must be satisfied in any refined model.

Some of these, which we term constraints, are inflexible requirements, that is, exact relations which must hold exactly for any valid solution. For example, the set of torsion angles along the backbone of one residue of an *A*-DNA helix must be such that the 'top' of the residue is continuous with the 'bottom' of the next. One way to treat this would be to find a (lesser) set of truly independent variables, but this would negate the physical significance of the parameters, which is the essence of *LALS*. We therefore formulate these constraints as linear or approximately linear functions of the parameters which are to be made zero, and apply them as already detailed.

LALS has the built-in capability of applying the following types of constraint.

(i) Coincidence of two 'atoms', of which either may be transformed with local or crystallographic symmetry; used, for example, to maintain helix continuity and to close chemical ring systems.

(ii) Direct linear relations between varied parameters; for example, the maintenance of a constant difference between two torsion angles about the same bond.

(iii) Constraint of the distance between any two atoms.

(iv) Constraint of the angle described by any three atoms.

(v) Constraint of the dihedral angle described by any four atoms.

(vi) Collinearity of any three atoms; used, for example, to control the bending of hydrogen bonds.

(vii) Linear relations between atomic Cartesian or cylindrical coordinates; used for example to vary helix pitch or turn angle separately.

(viii) Constraint of the inclination of the plane described by any three atoms, used for example in polynucleotide work to control the base-plane slope (Arnott, Chandrasekaran & Selsing, 1975).

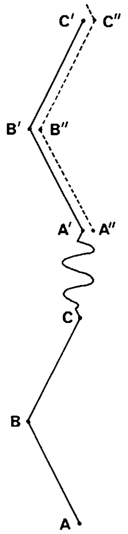

Fig. 2. *A'B'C'* is a triangle of dummy atoms in the structure shown. *A''B''C''* is a result of applying helix symmetry to *ABC*. The necessary constraints to maintain helix continuity are coincidence constraints between *A'* and *A''*, *B'* and *B''*, and *C'* and *C''*.

Some of these imply more than one constraint; for example, the first type requires three constraints on the differences between the *x*, *y* and *z* coordinates of the atoms concerned.

The commonest use of constraints in *LALS* is to make a helix backbone continuous or to close a ring. In the helix-backbone case, the structure is defined as the repeat unit plus three extra atoms of the following repeat (which are subsequently excluded from structure-factor calculations *etc.*). Coincidence constraints are then placed on the three pairs of atoms which comprise one of these extra atoms transformed by one helix repeat operation and the corresponding atom in the original structure (Fig. 2). Three atoms are used to preserve full control over all the bond angles and torsion angles in the overlapped region.

Notice that of these nine constraints, three are redundant (if the included angle is not varied) and cause singularity in the least-squares normal matrix. Whilst it is possible to select six of the nine which are linearly independent, the solution obtained is discrete but not unique and may not correspond to the fulfillment of the three unapplied constraints.

Constraints imply an exact parametric relation. Often we do not wish to impose such a drastic limitation but to include information about the expected value of some function of the system, perhaps with an associated probable deviation. This can best be done by including such terms in the least-squares summation: that is, considering them as data to be fitted to. We term these restraints, and have included in *LALS* the facility for restraints analogous to the constraints listed as (iii), (iv), (v) and (viii) above. By far the commonest use for this is to maintain hydrogen bonds close to known or ideal lengths.

An important type of restraint is that imposed on a single parameter, which we term elastic binding. Here we can build models whose conformation angles differ minimally from averages found from surveys of known structures. This is often an important preliminary step in constructing models. It can easily be shown that this type of data adds only to the diagonal of the normal matrix and for this reason it has been termed matrix augmentation. For parameters which have no 'preferred' value, binding to their current values has been found useful in preventing overshifting in poorly constrained situations: this adds to the normal-matrix diagonal but not to the right-hand-side vector.

Another available restraint used in modelbuilding studies is fitting to a known model by distance restraints between corresponding atoms in the studied and the known model. This allows the determination of a model with standard bond lengths and angles differing minimally from a given model, a process we term modelfitting. Modelfitting was used to generate the standard DNA and RNA coordinates published by Arnott, Smith & Chandrasekaran (1976).

6. Non-bonded contacts

We have already noted the desirability of automatically including distance restraints between atoms which are not bonded to each other and are too close.

The practical problems involved in this are many. Biopolymers – such as *A*-DNA – often have high molecular symmetry so that the effective asymmetric unit contains quite a small number of atoms. But each of the 11 residues of the *A*-DNA helix is in a different environment and each therefore makes different intermolecular contacts. Of course, the very fact that a close to regular 11-fold helix is observed indicates that intrachain rather than interchain interactions are the dominant conformation-determining forces. Nevertheless, the packing of chains and verification of the previous statement require that we make the investigation, and indeed in many other cases we can see that the packing symmetry is reflected in the molecular conformation [*e.g.* in hyaluronic acid (Guss *et al.*, 1975; Winter, Smith & Arnott, 1975)].

The first, and indeed the overriding, problem is then that, *a priori*, there are a very large number of interatomic distances to be tested. Order-of-magnitude estimates are sufficient to show that this number is in the 10^6 to 10^{10} range for many biopolymers, which makes the calculation prohibitively expensive.

A second problem is that we must exclude from consideration interactions between atoms bonded to each other or to a common third atom. Although these distances are not normally functions of the varied parameters, the magnitude of their contributions to Ψ is such that the important interactions are lost by numerical truncation.

LALS deals with these problems by a variety of sorting and optimizing procedures so that, typically, the number of interatomic distances calculated is of the order of 10^5, yielding some hundreds of 'useful' contacts after those unaffected by the varied parameters are eliminated. Bonding patterns are stored in a series of indexed look-up tables which reduce to an almost negligible level the time taken in checking for bonds and bonds to common atoms.

We therefore end up with a list of over-short interatomic distances which are included in our least-squares optimization (1). The cutoff distances d_i and the weights k_i may be derived by various methods and the final results of refinements with different sets are not greatly dependent on the exact numbers used. In Table 1 we list a set which has been found satisfactory. These parameters were derived by setting d_i equal to the sum of the van der Waals radii of the interacting atoms plus 0·02 nm and choosing k_i such that the interaction function best fits a conventional Buckingham non-bonded potential function over the range from the minimum energy distance down to the distance at which the energy exceeds the minimum energy by 2·5 kJ mol^{-1} (approximately thermal energy). The radii and potential function parameters are taken from the work of Chandrasekaran & Balasubramanian (1969) and Lakshminarayanan & Sasisekharan (1969).

Table 1. *Parameters for contacts*

Interaction	d_i (nm)	k_i (nm^{-2})
H–H	0·260	104
C–C	0·360	103
C–H	0·310	99
N–N	0·330	171
N–C	0·345	131
N–H	0·295	129
O–O	0·324	133
O–N	0·327	148
O–C	0·342	117
O–H	0·292	114
P–P	0·380	562
P–O	0·352	257
P–N	0·355	342
P–C	0·370	260
P–H	0·320	208
S–O	0·352	138
S–N	0·355	156
S–C	0·370	124
S–H	0·320	119

Certain short contacts may of course correspond to known non-bonded attractive interactions, particularly hydrogen bonds and coordination bonds to metal atoms. To cope with these, we have included a crude but useful modification to our interaction function. In this, contacts lying between two inner limits are elastically bound to an ideal value within these limits with a higher weight. The effective implied 'energy

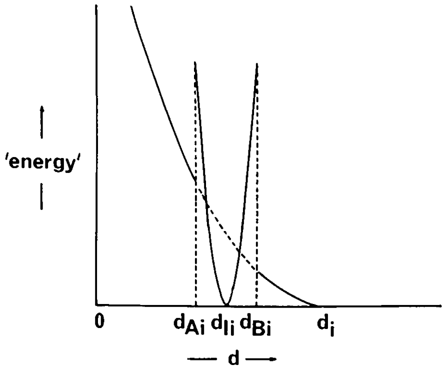

Fig. 3. The effective 'energy' function for attractive interactions (*e.g.* hydrogen bonds) is shown by the solid line. Interatomic separation d_i is the usual cutoff distance, d_{Ai} and d_{Bi} are the inner limits between which the attractive interaction occurs, d_{Ii} being the ideal separation. Notice that only the slope of the function has any meaning.

function' is shown schematically in Fig. 3. Although this may at first appear quite unrealistic, it should be realized that only the slope of the curve has any effect. Values for these inner limits and the weight are generally chosen to provide the desired degree of fit in a given problem. Use of limits 0·02 nm less and greater than the desired distance and a weight of 1000 nm^{-2} has proved satisfactory in some practical applications.

7. Structure factors

The structure-factor calculation in *LALS* is quite conventional in most respects and need not be described in great detail.

Two methods of calculating structure factors are provided, one using the usual trigonometric functions and the other using Bessel functions (Cochran, Crick & Vand, 1952). The latter has the advantage of separating molecular and crystal symmetry so that for situations where high molecular symmetry exists it is much faster.

The Bessel routine also calculates (but currently does not refine against) continuous intensity along layer lines which often exist in fibre diffraction patterns; it is also able to separate Bragg and continuous contributions on the basis of various idealized patterns of disorder (Arnott, 1973).

Scattering factors are included, calculated from the analytic approximation tabulated by Cromer & Waber (1974). Provision is made for including bonded hydrogen atoms with their heavy atom as a single scatterer and for using water-weighted scattering factors: that is, those relating to the scattering of atoms surrounded by a medium with a uniform electron density equal to the mean electron density of a water molecule. These have been successfully used for many structures in which relatively unstructured water fills much of the unit-cell voids, such as is the case for most nucleic acids under physiologically appropriate conditions of salt concentration and humidity. This method was introduced by Langridge, Wilson, Hooper, Wilkins & Hamilton (1960), and modified and detailed by Fuller (1961) and Arnott & Hukins (1973). The latter two, however, contain an error corrected here.

Water-corrected scattering factors, $g(\rho)$ where ρ is the reciprocal-space radius of a reflection, are related to normal scattering factors $f(\rho)$ according to the principle of Babinet, by:

$$g(\rho) = f(\rho) - V\sigma\varphi(\rho). \tag{24}$$

V is the volume of the scatterer, σ is the electron density of water (298·4 nm^{-3}) and $\varphi(\rho)$ is the scattering factor of a uniformly dense one-electron sphere of radius R, given by (25) (James, 1965).

$$\varphi(\rho) = 3\{[\sin(2\pi R\rho) - (2\pi R\rho)\cos(2\pi R\rho)]/(2\pi R\rho)^3\}. \tag{25}$$

To calculate V and φ we assume van der Waals radii R as follows: H 0·12, C 0·17, N 0·15, O 0·14, P 0·19, S 0·17 nm.

For atoms with n attached hydrogen atoms not separately included, the uncorrected scattering factors are taken as the sum of the heavy atom $f(\rho)$ and n times the hydrogen $f(\rho)$. To correct these combined scattering factors for water we modify V in (24) to include the volume of the hydrogen atoms. Because the spheres of the hydrogen atoms are partially embedded in that of the central atom, the added volume can be shown to be given by (26).

$$\begin{aligned} v_{\mathrm{H}} = {} & \tfrac{1}{3}\pi r_{\mathrm{H}}^3[2 + \cos\beta(\sin^2\beta + 2)] \\ & - \tfrac{1}{3}\pi r_A^3[2 - \cos\alpha(\sin^2\alpha + 2)] \end{aligned} \tag{26}$$

where $\alpha = \arccos[(r_A^2 + b^2 - r_{\mathrm{H}}^2)/2r_A b]$, $\beta = \arccos[(r_{\mathrm{H}}^2 + b^2 - r_A^2)/2r_{\mathrm{H}} b]$. The radius of the heavy atom is r_A, of hydrogen r_{H} (values as given above), b is the bond length (taken as C 0·109, N 0·101, O, 0·096 nm).

Thus in (24) we add nv_{H} to V and to calculate φ use an R corresponding to a sphere of volume V.

The terms included in (2) for the X-ray data are $[sF_o - F_c\exp(-b\rho^2/4)]$, where F_o is an observed structure amplitude on an arbitrary scale, s is a refinable scale factor, F_c is the corresponding calculated amplitude, and b is a refinable isotropic attenuation factor. In the case of reflections which overlap (a common problem in fibre diffraction) F_c is taken as the square root of the summed calculated intensities of the contributing reflections.

8. Solution of the normal equations

As has been already noted, the normal matrix is commonly singular and is therefore as a first step decomposed to its eigenvalues and eigenvectors. The method of Smith, Boyle, Garbow, Ikebe, Klema & Moler (1974) embodied in the *EISPACK* package is used in *LALS* and has proved both reliable and economical.

The eigenvalues are then scanned for redundancies, expressed as values some orders of magnitude less than the next larger ones, and these are eliminated. The matrix is then inverted and shifts in the parameters are calculated.

Because of non-linearity, it is generally dangerous to use very large shifts: typically we find about 20° in angular variables is a useful maximum. For this reason several modifications may be made to the shifts before they are applied. These are scaling to a maximum, damping, and constraint optimization.

Scaling involves reducing each shift by a constant factor so that the largest scaled shift is some specified figure, such as 20°. If all shifts are already less than this, no scaling is done. In practice, two maxima are used, one for angular and one for translational parameters. Scaling is a crude method, but it is fast and works well in controlling the early stages of refinement.

Damping is the multiplication of every shift by a fixed factor less than unity. It is occasionally useful when the refinement oscillates.

Constraint optimization implies a further scaling of all the shifts by a constant factor (possibly greater than unity) such that the solution best satisfies the Lagrange constraints. In practice, this is achieved by running trial solutions using the calculated shifts and twice the calculated shifts as far as the constraint calculation, fitting a quadratic to the three (zero, one and two times shifts) points, and taking the minimum of the quadratic as the best estimate of the shift scaling factor. This procedure proves worthwhile in poorly constrained situations, but has the disadvantage in well constrained systems of demanding very small shifts.

The modified shifts are added to the original parameter values and, subject to user control, refinement continues iteratively with the coordinate calculation.

9. Uncertainties

No refinement system can be considered complete without some indication of the precision of its results.

The theory of least-squares shows that the standard deviation, σ_j, of the refined (to convergence) parameter p_j is given by (27).

$$\sigma_j^2 = m_{jj} \qquad (27)$$

where m_{jj} is the appropriate diagonal element of the inverted normal matrix [$\mathbf{A}^{-1}$ in (20)]. This depends, as indeed does the validity of least-squares analysis, on two conditions: firstly, that the weighted errors of the data must be random (*i.e.* a finite subset of a normally distributed population), and secondly that each datum must be weighted by the reciprocal of its estimated variance, so that the mean value of the weighted squared differences between observed and calculated values is unity.

If the second condition is not met, that is, if the weights are proportional but not equal to the reciprocals of the variances, we can estimate the best value of the constant of proportionality as $1/v$ (28).

$$v = \sum w_i(c_i - d_i)^2/N \qquad (28)$$

where w_i are the applied weights to data d_i, c_i being the calculated value at convergence, and N is the excess of data over net varied parameters. Combining these results we obtain:

$$\sigma_j = (m_{jj}v)^{1/2}. \qquad (29)$$

Unfortunately, perhaps, this relies on the random-error condition and in doing so prevents us examining the value of v as a test of this condition being satisfied.

The above treatment also relies on a knowledge of N. Because of the contact analysis, it is far from trivial to assign a value to N. Obviously, the fact that some contact does not exist, or has been relieved, is a point in favour of the final model, yet assigning a value to N in the usual way increases the estimated error as fewer contacts remain. We have found no completely satisfactory solution to this problem, but have used as a pragmatic estimate the larger of the number of unrelieved contacts and the number of atoms in the asymmetric unit.

We therefore can in *LALS* get estimates of the uncertainties of our refined parameters. However, the more useful uncertainties are those of atomic positions. Because the effects on atomic positions of varying different parameters are correlated, it is necessary to use not only the parameter variances but also the pairwise covariances, which are derived from the inverted normal matrix in an analogous way. It is of note that the volume of calculation resulting is considerable.

10. Applications

It is beyond the scope of the present paper to list all past and possible applications of *LALS*. Since, however, many published instances of its use include useful practical details, we indicate some of them here.

Nucleic acid and polynucleotide structures have been extensively studied in this laboratory. X-ray data alone were used in studies by Arnott & Selsing (1974), Arnott, Chandrasekaran, Hukins, Smith & Watts (1974) and Selsing, Arnott & Ratliff (1975) on unusual DNA double helices, while a combination of modelling and comparison with continuous transform was employed for polyinosinic acid (Arnott, Chandrasekaran & Marttila, 1974). Several structures for which X-ray data were scarce or absent were investigated largely on the basis of their contact properties; these include *C*-DNA (Arnott & Selsing, 1975) and several triple-stranded polynucleotides (Arnott, Bond, Selsing & Smith, 1976). The structure of polycytidylic acid, a single-stranded polymer, was solved using a joint X-ray and contact refinement by Arnott, Chandrasekaran & Leslie (1976). Several non-helical model nucleic acid systems have also been investigated with *LALS*, such as Alden & Arnott's (1975) study of drug-molecule interaction in *B*-DNA.

In the polysaccharide field, *LALS* has helped the elucidation of several glycosaminoglycan structures and of some plant polysaccharides. X-ray data alone were used for keratan sulfate (Arnott, Guss, Hukins, Dea & Rees, 1974), ι-carrageenan (Arnott, Scott, Rees & McNab, 1974) and agarose (Arnott, Fulmer, Scott, Dea, Moorhouse & Rees, 1974). Hyaluronic acid has been extensively studied with joint X-ray and contact refinements (Guss *et al.*, 1975; Winter *et al.*, 1975;

Winter & Arnott, 1977), which have permitted the localization of cations and water molecules. A similar study on a bacterial polypentasaccharide has also been performed (Moorhouse, Winter, Arnott & Bayer, 1977).

Other applications have included the synthetic polymer poly(tetramethylene terephthalate) (Hall & Pass, 1976) and the use of modelfitting for the fitting of standard dimensions to measured coordinates for the coenzyme nicotinamide adenine dinucleotide bound to lactate dehydrogenase (White, 1976).

11. Program specifications

LALS is an overlaid Fortran program of about 7000 statements. It is coded specifically for a local variant of the CDC RUN compiler, but portability has been considered in its construction and no great difficulty is anticipated in adapting it for IBM 370 or similar equipment.

Performance figures are very problem dependent, but typical figures on our CDC 6500 computer under the Purdue MACE operating system are from 10 to 100 processor seconds per refinement cycle, with use of the order of 100 000 octal words of storage.

Further technical details and information on distribution of the program are available from the authors.

The *LALS* system has been developed over several years in close conjunction with many colleagues in this and other laboratories, to all of whom we express our thanks for their suggestions and comments.

Development of *LALS* was supported by grants from the National Science Foundation and the National Institutes of Health of the US.

References

Alden, C. J. & Arnott, S. (1975). *Nucleic Acids Res.* **10**, 1701–1717.

Arnott, S. (1973). *Trans. Am. Crystallogr. Assoc.* **9**, 31–56.

Arnott, S., Bond, P. J., Selsing, E. & Smith, P. J. C. (1976). *Nucleic Acids Res.* **3**, 2459–2470.

Arnott, S., Chandrasekaran, R., Hukins, D. W. L., Smith, P. J. C. & Watts, L. (1974). *J. Mol. Biol.* **88**, 523–533.

Arnott, S., Chandrasekaran, R. & Leslie, A. G. W. (1976). *J. Mol. Biol.* **106**, 735–748.

Arnott, S., Chandrasekaran, R. & Marttila, C. M. (1974). *Biochem. J.* **141**, 537–543.

Arnott, S., Chandrasekaran, R. & Selsing, E. (1975). *Structure and Conformation of Nucleic Acids and Protein–Nucleic Acid Interactions*, edited by M. Sundaralingam & S. T. Rao, pp. 577–595. Baltimore, Maryland: Univ. Park Press.

Arnott, S., Fulmer, A., Scott, W. E., Dea, I. C. M., Moorhouse, R. & Rees, D. A. (1974). *J. Mol. Biol.* **90**, 269–284.

Arnott, S., Guss, J. M., Hukins, D. W. L., Dea, I. C. M. Rees, D. A. (1974). *J. Mol. Biol.* **88**, 175–184.

Arnott, S. & Hukins, D. W. L. (1972). *Biochem. Biophys. Res. Commun.* **47**, 1504–1509.

Arnott, S. & Hukins, D. W. L. (1973). *J. Mol. Biol.* **81**, 93–105.

Arnott, S., Scott, W. E., Rees, D. A. & McNab, C. G. A. (1974). *J. Mol. Biol.* **90**, 253–267.

Arnott, S. & Selsing, E. (1974). *J. Mol. Biol.* **88**, 509–521.

Arnott, S. & Selsing, E. (1975). *J. Mol. Biol.* **98**, 265–269.

Arnott, S., Smith, P. J. C. & Chandrasekaran, R. (1976). *Handbook of Biochemistry and Molecular Biology*, edited by G. D. Fasman, *Nucleic Acids*, Vol. II, 3rd ed., pp. 411–422. Cleveland, Ohio: CRC Press.

Arnott, S. & Wonacott, A. J. (1966). *Polymer*, **7**, 157–166.

Bickley, W. G. & Thomson, R. S. H. G. (1964). *Matrices, their Meaning and Manipulation*, pp. 148–149. London: EUP.

Chandrasekaran, R. & Balasubramanian, R. (1969). *Biochim. Biophys. Acta*, **188**, 1–9.

Cochran, W., Crick, F. H. C. & Vand, V. (1952). *Acta Cryst.* **5**, 581–586.

Cromer, D. T. & Waber, J. T. (1974). *International Tables for X-ray Crystallography*, Vol. IV, pp. 71, 99–101. Birmingham: Kynoch Press.

Diamond, R. (1965). *Acta Cryst.* **19**, 774–789.

Eyring, H. (1932). *Phys. Rev.* **39**, 746–748.

Fuller, W. (1961). PhD Thesis, Univ. of London King's College.

Guss, J. M., Hukins, D. W. L., Smith, P. J. C., Winter, W. T., Arnott, S., Moorhouse, R. & Rees, D. A. (1975). *J. Mol. Biol.* **95**, 359–384.

Hall, I. H. & Pass, M. G. (1976). *Polymer*, **17**, 807–816.

James, R. W. (1965). *The Optical Principles of the Diffraction of X-rays*, p. 467. Ithaca: Cornell Univ. Press.

Lakshminarayanan, A. V. & Sasisekharan, V. (1969). *Biopolymers*, **8**, 475–488.

Langridge, R., Wilson, H. R., Hooper, C. W., Wilkins, M. H. F. & Hamilton, L. D. (1960). *J. Mol. Biol.* **2**, 38–64.

Moorhouse, R., Winter, W. T., Arnott, S. & Bayer, M. E. (1977). *J. Mol. Biol.* **109**, 373–391.

Selsing, E., Arnott, S. & Ratliff, R. L. (1975). *J. Mol. Biol.* **98**, 243–248.

Smith, B. T., Boyle, J. M., Garbow, B. S., Ikebe, Y., Klema, V. C. & Moler, C. B. (1974). *Lecture Notes in Computer Science*, Vol. 6, *Matrix Eigensystem Routines – EISPACK Guide*. Berlin: Springer.

White, J. L. (1976). PhD. Thesis, Purdue Univ.

Williams, D. E. (1969). *Acta Cryst.* A**25**, 464–470.

Winter, W. T. & Arnott, S. (1977). *J. Mol. Biol.* To be published.

Winter, W. T., Smith, P. J. C. & Arnott, S. (1975). *J. Mol. Biol.* **99**, 219–235.