Distance Geometry: Theory, Algorithms, and Chemical Applications

Timothy F. Havel

Harvard Medical School, Boston, MA, USA

Abbreviation

RMSD = root-mean-square coordinate difference.

1 INTRODUCTION

Distance geometry is the mathematical basis for a geometric theory of molecular conformation.[1] This theory plays a role in conformational analysis analogous to that played in statistical mechanics by a hard-sphere fluid ... which can in fact be regarded as the distance geometry description of a monoatomic fluid. More generally, a distance geometry description of a molecular system consists of a list of distance and chirality constraints. These are, respectively, lower and upper bounds on the distances between pairs of atoms, and the chirality of its rigid quadruples of atoms (i.e., *R* or *S* relative to some given order). The distance geometry approach is predicated on the assumption that it is possible to adequately define the set of all possible (i.e., significantly populated) conformations, or conformation space, of just about any nonrigid molecular system by means of such purely geometric constraints. By Occam's razor, we contend that any properties of the system that can be explained by such a simple model should be explained that way.

Distance geometry also plays an important role in the development of computational methods for analyzing distance geometry descriptions. The goal of these calculations is to determine the global properties of the entire conformation space, as opposed to the local properties of its individual members. This is done by deriving new geometric facts about the system from those given explicitly by the distance and chirality constraints, a process known more generally as geometric reasoning. Although numerous constraints can be derived from knowledge of the molecular formula, in many cases (e.g., globular proteins) additional noncovalent constraints are needed in order to define precisely the accessible conformation space. These must be obtained from additional experiments, and thus one of the best-known applications of distance geometry is the determination of molecular conformation from experimental data, most notably NMR spectroscopy. Other important applications include enumerating the conformation spaces of small molecules, ligand docking and pharmacophore mapping in drug design, and the homology modeling of protein structure.

2 THEORY

One of the most significant developments in distance geometry over the last few years has been the realization that the underlying theory is actually a special case of a more general theory, known as geometric algebra. This more general theory is certain to find manifold applications in computational chemistry, not only in the analysis of simple geometric models of molecular structure, but also in more complete classical and even quantum mechanical models. For these reasons we shall begin with a brief introduction to the geometric algebra of a three-dimensional Euclidean vector space; a more detailed account, including its applications in classical mechanics, may be found in Ref. 2.

2.1 Geometric Algebra

Although its origins can be traced back over 150 years to the work of the German schoolmaster Hermann Grassmann,[3] geometric algebra remains virtually unknown outside of a few specialized branches of mathematics and theoretical physics (where it is more commonly known as Clifford algebra, after one of its earliest developers). The three-dimensional Euclidean case is nevertheless not very difficult, and is all that is needed in order to begin to appreciate its utility. In reading the following introduction, it should be kept in mind that geometric algebra simultaneously generalizes, and hence unifies, most of the algebraic structures with which the reader is probably already familiar, including the real and complex numbers, vector algebra, and Hamilton's quaternions. It is not a substitute for linear algebra or differential calculus, but rather enriches them with new geometric content.

2.1.1 The Rules of the Game

The geometric product of vectors **a**, **b**, and **c** is an associative product, distributive with respect to vector addition, such that the square of any vector is the same as its length squared:

$$(\mathbf{ab})\mathbf{c} = \mathbf{a}(\mathbf{bc}), \quad \mathbf{a}(\mathbf{b}+\mathbf{c}) = \mathbf{ab}+\mathbf{ac},$$
$$(\mathbf{a}+\mathbf{b})\mathbf{c} = \mathbf{ac}+\mathbf{bc}, \quad \text{and} \quad \mathbf{a}^2 = \|\mathbf{a}\|^2 \tag{1}$$

Note we have not required this multiplication to be commutative, i.e., $\mathbf{ab} \neq \mathbf{ba}$ in general.

One immediate consequence of this definition is that every nonzero vector **a** has an inverse, namely

$$\mathbf{a}^{-1} = \mathbf{a}/\mathbf{a}^2 \tag{2}$$

Moreover, by the law of cosines, the usual inner product of vectors is

$$\begin{aligned}\mathbf{a}\bullet\mathbf{b} &= \tfrac{1}{2}(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a}-\mathbf{b}\|^2) \\ &= \tfrac{1}{2}(\mathbf{a}^2 + \mathbf{b}^2 - (\mathbf{a}-\mathbf{b})^2) \\ &= \tfrac{1}{2}(\mathbf{ab}+\mathbf{ba})\end{aligned} \tag{3}$$

which is just the symmetric part of their geometric product **ab**. Thus it is natural to define the outer product of two vectors as the antisymmetric part of their geometric product, i.e.,

$$\mathbf{a} \wedge \mathbf{b} = \tfrac{1}{2}(\mathbf{ab} - \mathbf{ba}) \tag{4}$$

This outer product is clearly anticommutative, meaning $\mathbf{a} \wedge \mathbf{b} = -\mathbf{b} \wedge \mathbf{a}$, and vanishes if and only if $\mathbf{a} = \alpha\mathbf{b}$ for some scalar α. The outer product $\mathbf{a} \wedge \mathbf{b}$ cannot be a vector, since inversion in the origin (i.e., multiplying all vectors by -1) does not change it; neither can it be a scalar, since then the geometric product $\mathbf{ab} = \mathbf{a}\bullet\mathbf{b} + \mathbf{a} \wedge \mathbf{b}$ would also be a scalar (which for linearly independent vectors **a** and **c** would contradict the above associativity condition). We can only conclude that the outer product of two vectors is a new entity, called a bivector.

Given an orthonormal basis $\mathbf{e}_1, \mathbf{e}_2$ of $\mathbf{e}_3$ our vector space, we can expand any outer product using anticommutivity as follows:

$$\begin{aligned}\mathbf{a} \wedge \mathbf{b} &= (a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + a_3\mathbf{e}_3) \wedge (b_1\mathbf{e}_1 + b_2\mathbf{e}_2 + b_3\mathbf{e}_3) \\ &= (a_1b_2 - b_1a_2)\mathbf{e}_1\mathbf{e}_2 + (a_3b_1 - b_3a_1)\mathbf{e}_3\mathbf{e}_1 \\ &\quad + (a_2b_3 - b_2a_3)\mathbf{e}_2\mathbf{e}_3 \end{aligned} \tag{5}$$

This shows that every bivector can be expanded in terms of the three elementary bivectors $\mathbf{e}_1 \wedge \mathbf{e}_2 = \mathbf{e}_1\mathbf{e}_2$, $\mathbf{e}_3 \wedge \mathbf{e}_1 = \mathbf{e}_3\mathbf{e}_1$, and $\mathbf{e}_2 \wedge \mathbf{e}_3 = \mathbf{e}_2\mathbf{e}_3$. Moreover, it is easily shown that

$$(\alpha\mathbf{e}_1\mathbf{e}_2 + \beta\mathbf{e}_3\mathbf{e}_1 + \gamma\mathbf{e}_2\mathbf{e}_3)^2 = -\alpha^2 - \beta^2 - \gamma^2 \tag{6}$$

and hence these three bivectors must be linearly independent, so that the expansion is unique. This shows, in particular, that the space of bivectors is likewise three-dimensional.

If we similarly define the outer product of three vectors as their antisymmetrized geometric product, i.e.,

$$\mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c} = \tfrac{1}{6}(\mathbf{abc} - \mathbf{acb} + \mathbf{cab} - \mathbf{cba} + \mathbf{bca} - \mathbf{bac}) \tag{7}$$

a similar though longer calculation shows that $\mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c} = \det(\mathbf{a}, \mathbf{b}, \mathbf{c})\mathbf{e}_1\mathbf{e}_2\mathbf{e}_3$, where the determinant is of the matrix whose columns are the coordinates of **a**, **b**, and **c** versus the basis $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$. It follows that all trivectors are multiples of the trivector $\iota = \mathbf{e}_1\mathbf{e}_2\mathbf{e}_3$, which will henceforth be called the unit pseudo-scalar. This has the interesting property of behaving like the imaginary unit, since

$$\begin{aligned}\iota^2 &= (\mathbf{e}_1\mathbf{e}_2\mathbf{e}_3)^2 = -(\mathbf{e}_1\mathbf{e}_2\mathbf{e}_3)(\mathbf{e}_3\mathbf{e}_2\mathbf{e}_1) \\ &= -(\mathbf{e}_1\mathbf{e}_2)\mathbf{e}_3^2(\mathbf{e}_2\mathbf{e}_1) = -\mathbf{e}_1\mathbf{e}_2^2\mathbf{e}_1 = -\mathbf{e}_1^2 = -1 \end{aligned} \tag{8}$$

In keeping with the fact that space is three-dimensional, all outer products of four or more vectors must be zero.

Any element of the algebra, or multivector, can be uniquely expanded as a sum of a scalar, a vector, a bivector, and a pseudo-scalar, for a total dimension of eight. The multilinearity of the geometric product shows that scalars and bivectors are unchanged on inversion in the origin, whereas vectors and trivectors change sign. This fact also implies that the product of an even number of vectors has no vector or trivector part, whereas odd products have no scalar or bivector part. This is described by saying that the product preserves parity. Another interesting transformation reverses the order of the vectors in a product, e.g., $\overline{\mathbf{abc}} = \mathbf{cba}$. It is easily seen that scalars and vectors are unchanged by reversion, while bivectors and trivectors change sign.

Together, these facts imply that $\mathbf{abc} - \overline{\mathbf{abc}}$ is equal to the trivector $2\mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c}$. More generally, if we define the outer product of a vector **a** with a bivector $\mathbf{b} \wedge \mathbf{c}$ to be the symmetric part

$$\tfrac{1}{2}(\mathbf{a}(\mathbf{b} \wedge \mathbf{c}) + (\mathbf{b} \wedge \mathbf{c})\mathbf{a}) \tag{9}$$

of their geometric product, a straightforward calculation shows that

$$\mathbf{a} \wedge (\mathbf{b} \wedge \mathbf{c}) = \mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c} = (\mathbf{a} \wedge \mathbf{b}) \wedge \mathbf{c} \tag{10}$$

Thus the outer product is also associative.

It is also possible to define various 'mixed' products, which are combinations of inner and outer products. Of particular interest is the inner product of a vector **a** with a bivector $\mathbf{b} \wedge \mathbf{c}$, which is defined as the antisymmetric part of their geometric product. In this case, a direct calculation yields the vector

$$\mathbf{a}\bullet(\mathbf{b} \wedge \mathbf{c}) = \tfrac{1}{2}(\mathbf{a}(\mathbf{b} \wedge \mathbf{c}) - (\mathbf{b} \wedge \mathbf{c})\mathbf{a}) = (\mathbf{a}\bullet\mathbf{b})\mathbf{c} - (\mathbf{a}\bullet\mathbf{c})\mathbf{b} \tag{11}$$

Similarly, the inner product of two bivectors is defined as the symmetric part of their geometric product, in which case we have:

$$\begin{aligned}(\mathbf{a} \wedge \mathbf{b})\bullet(\mathbf{d} \wedge \mathbf{c}) &= \mathbf{a}\bullet(\mathbf{b}\bullet(\mathbf{d} \wedge \mathbf{c})) \\ &= (\mathbf{a}\bullet\mathbf{c})(\mathbf{b}\bullet\mathbf{d}) - (\mathbf{a}\bullet\mathbf{d})(\mathbf{b}\bullet\mathbf{c}) \\ &= \det\begin{bmatrix}\mathbf{a}\bullet\mathbf{c} & \mathbf{a}\bullet\mathbf{d} \\ \mathbf{b}\bullet\mathbf{c} & \mathbf{b}\bullet\mathbf{d}\end{bmatrix} \end{aligned} \tag{12}$$

Finally, the inner product of two trivectors is:

$$(\mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c})\bullet(\mathbf{f} \wedge \mathbf{e} \wedge \mathbf{d}) = \det\begin{bmatrix}\mathbf{a}\bullet\mathbf{d} & \mathbf{a}\bullet\mathbf{e} & \mathbf{a}\bullet\mathbf{f} \\ \mathbf{b}\bullet\mathbf{d} & \mathbf{b}\bullet\mathbf{e} & \mathbf{b}\bullet\mathbf{f} \\ \mathbf{c}\bullet\mathbf{d} & \mathbf{c}\bullet\mathbf{e} & \mathbf{c}\bullet\mathbf{f}\end{bmatrix} \tag{13}$$

Determinants of matrices of inner products as above are commonly called Gramians.

2.1.2 *Geometric Interpretation*

The utility of the above, purely formal, algebraic system lies in the fact that all the quantities appearing in it have distinct geometric meanings. This enables us to translate such basic geometric relations as congruence, perpendicularity, incidence, etc., into algebraic equations, and then to use the machinery of geometric algebra to derive new relations. For example,

$$\begin{aligned}\mathbf{a}^2 = \mathbf{b}^2 &\Leftrightarrow \text{congruent, or} \\ \mathbf{a}\bullet\mathbf{b} = 0 &\Leftrightarrow \text{perpendicular}\end{aligned} \tag{14}$$

This geometric interpretation also enables the entities that appear in the algebra to represent many different physical quantities in very natural ways.[2]

Because most scientists are familiar with Gibbs' vector algebra, the easiest way to introduce this interpretation is to show how Gibbs' algebra fits in to the more general geometric algebra introduced above. Since the inner product is already part of Gibbs' vector algebra, we shall begin with the outer product of two vectors. Using the above formula for $\mathbf{a} \wedge \mathbf{b}$ together with the identity

$$-\iota\mathbf{e}_1\mathbf{e}_2 = (\mathbf{e}_3\mathbf{e}_2\mathbf{e}_1)\mathbf{e}_1\mathbf{e}_2 = (\mathbf{e}_3\mathbf{e}_2)\mathbf{e}_1^2\mathbf{e}_2 = \mathbf{e}_3\mathbf{e}_2^2 = \mathbf{e}_3 \tag{15}$$

(and similar identities involving $\iota\mathbf{e}_3\mathbf{e}_1$ and $\iota\mathbf{e}_2\mathbf{e}_3$), we find that

$$-\iota(\mathbf{a}\wedge\mathbf{b}) = (a_2b_3 - b_2a_3)\mathbf{e}_1 + a_3b_1 - b_3a_1)\mathbf{e}_2 + (a_1b_2 - b_1a_2)\mathbf{e}_3 \tag{16}$$

This is the usual expression for Gibbs' vector cross product $\mathbf{a}\times\mathbf{b}$ in terms of coordinates, which is of course a vector perpendicular to the plane containing $\mathbf{a}$ and $\mathbf{b}$, of length equal to the area of the parallelogram whose sides are $\mathbf{a}$ and $\mathbf{b}$.

Since vectors are viewed as directed line segments, this makes it natural to view a bivector as a directed plane segment. The direction, in this case, specifies which side of the plane is 'up'; in a right-handed coordinate system, this is the same as the side that the cross product points away from. Unlike the cross product, however, the direction of a bivector is not changed by inversion in the origin. Because we have shown that $\iota^2 = -1$, the inverse relation follows immediately:

$$\iota(\mathbf{a}\times\mathbf{b}) = \mathbf{a}\wedge\mathbf{b} \tag{17}$$

More generally, multiplication by the unit pseudo-scalar ι maps vectors (bivectors) to their duals, which lie in the planes (lines) perpendicular to them, and have the same magnitudes.

Let us now consider Gibbs' triple product, which is a scalar equal to the signed volume of the parallelopiped spanned by the three vectors. Using the fact that pseudo-scalars commute with everything in the algebra:

$$\begin{aligned}\mathbf{a}\bullet(\mathbf{b}\times\mathbf{c}) &= -\tfrac{1}{2}(\mathbf{a}\iota(\mathbf{b}\wedge\mathbf{c}) + \iota(\mathbf{b}\wedge\mathbf{c})\mathbf{a}) \\ &= -\tfrac{\iota}{4}(\mathbf{abc} - \mathbf{acb} + \mathbf{bca} - \mathbf{cba}) = -\iota(\mathbf{a}\wedge\mathbf{b}\wedge\mathbf{c})\end{aligned} \tag{18}$$

This makes it natural to interpret the outer product of three vectors as an directed space segment, where the direction now refers to its chirality relative to the coordinate axes. It further implies that the geometric and inner product of two trivectors both equal the negative of the product of the volumes of their space segments.

Now take an arbitrary plane in space, e.g., $\mathbf{e}_1\mathbf{e}_2 = \mathbf{e}_1\wedge\mathbf{e}_2$. If $\mathbf{a} = a_1\mathbf{e}_1 + a_2\mathbf{e}_2$ is any other vector in this plane, then $\mathbf{e}_1\mathbf{a} = \mathbf{e}_1\bullet\mathbf{a} + \mathbf{e}_1\wedge\mathbf{a} = a_1 + a_2\mathbf{e}_1\mathbf{e}_2$, where

$$(\mathbf{e}_1\mathbf{e}_2)^2 = -(\mathbf{e}_1\mathbf{e}_2)(\mathbf{e}_2\mathbf{e}_1) = -1 \tag{19}$$

It follows that, relative to some fixed vector in the plane, every other vector in the plane can be regarded as a 'complex number'. In particular, multiplication by the imaginary unit of the plane $\mathbf{e}_1\mathbf{e}_2$ rotates all vectors in the plane by one quarter turn. There is a different imaginary unit for every plane, however, which is why we do not give it a special symbol as we did the pseudo-scalar ι.

The inner product of a vector and a bivector can thus be interpreted as a projection onto the plane of the bivector followed by a quarter turn and scaling by the magnitude of the bivector. The projection itself can be recovered by multiplying it by the inverse bivector, i.e.,

$$(\mathbf{a}\bullet(\mathbf{e}_1\mathbf{e}_2))(\mathbf{e}_2\mathbf{e}_1) = (a_1\mathbf{e}_2 - a_2\mathbf{e}_1)(\mathbf{e}_2\mathbf{e}_1) = a_1\mathbf{e}_1 + a_2\mathbf{e}_2 \tag{20}$$

where $a_1 = \mathbf{a}\bullet\mathbf{e}_1$ and $a_2 = \mathbf{a}\bullet\mathbf{e}_2$. To interpret the inner product of two bivectors, on the other hand, we use the fact that any two planes intersect in a line, and let one of the factors of each bivector lie on this line. Then by the law of cosines for spherical trigonometry,

$$\begin{aligned}(\mathbf{a}\wedge\mathbf{b})\bullet(\mathbf{b}\wedge\mathbf{c}) &= (\mathbf{a}\bullet\mathbf{c})\|\mathbf{b}\|^2 - (\mathbf{a}\bullet\mathbf{b})(\mathbf{b}\bullet\mathbf{c}) \\ &= (\cos(\theta_{\mathbf{ac}}) - \cos(\theta_{\mathbf{ab}})\cos(\theta_{\mathbf{bc}}))\|\mathbf{a}\|\,\|\mathbf{b}\|^2\|\mathbf{c}\| \\ &= -(\sin(\theta_{\mathbf{ab}})\sin(\theta_{\mathbf{bc}})\cos(\phi))\|\mathbf{a}\|\,\|\mathbf{b}\|^2\|\mathbf{c}\| \\ &= -\|\mathbf{a}\wedge\mathbf{b}\|\,\|\mathbf{b}\wedge\mathbf{c}\|\cos(\phi)\end{aligned} \tag{21}$$

where ϕ is the dihedral angle between the planes $\mathbf{a}\wedge\mathbf{b}$ and $\mathbf{b}\wedge\mathbf{c}$.

Now take an arbitrary unit vector $\mathbf{e}$, and consider the expression

$$-\mathbf{eae} = -\mathbf{e}(\mathbf{a}\bullet\mathbf{e} + \mathbf{a}\wedge\mathbf{e}) = \mathbf{a} - 2(\mathbf{a}\bullet\mathbf{e})\mathbf{e} \tag{22}$$

This is just the reflection of $\mathbf{a}$ in the plane perpendicular to $\mathbf{e}$. It is well known that the product of two reflections is a rotation about the axis in which their planes intersect, and by twice the smaller angle between the planes. Hence for any two unit vectors $\mathbf{e}$, $\mathbf{f}$, the expression

$$\begin{aligned}(\overline{\mathbf{ef}})\mathbf{a}(\mathbf{ef}) &= (\mathbf{e}\bullet\mathbf{f} - \mathbf{e}\wedge\mathbf{f})\mathbf{a}(\mathbf{e}\bullet\mathbf{f} + \mathbf{e}\wedge\mathbf{f}) \\ &= \overline{\mathbf{R}}\mathbf{a}\mathbf{R}\ (\mathbf{R}\equiv\mathbf{e}\bullet\mathbf{f} + \mathbf{e}\wedge\mathbf{f})\end{aligned} \tag{23}$$

describes a rotation of the vector $\mathbf{a}$, where the angle of rotation is $2\arccos(\mathbf{e}\bullet\mathbf{f})$. It follows that any rotation can be represented by the sum of a scalar and a bivector

$$\mathbf{R} = \sigma + \rho\cdot\iota\mathbf{r}, \quad \text{where } \sigma^2 + \rho^2 = 1 \Leftrightarrow \overline{\mathbf{R}}\mathbf{R} = 1 \tag{24}$$

and $\mathbf{r}$ is a unit vector along the axis of rotation. This representation is unique up to sign.

The products of even numbers of vectors form a subalgebra of the geometric algebra, called the even subalgebra. If we let $\mathbf{I} = \mathbf{e}_2\mathbf{e}_3$, $\mathbf{J} = \mathbf{e}_3\mathbf{e}_1$ and $\mathbf{K} = \mathbf{e}_1\mathbf{e}_2$ be our standard bivector basis, we find that

$$\begin{aligned}&\mathbf{I}^2 = \mathbf{J}^2 = \mathbf{K}^2 = -1, \quad \mathbf{JI} = \mathbf{K} = -\mathbf{IJ}, \\ &\mathbf{IK} = \mathbf{J} = -\mathbf{KI}, \quad \text{and} \quad \mathbf{KJ} = \mathbf{I} = -\mathbf{JK}\end{aligned} \tag{25}$$

Up to sign, these are the relations that define Hamilton's quaternions. If we represent each vector $\mathbf{a}$ by its dual $\mathbf{A} = \iota\mathbf{a}$, we thus obtain Cayley's formula

$$\overline{\mathbf{R}}\mathbf{A}\mathbf{R} = \iota\overline{\mathbf{R}}\mathbf{a}\mathbf{R} \tag{26}$$

for rotations in terms of quaternions. It follows that the even subalgebra can be identified with Hamilton's quaternions, and used to describe the rotations of vectors in the same way.

2.2 Invariant Theory

The scalar-valued expressions in the geometric algebra generated by the interpoint vectors are of particular interest, because they are automatically invariant under translations and rotations. The theory of such invariant expressions shows that they can always be reduced to multivariate polynomials in the squared interpoint distances together with the signed volumes of tetrahedra. The relations among these fundamental invariants, obtained from Cayley–Menger determinants,[4,5] provide general equations for use in geometric reasoning, and these determinants can likewise be viewed as entities within geometric algebra. In this section we shall demonstrate these facts, and show how these quantities can be interpreted within the geometric algebra of three dimensions.

2.2.1 The Fundamental Invariants

Invariant theory is a branch of mathematics that deals with polynomials that are invariant (i.e., whose values are

preserved) under a group of transformations of their variables.[6] In the case of the three-dimensional Euclidean group consisting of all translations and proper rotations of the Cartesian coordinates, the squared distance between points, and the signed volumes spanned by tetrahedra, are obvious examples of such invariants. The first fundamental theorem of invariant theory for the Euclidean group states that these two types of invariants constitute a complete system, meaning that any invariant can be written as a multivariate polynomial in the squared distances and signed volumes.[7]

The first item of business is to show how any set of multivectors in the geometric algebra of three dimensions can be characterized, up to rotation, by a system of scalar-valued expressions in these fundamental invariants. Any multivector can always be separated into its scalar, vector, bivector, and trivector parts. The scalar part is ready to go, while the trivector part can be converted to a scalar simply by multiplying it by the unit pseudo-scalar. We next observe that any set of vectors is determined, up to rotation, by their Gram matrix of inner products. This is easily seen by taking any maximal linearly independent subset $\mathbf{a}, \mathbf{b}, \ldots, \mathbf{c}$ (whose Gramian is necessarily nonzero), and noting that the inner products of any other vector $\mathbf{x}$ with this basis determine the coordinates of that vector versus the basis, via the normal equations:

$$\begin{bmatrix} \mathbf{a}^2 & \mathbf{a}\bullet\mathbf{b} & \ldots & \mathbf{a}\bullet\mathbf{c} \\ \mathbf{a}\bullet\mathbf{b} & \mathbf{b}^2 & \ldots & \mathbf{b}\bullet\mathbf{c} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}\bullet\mathbf{c} & \mathbf{b}\bullet\mathbf{c} & \ldots & \mathbf{c}^2 \end{bmatrix} \begin{bmatrix} x_a \\ x_b \\ \vdots \\ x_c \end{bmatrix} = \begin{bmatrix} \mathbf{a}\bullet\mathbf{x} \\ \mathbf{b}\bullet\mathbf{x} \\ \vdots \\ \mathbf{c}\bullet\mathbf{x} \end{bmatrix} \tag{27}$$

Mixed sets of vectors and bivectors can likewise be characterized by the Gram matrices of the vectors together with the duals of the bivectors; as shown previously, the inner product of a vector with the dual of a bivector is a triple product.

Using the distributive properties of inner and outer products, any products of linear combinations can be expanded into linear combinations of products. Thus any scalar-valued expression can be expanded into a polynomial in the inner products of pairs of vectors, bivectors, or trivectors, together with scalars and triple products. Inner products of bivectors and trivectors can be further expanded into polynomials in the inner products of vectors only, using the equivalence to Gramians derived in the previous section. Moreover, if two triple products occur in any term, we can likewise expand them as a Gramian into a polynomial in the vector inner products, since

$$\begin{aligned}(\imath(\mathbf{a}\wedge\mathbf{b}\wedge\mathbf{c}))(\imath(\mathbf{d}\wedge\mathbf{e}\wedge\mathbf{f})) &= \imath^2(\mathbf{a}\wedge\mathbf{b}\wedge\mathbf{c})(\mathbf{d}\wedge\mathbf{e}\wedge\mathbf{f}) \\ &= (\mathbf{a}\wedge\mathbf{b}\wedge\mathbf{c})\bullet(\mathbf{f}\wedge\mathbf{e}\wedge\mathbf{d})\end{aligned} \tag{28}$$

By this means we can reduce any scalar-valued expression to a polynomial in the inner products of single pairs of vectors, together with vector triple products, that is only linear in the triple products. Thus any set of multivectors can be characterized, up to rotation, by specifying the scalar values of a system of such multivariate polynomials.

Any translation-independent expression in geometric algebra can be written in terms of the interpoint vectors only, but it can be quite difficult to actually perform this rewriting. Thus to be assured of translation-independence, it is best to work only with interpoint vectors (rather than vectors from an arbitrary origin) throughout the calculations. Because of the relation $(\mathbf{a}-\mathbf{b})+(\mathbf{b}-\mathbf{c})=(\mathbf{a}-\mathbf{c})$, one can choose one's interpoint vectors in many different ways. This in turn leads to many different possible systems of polynomials for the same set of multivectors, some of which may be far more complicated than necessary.

These ambiguities in our choice of interpoint vectors can be eliminated by replacing all the inner products of interpoint vectors by linear combinations of squared distances, using the generalized law of cosines

$$\begin{aligned}(\mathbf{a}-\mathbf{b})\bullet(\mathbf{c}-\mathbf{d}) &= \tfrac{1}{2}(D_{ad}+D_{bc}-D_{ac}-D_{bd}) \\ &= \tfrac{1}{2}(\|\mathbf{a}-\mathbf{d}\|^2+\|\mathbf{b}-\mathbf{c}\|^2-\|\mathbf{a}-\mathbf{c}\|^2-\|\mathbf{b}-\mathbf{d}\|^2)\end{aligned} \tag{29}$$

(where $D_{ad}=\|\mathbf{a}-\mathbf{d}\|^2$, etc.), which is easily verified by expanding both sides and cancelling terms. Thus we are able to reduce any polynomial in the inner products of interpoint vectors to a polynomial in the squared distances. In the event that we are working with vectors from a common origin, application of the above relation with $\mathbf{b}=\mathbf{d}$ set to the origin will result in a polynomial with no squared distances to the origin in it if and only if the original expression was translation-independent.

In summary, we have shown that any scalar-valued, translation and rotation-independent expression in the geometric algebra of three-dimensions can be written in terms of the squared distances and signed volumes, and hence any set of multivectors can also be characterized, up to translation and rotation, by the values of a system of such expressions.

2.2.2 The Fundamental Syzygies

The number of squared distances and signed volumes among a set of N three-dimensional points generally exceeds the number of internal degrees of freedom $3N-6$. For example, if one fixes all but one of the ten distances among a set of five points, the remaining distance can assume at most two possible values. The second fundamental theorem of invariant theory states that these algebraic relations, or syzygies as they are called, among the squared distances and oriented volumes can be written as multivariate polynomials in a complete system of syzygies. While it has never actually been proven in full generality, a safe bet is that certain determinants in the squared distances, known as Cayley–Menger determinants, together with a few auxilliary relations connecting them to the signed volumes, constitute a complete system of syzygies for Euclidean geometry.

The most important such relations turn out to be consequences of the fact that, in three dimensions, the hypervolume spanned by any four interpoint vectors is always zero, i.e.,

$$(\mathbf{b}-\mathbf{a})\wedge(\mathbf{c}-\mathbf{a})\wedge(\mathbf{d}-\mathbf{a})\wedge(\mathbf{e}-\mathbf{a})=0 \tag{30}$$

We can convert this into a scalar-valued expression by taking the inner product with its reverse. The resulting Gramian can be Laplace expanded to a polynomial in the inner products of the four vectors, which in turn can be converted into a polynomial in the squared interpoint distances as above. The vanishing of this polynomial is our first syzygy among the squared distances.

These polynomials tend to be rather complicated, and hence it is fortunate that they can be written simply in terms of Cayley–Menger determinants. To see how this works, consider the inner square of a single pair of interpoint vectors $\mathbf{x}=\mathbf{b}-\mathbf{a}$

and $\mathbf{y} = \mathbf{c} - \mathbf{a}$. This can be expanded as a Gramian which, by the Cauchy–Schwarz inequality, is always nonnegative:

$$0 \le (\mathbf{x} \wedge \mathbf{y}) \bullet (\mathbf{y} \wedge \mathbf{x}) = \det \begin{bmatrix} \mathbf{x}^2 & \mathbf{x} \bullet \mathbf{y} \\ \mathbf{x} \bullet \mathbf{y} & \mathbf{y}^2 \end{bmatrix} \tag{31}$$

We can augment the determinant with a pair of unit row/columns without changing its value, and then use elementary row/column operations to change variables, as follows:

$$\begin{aligned} &\det \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \mathbf{x}^2 & \mathbf{x} \bullet \mathbf{y} \\ 0 & 0 & \mathbf{x} \bullet \mathbf{y} & \mathbf{y}^2 \end{bmatrix} \\ &= -\det \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & \mathbf{x}^2 & \mathbf{x} \bullet \mathbf{y} \\ 1 & 0 & \mathbf{x} \bullet \mathbf{y} & \mathbf{y}^2 \end{bmatrix} \\ &= -\det \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & -\frac{1}{2}\mathbf{x}^2 & -\frac{1}{2}\mathbf{y}^2 \\ 1 & -\frac{1}{2}\mathbf{x}^2 & 0 & \mathbf{x} \bullet \mathbf{y} - \frac{1}{2}(\mathbf{x}^2 + \mathbf{y}^2) \\ 1 & -\frac{1}{2}\mathbf{y}^2 & \mathbf{x} \bullet \mathbf{y} - \frac{1}{2}(\mathbf{x}^2 + \mathbf{y}^2) & 0 \end{bmatrix} \end{aligned} \tag{32}$$

This last determinant can be rewritten as a three-point Cayley–Menger determinant:

$$\begin{aligned} &-\det \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & -D_{ab}/2 & -D_{ac}/2 \\ 1 & -D_{ab}/2 & 0 & -D_{bc}/2 \\ 1 & -D_{ac}/2 & -D_{bc}/2 & 0 \end{bmatrix} \\ &= -\tfrac{1}{4}\det \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & d_{ab}^2 & d_{ac}^2 \\ 1 & d_{ab}^2 & 0 & d_{bc}^2 \\ 1 & d_{ac}^2 & d_{bc}^2 & 0 \end{bmatrix} \end{aligned} \tag{33}$$

where $\mathbf{x}^2 = (\mathbf{b} - \mathbf{a})^2 = D_{ab} = d_{ab}^2$, etc. Finally, the polynomial in the distances obtained by expanding this determinant can be factorized as:

$$\begin{aligned} &\tfrac{1}{4}(d_{ab} + d_{ac} + d_{bc})(d_{ab} + d_{ac} - d_{bc})(d_{ac} + d_{bc} - d_{ab}) \\ &\quad \times (d_{bc} + d_{ab} - d_{ac}) \end{aligned} \tag{34}$$

The nonnegativity of the determinant is thus equivalent to the three triangle inequalities among the points, i.e.,

$$d_{ab} \le d_{ac} + d_{bc} \qquad d_{ac} \le d_{bc} + d_{ab} \qquad d_{bc} \le d_{ab} + d_{ac} \tag{35}$$

In the following, we shall denote such a symmetric three-point Cayley–Menger determinant by $D(a, b, c)$. The general definition of an N-point Cayley–Menger determinant is

$$\begin{aligned} &D(a, b, \ldots, c; p, q, \ldots, r) \\ &= 2\left(\frac{-1}{2}\right)^N \det \begin{bmatrix} 0 & 1 & 1 & \ldots & 1 \\ 1 & D_{ap} & D_{aq} & \ldots & D_{ar} \\ 1 & D_{bp} & D_{bq} & \ldots & D_{br} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & D_{cp} & D_{cq} & \ldots & D_{cr} \end{bmatrix} \end{aligned} \tag{36}$$

A derivation similar to that used above for the symmetric three-point case shows that

$$\begin{aligned} &D(a, b, \ldots, c; p, q, \ldots r) \\ &\quad = ((\mathbf{b} - \mathbf{a}) \wedge \ldots \wedge (\mathbf{c} - \mathbf{a})) \bullet (\overline{(\mathbf{q} - \mathbf{p}) \wedge \ldots \wedge (\mathbf{r} - \mathbf{p})}) \end{aligned} \tag{37}$$

Thus a four-point Cayley–Menger determinant is equal to the product of the signed volumes spanned by the parallelepipeds whose sides are the vectors from one of the points to the other three (in each of the two point sets separately). This is the syzygy that connects the squared distances to the signed volumes. The five-point Cayley–Menger determinants, on the other hand, are the Gramians of four interpoint vectors, and hence vanish identically whenever the distances are three-dimensional, as described at the beginning of this section.

Although it is much more difficult to prove, it turns out that the nonnegativity of the symmetric two-, three-, and four-point Cayley–Menger determinants, together with the vanishing of all higher determinants, is also sufficient for any symmetric matrix of real numbers with zeros down the diagonal to be a matrix of squared distances in a three-dimensional Euclidean space. In fact, assuming that none of the distances are zero, it is sufficient if all five-point Cayley–Menger determinants vanish, with the exception of a single type of six-point counterexample.[1,4,5]

Note that our derivation of the equivalence of Cayley–Menger determinants and the Gramians of interpoint vectors augmented the Gram matrix by two additional rows and columns. Thus the matrices in Cayley–Menger determinants can likewise be regarded as Gram matrices among null vectors in a five-dimensional space, which turns out to have a Minkowski metric (like space-time in the theory of relativity). The vectors in this five-dimensional space constitute homogenous coordinates for an extension of Euclidean geometry, known as Mobius sphere geometry, which includes spheres and planes as its elemental objects along with points. A plane is a special case of a sphere, whose center lies 'at infinity', and this point-at-infinity corresponds to the border of ones in a Cayley–Menger determinant. The translations of Euclidean points in this space are rotations relative to the Minkowski metric! Further details may be found in Refs. 8–10 and references cited therein.

In closing, we note that the signed volumes themselves are not independent, but satisfy a linear relation among each set of five points. For ease of presentation we shall assume that one of these points is located at the origin. Then this relation is readily derived by expanding the (dual of) the outer product:

$$\begin{aligned} (\mathbf{b} - \mathbf{a}) \wedge (\mathbf{c} - \mathbf{a}) \wedge (\mathbf{d} - \mathbf{a}) = \mathbf{b} \wedge \mathbf{c} \wedge \mathbf{d} - \mathbf{a} \wedge \mathbf{c} \wedge \mathbf{d} \\ + \mathbf{a} \wedge \mathbf{b} \wedge \mathbf{d} - \mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c} \end{aligned} \tag{38}$$

That is, if we order the five points arbitrarily, and alternately add and subtract the signed volumes of the quadruple obtained by taking out each point in turn, we obtain zero. Suprisingly, this relation turns out to be a consequence of the foregoing relations between the signed volumes and squared distances, and hence does not need to be imposed as an independent condition.[1]

3 ALGORITHMS

The above theory is useful in gaining insight into, and closed form solutions for, simple problems involving small

molecules, but the scope of its applications is limited by sheer algebraic complexity. Thus we are forced to resort to numerical methods for finding isolated solutions to larger problems. In order to gain at least some insight into the structure of the conformation space as a whole, we generally attempt to find a number of different solutions (i.e., conformations) that satisfy the constraints. Such a set of conformations is called a conformational ensemble. Providing this ensemble is sufficiently large and random, any geometric features that are common to all its members can be assumed to be necessary consequences of the constraints. This constitutes an inductive approach to geometric reasoning, as opposed to the deductive approaches considered above.

While one could require these conformations to satisfy a wide variety of geometric conditions by means of constraints on suitable polynomials in the interatomic squared distances and signed volumes, it turns out that the simplest possible such constraints are also the most widely useful. These are lower and upper bounds on the interatomic squared distances themselves, together with the signs ($+1, -1$, or 0) of the volumes of selected quadruples of atoms. The latter, called chirality constraints, determine the chirality of the quadruple, or force it to be planar if the sign is zero. The totality of constraints of this form is called a distance geometry description. Experience has shown that most 'conformation spaces' of practical interest in chemical problems can be accurately described by means of these simple constraints alone.

In this section we shall present several algorithms, which collectively provide a means of computing conformational ensembles consistent with distance geometry descriptions. The overall procedure is often referred to as 'the EMBED algorithm', although that term, strictly speaking, applies only to the coordinate generation process in step (2) below. The procedure, in any case, consists of the following three, broadly defined, steps:

(1) Extrapolating a complete set of lower and upper limits on all the distances from the sparse set of lower and upper bounds that are usually available, a process known as bound smoothing.
(2) Choosing a random distance matrix from within these limits, and computing coordinates that are a certain best-fit to the distances, in a process itself called embedding.
(3) Optimizing these coordinates versus an 'error' function which measures the total violation of the distance and chirality constraints, usually by some form of simulated annealing.

3.1 Bound Smoothing

The three-dimensional distance limits are the minimum and maximum value that each distance can assume in any three-dimensional structure all the distances of which lie between their given lower and upper bounds. The computation of these distance limits is a difficult and unsolved problem, but loose approximations are available. By 'loose', we mean that each approximate upper limit is at least as large as the true upper limit, while each approximate lower limit is no larger than the true lower limit. The most important of these approximate limits are called the triangle inequality limits. The reason they are the most important is that they can be computed rapidly and reliably, even for very large problems. Unfortunately, the triangle inequality limits are often very poor approximations to the true three-dimensional limits, particularly with regard to the lower limits. By using the tetrangle inequality obtained from the nonnegativity of the symmetric four-point Cayley–Menger determinants, it is sometimes possible to obtain substantially 'tighter' limits. The computational expense of doing so, however, rapidly becomes prohibitive past 100 to 200 atoms. The process of calculating such distance limits is generally known as bound smoothing. It is the most important, and least well-solved, step of the overall procedure.

3.1.1 The Triangle Inequality Limits

The triangle inequality limits are the minimum and maximum values that the distances can assume in any metric space consistent with the bounds. A metric space, for our purposes, is simply a distance matrix whose distances satisfy the triangle inequality. The limits are attained in certain extremal metric spaces, with characteristic patterns of distances equal to their bounds. The patterns that can occur in a three-point metric space are shown in Figure 1. In addition to providing some finite range of values for every distance in a molecule, triangle inequality bound smoothing can locate certain contradictions in the bounds, called triangle inequality violations.

In order to compute the triangle inequality limits efficiently, we reduce their calculation to an all pairs shortest paths problem in a certain digraph (i.e., a collection of 'nodes' connected by 'arcs' with an arrow at one or both ends). A 'path' in such a digraph is a sequence of nodes such that any two consecutive nodes in the sequence are connected by an arc in the digraph, whose arrow points from the first to the second; the 'length' of the path is the sum of the lengths assigned to its arcs. It is easily seen that the upper triangle limits are equal to the lengths of the shortest paths in an undirected digraph (i.e., one whose arcs are all two-headed), whose arc lengths are equal to the given upper bounds. It is a little more difficult to show that all the lower triangle limits are of the form

$$\overline{\ell}_{ij} = \ell_{km} - \overline{u}_{ik} - \overline{u}_{jm} \tag{39}$$

where i, j, k, m are not necessarily distinct atom indices, ℓ_{ij}, u_{ij} respectively denote lower and upper bounds on the corresponding indexed atoms, and overbars indicate the corresponding triangle inequality limits.[11] This shows, first of all, that the upper limits can be computed independently of the lower, and secondly, that the greatest lower limit cannot exceed the greatest lower bound. It also shows that the negatives of the lower limits are the lengths of shortest paths in a certain digraph which is guaranteed to have at most one negative lower bound in each path.

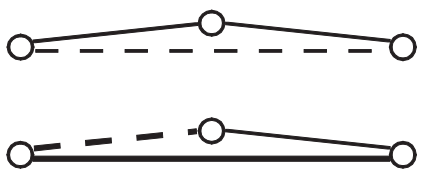

Figure 1 The upper (top) and lower (bottom) triangle inequality limits. The heavy solid line denotes the distance at its lower bound, while a light solid line denotes the distance at its upper bound; the dashed line denotes the associated limit

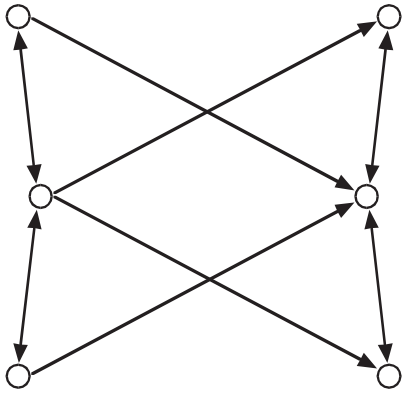

Figure 2 The digraph whose shortest paths determine the triangle inequality limits. The two-headed arrows in the left and right halves have length equal to the upper bound between the corresponding pair of atoms, while the one-headed arcs going from left to right have length equal to the negative of the lower bound. Note that not all possible arcs are present

This digraph consists of two sets of nodes, where each set contains one node for every atom in the system. Within each set, the two-headed arcs connecting pairs of nodes have length equal to the upper bounds between the corresponding pairs of atoms, while between the two node sets, the one-way arcs have length equal to the negatives of the lower bounds between the corresponding pairs of atoms. Figure 2 illustrates such a digraph. The fact that all the negative arcs go in one direction between the two node sets ensures that this digraph contains no negative cycles, which would lead to shortest paths with a length of negative infinity! The shortest paths within each node set are clearly the triangle inequality upper limits, while those between the two node sets are the negatives of the lower limits (if less than zero). Moreover, if a triangle inequality violation $\ell_{ij} > \overline{u}_{ij}$ is found, the erroneous constraints must lie on the shortest paths that determine these limits.

Perhaps the simplest shortest paths algorithm is Floyd's algorithm. This algorithm takes each node k of the digraph in turn, and then makes a pass through all ordered pairs of other nodes (i, j). If the length of the path $i \to k \to j$ is shorter than the length of the direct path $i \to j$, the latter is set to the former. This ensures that after each pass all the path lengths are at least as short as any path that goes through the node k, and hence iterating on this procedure for $k = 1, \ldots, N$ produces the desired matrix of shortest paths. The following pseudocode implements this procedure for a digraph of the form described above.

```
procedure Floyd( Natom,Lower,Upper )
   for k from 1 to Natom do
      for i from 1 to Natom − 1 do
         for j from i + 1 to Natom do
comment: Path lengths in left-hand network.
            if Upper[i,j] > Upper[i,k] + Upper[k,j] then
               Upper[i,j] :=Upper[i,k] + Upper[k,j];
comment: Path lengths from left to right-hand network.
            if Lower[i,j] < Lower[i,k] − Upper[k,j] then
               Lower[i,j] :=Lower[i,k] − Upper[k,j];
            else
            if Lower[i,j] < Lower[j,k] − Upper[k,i] then
               Lower[i,j] :=Lower[j,k] − Upper[k,i];
comment: Check for triangle inequality violations.
            if Lower[i,j] > Upper[i,j] then
               exit( ''bad bounds'' );
   endfor endfor endfor
endproc
```

Clearly, this algorithm requires time proportional to the cube of the number of atoms in every case. This algorithm is simple and general, but does not take advantage of the fact that the bounds are generally very sparse, meaning that no explicit lower and upper bounds are available for most pairs of atoms. A much faster algorithm takes advantage of sparsity by constructing a shortest paths tree, one from each node on the 'left' side of the network in turn, which contains only those arcs whose lengths are equal to the available bounds. Such shortest path algorithms were first applied to triangle inequality bound smoothing in the DISGEO program,[12] but ignored the hard-sphere lower bounds to do so. These lower bounds are an obvious exception to the sparsity of the constraints, but because they are very small they are seldom involved in any other lower triangle limits. Nevertheless, it has recently been found that by simply sorting the atoms by radii, and ceasing to look at the implicit arcs whose lengths are the negative sum of the corresponding radii once that sum falls below the current path length, one can also handle large numbers of small hard-sphere lower bounds with no significant increase in running time.

3.1.2 The Tetrangle Inequality Limits

The tetrangle inequality is a consequence of the nonnegativity of the symmetric four-point Cayley–Menger determinants. These do not factorize like the three-point determinants, but they can be expanded as a quadratic function of one of the squared distances, e.g., $D(3,4)$, as follows:

$$D(1,2,3,4) = D(1,2,3,4)|_{D(3,4)=0} - D^2(3,4)D(1,2)/4 + D(3,4)D(1,2,3;1,2,4)|_{D(3,4)=0} \tag{40}$$

Here, the notation

$$D(1,2,3,4)|_{D(3,4)=0} \tag{41}$$

means that the determinant is evaluated with the one squared distance $D(3,4)$ replaced by zero, and similarly for $D(1,2,3;1,2,4)$. The discriminant of this quadratic is

$$D(1,2,3)D(1,2,4), \tag{42}$$

and hence nonnegative whenever all the triangle inequalities not involving $D(3,4)$ are satisfied. Since the leading coefficient is negative, the four-point Cayley–Menger determinant is nonnegative for all values of $D(3,4)$ in a closed interval between the nonnegative roots of this quadratic. These roots are called the lower and upper tetrangle inequality limits on $D(3,4)$. In the case of a chain of four covalently bonded atoms 3–1–2–4, they correspond to the *cis/trans* limits.

Like the triangle inequality limits, all possible tetrangle inequality limits on any one distance among each four points, given bounds on the remaining five distances, can be explicitly enumerated. Up to reindexing, there are two possible combinations of bounds for the lower limits, and two for the upper limits; these are shown in Figure 3. Unfortunately, there is no equivalent shortest paths characterization for the tetrangle inequality limits. This forces us to use a comparatively brute force approach, in which we initially set the lower and upper limits to the bounds (or zero and infinity where no bounds are available), and iteratively scan all quadruples replacing the current limits by the triangle and tetrangle limits on one

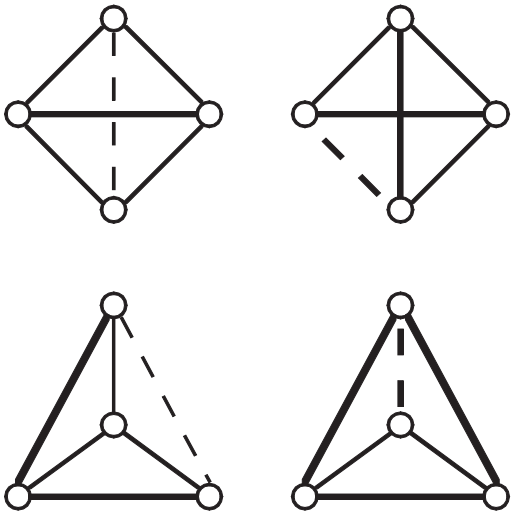

Figure 3 The upper (left) and lower (right) tetrangle inequality limits. A heavy solid line denotes the distance at its lower bound, while a light solid line denotes the distance at its upper bound; a dashed line denotes the associated limit

distance in the quadruple (if tighter) until no further changes occur. This procedure can be summarized as follows:

```
procedure Easthope( Natom,Lower,Upper )
  for i from 1 to Natom do for j from i+1 to Natom do
    for k from 1 to Natom do for m from k+1 to Natom do
comment: See if k and m can be made collinear with i or j.
          if not Collinear( i,k,m,Lower,Upper )
          and not Collinear( j,k,m,Lower,Upper ) then
comment: Tighten k,m upper limit by tetrangle limit.
            test := UpTetLim( i,j,k,m,Lower,Upper );
            if test < Upper[k,m] then Upper[k,m] := test;
comment: Tighten k,m lower limit by tetrangle limit.
            test := LoTetLim( i,j,k,m,Lower,Upper );
            if test > Lower[k,m] then Lower[k,m] := test;
          endif
comment: Check for tetrangle inequality violations.
          if Lower[k,m] > Upper[k,m] then
            exit( ''bad bounds'' );
  endfor endfor endfor endfor
comment: Test for convergence.
  if any changes were made to the limits then
    return TRUE else return FALSE;
endproc
```

The subprocedure `Collinear` checks if the corresponding triple of atoms can be made collinear, given the current limits among the atoms; if so, it updates the current limits by the corresponding triangle inequality limits and exits `FALSE`. The subprocedures `LoTetLim` and `UpTetLim`, which are valid only if these triples cannot be made collinear, return the tightest tetrangle inequality limits enumerated as above. This procedure (named after the programmer who first implemented it[13]) is called iteratively until no further changes occur, as indicated by its return status.

Each call to `Easthope` requires time proportional to the fourth power of the number of atoms, and hence tetrangle inequality bound smoothing is too time consuming to be used for more than 100 to 200 atoms. In addition, large improvements in the limits are generally obtained only in subsets of atoms wherein the bounds force the atoms to be approximately coplanar. Since such subsets do not account for the majority of atoms in most molecules, tetrangle inequality bound smoothing is of limited practical importance. The next logical step in the progression would be to use pentangle inequality bound smoothing, but while we would expect significant improvements from such a procedure, it would be even more computationally intensive.

Some indication of the difficulty of developing efficient procedures for computing higher-order limits may be obtained from the following considerations. The tetrangle limits are properly defined as the minimum and maximum values that the distances can assume in any metric space consistent with both the bounds and the tetrangle inequality. A procedure that iterates over all quadruples tightening the limits according to the triangle and tetrangle inequalities (as above) will not, however, find these limits generally. A set of lower and upper bounds among five points, known as the Cauchy pentagon (Figure 4), constitutes a counterexample.[1]

3.2 The EMBED Algorithm

The next step in the overall calculation of coordinates is to choose a random distance matrix from within the distance limits, and fit a set of coordinates to it. By using a different random number seed in multiple passes through this step, one obtains an ensemble of different conformations. Early applications of this procedure to conformational calculations simply chose uniformly distributed, independent random numbers from within the limits, in order to obtain the random distance matrix. This had the effect of producing structures that were relatively expanded, and also seemed rather more similar to one another than they should be, especially when the limits were very 'loose'. This in turn led to concerns that the overall procedure might not be sampling conformation space adequately. Subsequently it was found that the degree of expansion could be controlled at will by biasing the distances towards larger or smaller values, and a procedure called metrization was developed which greatly improved the sampling, by forcing the random distances to satisfy the triangle inequality as well as the given limits.[15] The actual process of fitting coordinates to the distances, called embedding, has proven remarkably robust, and although a number of variations have been developed,[16,17] it seems safe to say that it has not been greatly improved since it was first introduced.[18]

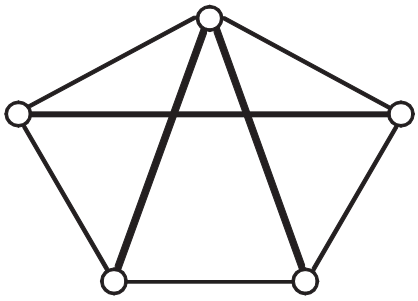

Figure 4 The Cauchy pentagon is a tensegrity framework,[14] whose 'struts' (heavy lines) constitute lower bounds, and whose 'cables' are upper bounds, on the associated distances. The upper limit on any one pair of nodes separated by a strut, or the lower limit on any one pair connected by a cable, are tetrangle inequality limits implied by the bounds that correspond to the remaining struts and cables. Such five-point tetrangle inequality limits would not be found by an algorithm that only iterates over all quadruples of atoms as above

3.2.1 *Metrization*

Metrization is based on the following facts:

(1) The triangle inequality limits are the minimum and maximum values that the distances can assume in any metric space consistent with the triangle inequality limits themselves.
(2) The set of all metric spaces satisfying the distance limits, being the intersection of the 'box' defined by the limits and the convex cone defined by the triangle inequalities, is itself a convex set. Thus every value of every distance between its lower and upper limits is attained in some metric space consistent with all the limits.

The procedure is now almost obvious. First, one takes one of the distances and sets it to some random number between its lower and upper limits. One then sets its lower and upper limits to this number, and recomputes the triangle inequality limits using these modified limits as the input bounds. Repeating this procedure for each and every distance in turn eventually yields a set of lower and upper triangle inequality limits that are equal to each other, and also lie within the original limits. These limits therefore are also equal to the desired matrix of distances satisfying both the triangle inequality and the original limits.

This procedure, though straightforward, would require time proportional to the fifth power of the number of atoms. Fortunately, if one has computed a shortest paths tree from a root node (utilizing, in this case, all the distance limits, and not just those given by the relatively sparse set of distance bounds), it is possible to update the limits after tightening one limit between the root of the tree and any other node in time that is only linear in the number of nodes. This reduces the total time required for metrization to only the cube of the number of atoms. It has the disadvantage, nonetheless, of restricting the order in which one chooses the distances to all those involving one atom, and then all the rest involving some other atom, etc. This propsective form of metrization fills up the above-diagonal half of a distance matrix by row. An alternative, called retrospective metrization, fills up the above-diagonal half by column, with quite similar overall results. If we use negative indices for nodes in the 'right' half of the digraph introduced in Section 3.1 above, the following pseudocode illustrates the prospective case:

```
procedure Metrize( Natom,Lower,Upper )
   for r from 1 to Natom-1 do
comment: Make shortest paths tree using current limits.
      tree :=MakeTree( r,Lower,Upper );
      for s from r+1 to Natom do
         lub :=tree.path_length[s];
         glb :=-tree.path_length[-s];
comment: Choose random number between current limits.
         distance :=Random( glb,lub );
comment: Set the limits to it, and update the tree by it.
         Lower[r,s] :=Upper[r,s] :=distance;
         tree :=UpdateTree( s,distance,tree );
      endfor
   endfor
endproc
```

The subprocedure `MakeTree` computes a shortest paths tree using the given root `r` and the given `Lower, Upper` limits, while `UpdateTree` updates the tree on setting the limits between the root and each other node `s` to the given distance; `Random` just returns a random number between its arguments.

Because the triangle inequality is a great deal more effective at reducing upper limits than it is at raising lower limits, the above restriction on the order causes those atoms that serve as the root of the tree early in the procedure to come out much closer together than those chosen later. By randomizing the order in which the atoms serve as tree roots, one can nevertheless avoid any net bias in the final set of structures. The cubic dependence of the time required for metrization on the number of atoms means that for very large systems (above ca. 1000 atoms) metrization can take a significant fraction of the overall time required to compute a conformational ensemble. It has, however, been observed that much of the gain can be obtained by using the procedure to choose only the distances between a few of the atoms and all others, and then filling in the rest of the distance matrix with independent random numbers between the resulting limits.[19]

Another interesting parameter to vary is the distribution with which the distances are chosen from within their limits during the process. In many cases, particularly when dealing with chain molecules and distance bounds consisting primarily of interatomic contacts (as in protein structure determination from NMR data), the distances in the resulting structures are strongly correlated with the triangle inequality upper limits. In such cases a bias towards the upper limits significantly improves the quality of the embedded coordinates, albeit at the expense of reduced sampling. This bias can be introduced by generating the distances with an exponential distribution, whose mean value is obtained from the lower and upper limits via the formula

$$\overline{d} = ((1-\beta)\overline{\ell}^{\,\alpha} + \beta\overline{u}^{\,\alpha})^{1/\alpha} \tag{43}$$

in which one typically sets $\beta = 1/2$ and $\alpha = 4$. As a general rule, one has to find a compromise between the quality of the individual coordinate sets and the sampling as a whole.

3.2.2 *Embedding*

The key to the EMBED algorithm lies in the fact that coordinates that are a certain best-fit to the estimated distances obtained via metrization can be found rapidly and reliably by eigenvalue methods, with no problems at all from local minima. The most obvious way to fit coordinates to distances is to minimize either the so-called STRESS

$$\sum_{1=i<j}^{N,N} (w_{ij}(\|\mathbf{x}_i - \mathbf{x}_j\| - d_{ij}))^2 \tag{44}$$

or else the smoother SSTRESS

$$\sum_{1=i<j}^{N,N} (w_{ij}(\|\mathbf{x}_i - \mathbf{x}_j\|^2 - D_{ij}))^2 \tag{45}$$

with respect to the coordinates $\mathbf{x}_1, \ldots, \mathbf{x}_N$, where $D_{ij} = d_{ij}^2$ are the estimated squared distances and the $w_{ij} \geq 0$ are weights. Although some nice characterizations of the stationary points of these functions are known,[17,20] no fast or reliable algorithm for finding their global minima is available.

If we expand the SSTRESS as

$$4\sum_{1=i<j}^{N,N} w_{ij}^2((\mathbf{x}_i \bullet \mathbf{x}_j) - \frac{1}{2}(\mathbf{x}_i^2 + \mathbf{x}_j^2 - D_{ij}))^2 \tag{46}$$

we see that it can be regarded as a weighted sum of squares of the differences between the inner products calculated from the coordinates and an estimate thereof, namely

$$\tfrac{1}{2}(\mathbf{x}_i^2 + \mathbf{x}_j^2 - D_{ij}) \tag{47}$$

In order to obtain an estimate that is independent of the coordinates we are trying to calculate, we use the following formula to calculate the squared distances to the center of mass from the squared distances among the points, i.e.,

$$D_{0i} = M^{-1}\sum_{j=1}^{N} m_j D_{ij} - M^{-2}\sum_{1=j<k}^{N,N} m_j m_k D_{jk} \tag{48}$$

where the m_j are the masses of the points, and $M = \sum_j m_j$. It is easily shown that when the estimated distances among the points are exact, then so are the distances to the center of mass, i.e., for center of mass coordinates,

$$D_{ij} = (\mathbf{x}_i - \mathbf{x}_j)^2 \Rightarrow D_{0i} = \mathbf{x}_i^2 \qquad \text{for all } i, j \tag{49}$$

Because it involves an averaging process, the above equation for the squared distances to the center of mass D_{0i} has the important feature of being very tolerant to errors in the estimated squared distances D_{ij}.

We now restrict ourselves to weights of the form $w_i w_j$, and consider the problem of minimizing the so-called STRAIN

$$\frac{1}{2}\sum_{i,j=1}^{N,N} (w_i w_j(\mathbf{x}_i \bullet \mathbf{x}_j - a_{ij}))^2 \tag{50}$$

where $a_{ij} = \frac{1}{2}(D_{0i} + D_{0j} - D_{ij})$. If we let $\mathbf{A} = [a_{ij}]$ and $\mathbf{W} = \mathbf{diag}(w_1, \ldots, w_N)$ be a diagonal matrix of weights, the STRAIN can be expressed as a squared Frobenius norm, i.e.,

$$F(\mathbf{X}) = \tfrac{1}{2}\|\mathbf{W}(\mathbf{X}\mathbf{X}^{\mathrm{T}} - \mathbf{A})\mathbf{W}\|^2 \tag{51}$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^{\mathrm{T}}$ is an $N \times 3$ matrix of coordinates. It can be shown that the matrix $\mathbf{A}$ is related to $\mathbf{D} = [D_{ij}]$ by a two-sided projection, namely

$$\mathbf{A} = -\tfrac{1}{2}(\mathbf{I} - \mathbf{1}\mathbf{m}^{\mathrm{T}}/M)\mathbf{D}(\mathbf{I} - \mathbf{m}\mathbf{1}^{\mathrm{T}}/M) \tag{52}$$

where $\mathbf{1} = [1, \ldots, 1]^{\mathrm{T}}$, $\mathbf{m} = [m_1, \ldots, m_N]^{\mathrm{T}}$, and $\mathbf{I}$ is the identity matrix; the matrix $\mathbf{A}_X = \mathbf{X}\mathbf{X}^{\mathrm{T}}$ is similarly related to $\mathbf{D}_X = [(\mathbf{x}_i - \mathbf{x}_j)^2]$. Thus the STRAIN can be written as

$$F(\mathbf{X}) = \tfrac{1}{2}\|\mathbf{P}(\mathbf{D}_X - \mathbf{D})\mathbf{P}^{\mathrm{T}}\|^2 \tag{53}$$

where $\mathbf{P} = \mathbf{W}(\mathbf{I} - \mathbf{1}\mathbf{m}^{\mathrm{T}}/M)/\sqrt{2}$. In this sense it represents a direct fit to the (squared) distances.[21,22]

If we change variables to $\mathbf{Y} = \mathbf{W}\mathbf{X}$ and let $\mathbf{B} = \mathbf{W}\mathbf{A}\mathbf{W}$, the STRAIN can also be written as

$$F(\mathbf{Y}) = \tfrac{1}{2}\|\mathbf{Y}\mathbf{Y}^{\mathrm{T}} - \mathbf{B}\|^2 \tag{54}$$

The gradient of the STRAIN, of course, must vanish at its global minimum. If we arrange this gradient in matrix form, we obtain

$$[\partial F/\partial y_{ij}] = (\mathbf{Y}\mathbf{Y}^{\mathrm{T}} - \mathbf{B})\mathbf{Y} = \mathbf{0} \tag{55}$$

or

$$\mathbf{B}\mathbf{Y} = \mathbf{Y}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y}) \tag{56}$$

where $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ is a 3×3 matrix that we will call the inertial tensor. Without loss of generality, we can assume that the coordinates are rotated in space so that the inertial tensor is diagonal, i.e.,

$$\mathbf{Y}^{\mathrm{T}}\mathbf{Y} = \mathbf{diag}(\lambda_1, \lambda_2, \lambda_3) \tag{57}$$

Then if $\mathbf{Y}_i$ is the ith column of the scaled coordinate matrix $\mathbf{Y}$, we have $\mathbf{B}\mathbf{Y}_i = \lambda_i \mathbf{Y}_i$ for $i = 1, \ldots, 3$. It follows that these columns are proportional to eigenvectors of the scaled estimated Gram matrix $\mathbf{B}$, while the moments of inertia $\lambda_1, \lambda_2, \lambda_3$ are the corresponding eigenvalues. Since the eigenvectors have unit norm, the diagonal form of the inertial tensor implies that the constant of proportionality is $\sqrt{\lambda_i}$.

We now expand the squared Frobenius norm as follows:

$$\begin{aligned} F(\mathbf{Y}) &= \mathrm{tr}((\mathbf{Y}\mathbf{Y}^{\mathrm{T}} - \mathbf{B})^2) \\ &= \mathrm{tr}(\mathbf{B}^2 - 2\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{B} + \mathbf{Y}\mathbf{Y}^{\mathrm{T}})^2) \\ &= \mathrm{tr}(\mathbf{B})^2 - \mathrm{tr}(2\mathbf{Y}^{\mathrm{T}}\mathbf{B}\mathbf{Y} - (\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^2) \end{aligned} \tag{58}$$

The foregoing considerations imply that at any stationary point of the STRAIN, we have

$$\mathbf{Y}^{\mathrm{T}}\mathbf{B}\mathbf{Y} = (\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^2 = (\mathbf{diag}(\lambda_1, \lambda_2, \lambda_3))^2 \tag{59}$$

and hence

$$F(\mathbf{Y}) = \mathrm{tr}(\mathbf{B}^2) - \lambda_1^2 - \lambda_2^2 - \lambda_3^2 \tag{60}$$

It follows at once that among all stationary points, the global minimum of the STRAIN is characterized by the three eigenvalues of the matrix $\mathbf{B}$ of largest magnitude. Of course, we have assumed that our inner product is positive-definite, which holds only if these eigenvalues are nonnegative. Therefore, the global minimum $\mathbf{Y}$ subject to this condition is obtained by taking the three largest nonnegative eigenvalues of $\mathbf{B}$, and scaling the corresponding eigenvectors by their squareroots. These are then scaled back to the original coordinates $\mathbf{X} = \mathbf{W}^{-1}\mathbf{Y}$.

It is important to note that a variety of iterative methods exist by which the three largest eigenvalues of a symmetric matrix can be rapidly found without fully diagonalizing it. The simplest, called the power method, consists of little more than taking a random unit vector, then iteratively multiplying it by the matrix and renormalizing until convergence. The eigenspace is then subtracted from the matrix and the process repeated to get the next largest eigenvalue/vector. Providing that the three largest eigenvalues differ from one another by some fixed percentage (in practice, a few percent is sufficient), this procedure requires time proportional to only the square of the number of atoms (i.e., the size of the Gram matrix). This assumption seems to hold quite well in most chemical applications, and even if nearly degenerate eigenvalues are encountered, the solution that is obtained on terminating prior to convergence is still quite good. In the case of perfectly degenerate eigenvalues, the procedure just converges to an arbitrary eigenvector within the degenerate eigenspace, and the optimum is still obtained. For these reasons we have never bothered to implement eigenvalue procedures with higher-order convergence, especially since each iteration of these more complex procedures generally requires time proportional the cube of the number of atoms. As it stands, embedding requires only a very small fraction of the time needed for

coordinate refinement, so there is little point in trying to make it faster.

3.3 Coordinate Refinement

The coordinates obtained from metrization and embedding, although they fit the distance constraints well in the root-mean-square sense, will generally have significant constraint violations in them. This is particularly noticeable for the covalent constraints such as bond lengths, where a violation of anångstöm has disastrous energetic consequences. In addition, the chirality constraints are completely ignored during embedding, and while the overall handedness can be corrected by a simple reflection operation, the relative handedness of various parts of the system are generally also not consistent with the chirality constrants. Correction of these violations is done by minimizing an error function, which measures the total violation of both the distance and chirality constraints. Even though these functions generally have numerous local minima, the coordinates obtained by embedding are good enough to enable solutions to be found with simple descent-based minimization algorithms even on problems with as many as a hundred atoms. In larger problems such as proteins, it is still necessary to resort to global minimization procedures like simulated annealing, but even in these cases the ready availability of the good starting structures produced by embedding makes the job much easier. In this section we shall first describe the usual error functions used in distance geometry calculations, and subsequently some of the more common methods of minimizing them.

3.3.1 Error Functions

An error function[23] is a real-valued function of the atomic coordinates that is:

(1) nonnegative;
(2) translation and rotation independent;
(3) everywhere at least twice differentiable;
(4) zero if and only if all of the geometric constraints are fully satisfied.

The following is the error function favored by this author for the distance constraints:

$$E_{\mathrm{d}}(\mathbf{X}) = \sum_{\{i,j\}} \max^2\left(0, \frac{(\mathbf{x}_i - \mathbf{x}_j)^2 - u_{ij}^2}{\varepsilon_u^2 + u_{ij}^2}\right) + \sum_{\{i,j\}} \max^2\left(0, \frac{l_{ij}^2 - (\mathbf{x}_i - \mathbf{x}_j)^2}{\varepsilon_l^2 + (\mathbf{x}_i - \mathbf{x}_j)^2}\right) \tag{61}$$

The symbol $\{i, j\}$ indicates that the sums are only over those pairs of atoms for which explicit distance constraints exist. Since in most problems the distance constraints are very sparse, such an error function can be computed much more rapidly than semi-empirical energy functions. The parameters $\varepsilon_l, \varepsilon_u$ are made large enough to avoid having any single term become much bigger than the others when the distance or upper bound is very small, respectively. Each term in this error function is called a distance restraint.

The exception to constraint sparsity is of course the hard-sphere lower bounds, which apply to most pairs of atoms. Because the sphere radii tend to be small compared to the overall dimensions of the system, however, a list of all pairs of atoms that are close enough together to collide can be constructed quite rapidly. There exist a number of methods for doing this, but the simplest is a standard method for range query checking known as the plane sweep method. This may be summarized as follows:

```
procedure Sweep( Natom,Radius,Coord )
   for i from 1 to N do
comment: Store the endpoints & indices of their intervals.
      Term[2*i].value :=Coord[i,1] − Radius[i];
      Term[2*i+1].value :=Coord[i,1] + Radius[i];
      Term[2*i].index :=Term[2*i+1].index :=i;
   endfor
comment: Sort the list of endpoints by their values.
   SortByValue( Term );
comment: Scan the list of endpoints recording hits.
   for i from 1 to 2*N do
      if not OnList( Term[i].index ) then
         EnList( Term[i].index );
      else
         DeList( Term[i].index );
comment: All intervals now on the list overlap the i-th.
         while NextIn( other_index ) do
            NewHit( Term[i].index,other_index );
      endif
   endfor
endproc
```

The first loop records the left and right termini of the intervals obtained by projecting the spheres onto the first coordinate axis. This array of termini is then sorted by value, and scanned from left to right. Each time the left terminus of an interval is encountered, its index is stored on a list. Each time the right terminus is encountered, the index is removed from the list, and the list scanned. Every interval with an index on the list then overlaps with the delisted interval, and is recorded as a hit. In actual practice, pairs are only considered hits when the intervals obtained by projecting the spheres onto the second and third axes also overlap, so that each hit corresponds to the intersection of a pair of cubes parallel the coordinate axes, whose side lengths are twice the radii of the corresponding spheres. Only these pairs must subsequently be checked for hard-sphere overlaps. By using radii that are, e.g., 1 Å larger than the true radii, one can avoid having to rebuild the list until at least one atom has moved by that amount, albeit at the expense of having to check a slightly larger number of hits each time. The exact amount of extra radius that optimizes the overall speed depends on several factors, and is usually determined empirically.

We next turn our attention to the chirality error function. This is generally defined using the signed volumes, as follows:

$$E_{\mathrm{c}}(\mathbf{X} = \sum_{\{i,j,k,m\}} \max^2(0, \mathrm{vol}[\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_m] - u_{ijkm}) + \sum_{\{i,j,k,m\}} \max^2(0, l_{ijkm} - \mathrm{vol}[\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_m]) \tag{62}$$

Here, the function $\mathrm{vol}[\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_m] = (\mathbf{x}_j - \mathbf{x}_i) \bullet ((\mathbf{x}_k - \mathbf{x}_i) \times (\mathbf{x}_m - \mathbf{x}_i))$ is the signed volume, while l_{ijkm} and u_{ijkm} are lower and upper bounds on it. Since the chirality constraints by definition specify only the sign of the volume, these lower and upper bounds are derived from the distance constraints among the quadruple of atoms, by minimizing and maximizing the corresponding four-point Cayley–Menger determinant subject to those constraints. The absolute volume so obtained is given the sign of the chirality, if the chirality is nonzero. A chirality of zero, on the other hand, denotes a planarity constraint,

in which case the minimum absolute volume consistent with the distance constraints should also be zero, and the upper and lower bounds on the volume are the maximum volume allowed by the distance constraints and its negative, respectively. Bounds on the signed volumes can also be derived from ranges in the values of the torsion angles about rotatable bonds, and used to constrain these angles. In this case the torsion angles between all pairs of substituents, one on either end of the bond, should be constrained in order to obtain the desired range of rotomeric states in all cases.

3.3.2 Optimization Procedures

Despite a superficial resemblence to semi-empirical energy functions, it is much easier to locate a global minimum of the error functions used in distance geometry than it is to locate global energy minima. There are several reasons for this, of which the above-mentioned speed with which error functions and their gradients can be calculated is probably the least important. Unlike energy, where the primary criterion is to model the physics as accurately as possible, one has a great deal of lattitude in ones choice of error function, and we have utilized this lattitude so as to make it as smooth and well-behaved as possible. Thus the individual terms of the above distance error function contain no inflection points (as functions of the distances), and the intrinsic weighting of the terms tends to equalize their contributions. In particular, the terms due to the lower bound violations are bounded, and do not become infinite as the distance goes to zero. The force due to the upper bound violations, on the other hand, increases steadily as the violations increase, so that the gradient tends to point in the general direction of the global minimum even when one is far away from it. Last but not least, the value of the global minimum is known to be zero (assuming that the constraints are mutually consistent), so that one can at least tell when one has reached it!

The earliest applications of distance geometry to chemical problems used the conjugate gradients minimization algorithm.[23] This is a very robust and general-purpose minimization algorithm, which requires only first derivatives of the error, and is still probably the method of choice for solving small problems. Even though it is difficult to attain true quadratic convergence with the conjugate gradients method, it has been found to perform well even when far from a minimum. Higher-order methods relying on second derivative information have not been used as extensively, primarily because the starting coordinates obtained from embedding are usually sufficiently far away from the minimum that quadratic convergence does not set in for a very long time, during which one has the additional overhead of calculating second derivatives and solving Newton's equations. Among the numerous variations on the conjugate gradients algorithm that we have tried, the best performance appears to be obtained from Shanno's version.[24]

In computational chemistry generally, minimization algorithms like conjugate gradients tend to perform best when the torsion angles about single bonds are used as the variables, instead of the Cartesian coordinates of the individual atoms, for the simple reason that the number of torsion angles generally runs about one tenth the number of Cartesian coordinates. The variable target function method uses the conjugate gradients algorithm with torsion angles to solve distance geometry problems, starting from random conformations.[25,26] This method was designed for the specific problem of fitting polypeptide chains to the distance and torsion angle constraints that are available from NMR spectroscopy (see below). It operates by first minimizing with respect to the 'short-range' constraints connecting atoms separated by at most one amino acid residue, and gradually adding on longer and longer-range constraints until all the distance constraints are included.

When it works, the variable target function method is quite fast, but how well it works depends strongly on the way the constraints are distributed along the chain. For example, much better results are obtained with proteins that are primarily alpha helical than with proteins that contain large amounts of beta sheet. The method also does not extend readily to other kinds of molecules, particularly those containing complicated flexible ring structures. Unfortunately, the EMBED algorithm cannot be used to get good initial torsion angles from which to start the optimization, because the distorted covalent geometry in the embedded structures renders the torsion angles in them meaningless. The lack of a fast and general method of finding good initial torsion angles, together with the problems of handling ring molecules, has limited the scope of the applications of torsion angle based algorithms to distance geometry. Recent programs for performing dynamical simulated annealing (below) using torsion angles should nevertheless significantly improve this situation.[27,28]

Presently, the most reliable method of fitting coordinates to distance and chirality constraints is dynamical simulated annealing.[29–31] In this method one performs a molecular dynamics simulation in which one treats the error function as if it were the energy, and gradually cools the system down, starting from a high temperature. This tends to drop one into a good, in fact often global, minimum of the error. Because one does not need to be physically realistic, one can make all the atomic masses equal, and use the largest time step consistent with numerical stability. This, together with the above-mentioned smoothness of the error function, enables one rapidly to make very large changes to the coordinates. The amount by which the individual atoms move on each time step, in fact, can exceed 1 Å.

With appropriate parameters and a sufficiently long cooling period, the convergence obtained by this procedure can approach 100%.[32] It is even possible to obtain good convergence starting from random coordinates for large molecules such as proteins.[33] The amount of time required to obtain a given convergence ratio, however, is substantially less when the starting coordinates are obtained from the EMBED algorithm. In addition, there is always a compromise between the failure rate and the amount of time spent annealing each structure. Generally speaking, the overall time required to obtain a given number of converged conformations is minimized at about an 80% convergence ratio.

One interesting trick, which significantly improves the convergence ratio obtained for a given computational investment with both conjugate gradients as well as simulated annealing, involves using four-dimensional coordinates. Minimization in higher dimensions was first pioneered by G. M. Crippen, using simple potential functions on reduced polypeptide models.[34] Subsequently, several groups found that by embedding four-dimensional coordinates and adding a dimensionality error, given by the sum of the squares of the fourth coordinates, to

the error function, one could avoid being trapped in certain kinds of local minima.[30–32] Of course, the final structures that converge to zero error are assured of being three-dimensional, by virtue of the added dimensionality error.

A more recent variation on this procedure developed by the author, known as the WARP procedure, uses conjugate gradients to minimize the three-dimensional conformation obtained by projecting a fixed four (or higher)-dimensional structure into three dimensions, where the entries of the 4×3 matrix which maps the four-dimensional structure into three dimensions are the variables for the minimization. Using four-dimensions, one obtains only about a factor of two improvement in the error, but the coordinate change can be quite large. As one extends this procedure to higher dimensions, one uses more variables, reaching all $3N - 3$ internal and rotational degrees of freedom at $N - 1$ dimensions. Thus this procedure gives us a systematic way of defining a nested sequence of subspaces of the internal configuration space, in a fashion reminscent of normal modes. It remains to be seen if this may have wider applicability in, e.g., energy minimization problems.

4 APPLICATIONS

The utility of distance geometry depends on discovering innovative ways of formulating conformational problems in terms of distance and chirality constraints. In this section we describe several chemical applications and their associated problem formulations, which demonstrate that the distance geometry approach is surprisingly general even though it relies upon simple 'black-and-white' constraints rather than 'grey-scale' probabilities derived from, e.g., a semi-empirical energy function. This generality is one of distance geometry's most important assets: once one has mastered this one tool, one is ready to solve a wide variety of problems. It is also worth noting that the 'black-and-white' model enables one to answer some questions that would be difficult or impossible to answer with a probablistic model. The reason is that given just about any set of consistent distance constraints, the distance geometry algorithms described above are reliable enough to enable one to find a conformation that satisfies them all. Thus if one is unable to find such a conformation, one can be reasonably confident that ones geometric assumptions (constraints) are incorrect, and providing there are not too many errors in the constraints, one will generally also be able to figure out what is wrong. In energy minimization, by contrast, it is often difficult just to tell if something is wrong (e.g., one is trapped in a local minimum, the energy function parameters do not extend to the molecule of interest, etc.). Distance geometry is much more than molecular modeling with a simplified energy function!

4.1 Conformational Analysis

It is, of course, straightforward to describe the covalent structure in terms of distance and chirality constraints. The bond lengths are all constrained to their standard values, while the bond angles can be fixed by constraining the corresponding geminal distances. The vicinal distances across rotatable bonds are usually set to their *cis*/*trans* limits, to enable free rotation, and all the distances within any known rigid group of atoms (e.g., phenyl rings) are constrained to their known values. Hard sphere lower bounds are imposed on all other distances, where the sphere radii are generally chosen about 10% below the van der Waals radii (or determined empirically from crystal packing studies). Finally, chirality constraints should be imposed on every rigid quadruple of atoms in the molecule, even if it is not chiral in the usual chemical sense, in order to enforce rigidity of these groups during optimization. This is particularly important for planar groups of atoms, since to a first-order approximation the distances are independent of out-of-plane distortions.

4.1.1 General Methods

All of these constraints are readily generated from covalent connectivity tables and a geometric database. Thus one obvious application of distance geometry is as a means of constructing three-dimensional molecular models from connectivity data, either entered interactively from a graphical monitor, or else contained in a database of chemical compounds. Programs that use distance geometry to accomplish these tasks are commercially available, and in widespread use. The RUBICON program from Daylight Chemical Information Systems is of particular interest, because it contains a user extensible language for generating constraints from substructures. The further development and standardization of such languages promises to broaden substantially the scope of the applications of distance geometry.

One advantage of the distance geometry approach to model building is that the conformers are assured of having no broken rings or long-range van der Waals clashes between atoms. Another advantage is that, when the molecule is flexible, one can easily generate a large number of random conformers, rather than just one, simply by using a different random number seed during metrization for each calculation. The random nature of the sampling sometimes leads to the discovery of surprising conformational possibilities. Moreover, in combination with conformational metrics such as the RMSD and simple clustering algorithms, one can select a small but diverse set of conformers spanning all of conformation space from such a conformational ensemble. The earliest example of this approach was a conformational analysis of cycloalkane, crown ether, and steroid rings.[35] Other examples, including several cyclic oligopeptides, may be found in a recent review.[36]

By imposing additional constraints, it is often possible to direct the search to particularly interesting regions of the conformation space. For example, if one believes that a certain hydrogen bond is important, it is easy to constrain the donor–acceptor pair to be together in all members of the ensemble. One can even systematically enumerate all possibilities according to some preconceived classification scheme. The most interesting case here is probably to systematically impose ranges on the torsion angles that cover all possible rotomeric states of the molecule, and see which ones can be built with the desired properties. Unlike torsion angle grid searches, this procedure involves only one calculation per rotomer, and so may be useful in cases where grid searches could only be performed with a grid too coarse to catch all sterically allowed conformations. Another classification scheme, involving interatomic contacts, has been used to search for the global minimum of a square-well energy function,[37] in a fashion similar to more recent protein threading algorithms.[38]

4.1.2 Special Cases

One nice feature of distance geometry is that, in special cases, it is possible describe the entire conformation space, either by enumeration if it is discrete, or by finding an explicit parametrization if it is continuous. The first, and perhaps still most elegant, such analysis was carried out on cyclohexane by Andreas Dress. By expanding the five- and six-point Cayley–Menger determinants as functions of the three variable vicinal distances, he was able to prove rigorously that the conformation space of this molecule consists of a rigid chair form as well as a flexible 'circle' of conformers related to the boat form by pseudo-rotation. This work was never published, but an account may be found in Ref. 1. To date, this mathematical analysis has not been successfully extended to any larger cycloalkanes, although a combination of symbolic and numerical calculations have succeeded in characterizing the conformation space of cycloheptane.[39]

Another case in which it has proven possible to enumerate all conformations, and which is of particular interest in restoring ring closure after changing its torsion angles, is known as the local deformation problem. In this problem, originally studied by Gõ and Scheraga,[40] one is given a chain of six single bonds in some standard (e.g., tetrahedral) geometry, and seeks values for the six torsion angles such that the net rotation and translation places a rigid body attached to the last bond in a given position and orientation with respect to a rigid body attached to the first bond. These conditions imply that the positions and orientations of the first and last bonds themselves are fixed, and hence the four distances between the atoms of the two bonds, together with their chirality, are likewise fixed. By breaking the chain at the middle atom and duplicating it, we can rephrase the problem as a problem of satisfying four distance constraints with respect to only four of the six bond rotations. Figure 5 illustrates this situation.

Using geometric algebra to describe the bond rotations, together with a symbolic algebra package for Gibbs' vector algebra, it was possible to derive closed form expressions for these squared distances indicated with dashed lines in Figure 5 as functions of the tangents of the half-angles of the bond rotations. Setting these expressions to the known values of the squared distances then gives us four equations, three of which depend on only two of the tangents each, while the fourth (that which is a consequence of the vanishing of the distance between the duplicate atoms) depends on all four tangents simultaneously. All four equations are only quadratic in each tangent separately, and their coefficients can be expressed as polynomials in the squared distances and signed volumes, in accord with the basic theory of distance geometry. It is possible to eliminate all but one of the tangents from these equations, and so reduce the problem to a univariate one, which can then be solved by grid search and interpolation. The result is the first new algorithm for the local deformation problem to be developed in over 20 years.[41]

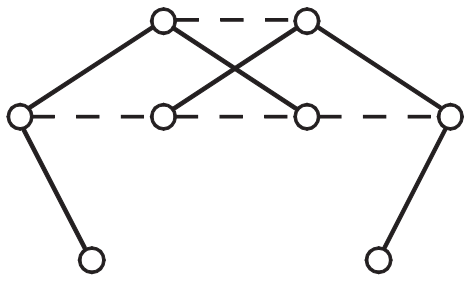

Figure 5 The two chains obtained by fixing the positions and orientations of the first and last bonds in the original chain, breaking it at the middle atom, and duplicating that atom. The solid lines indicate chemical bonds, while the dashed lines indicate the distance constraints that the bond rotations must satisfy. In particular, the distance between the duplicates of the middle atom must be zero in any valid solution to the local deformation problem

4.2 Drug Design

Distance geometry has been applied to drug design in two distinct ways. The first is as a means of docking ligands into binding pockets on receptor proteins of known structure, while the second is as a means of identifying the 'pharmacophore' in a series of active analogues even in cases in which the receptor structure is not known. As always, the advantages of distance geometry in these applications include speed, the ability to incorporate diverse types of information, and the possibility of falsifying incorrect assumptions.

4.2.1 Ligand Docking and Screening

Distance geometry was first used as an adjunct to computer graphics methods to fit ligands into binding sites by Jeff Blaney, as reviewed in Ref. 36. In this work, one typically hypothesizes a few energitically favorable ligand–receptor contacts, and attempts to generate randomly a number of models of the complex in which these contacts occur. Of course, the entire protein structure does not need to be calculated, but only those atoms in the binding site. As a rule, the binding site geometry is fixed by constraining all of the distances between its atoms to their values in the crystal structure of the receptor, although it is not difficult to allow some of the binding site atoms to vary along with the ligand atoms. The resulting models are then evaluated and ranked by energy minimization and computer graphics.

This approach works best if the binding site is a deep pocket. The packing constraints, i.e., the small but numerous lower bounds due to the hard-sphere radii, then play a major role in determining the ligand conformation. The joint effects of all the distance constraints together is difficult to predict even with the aid of computer graphics, and can lead to a surprising degree of conformational confinement. The packing constraints may even turn out to be incompatible with the hypothesized contacts, as indicated by repeated failures to find a conformation that satisfies all the constraints together. In this case one is forced to consider new patterns of contacts, or binding modes. This provides an effective screen for a chemist's intuition, and can lead to the discovery of binding modes that were not originally anticipated.

Using the EMBED algorithm, this approach has been used to predict binding modes in a wide variety of ligand–receptor systems, including phenylhippurate–chymotrypsin, phospholipid–phospholipase-A2, and bis-acridine–double helical DNA.[36] A similar approach has also been demonstrated using an unusual optimization algorithm called the ellipsoid algorithm, together with the torsion/Euler angles as the conformational variables, which have the advantage of intrinsically keeping the active site rigid.[42] More recently, an extension of the EMBED algorithm has been developed that also keeps the

coordinates of a subset of the atoms fixed, and promises to be very useful in such problems.[43]

Another interesting extension of the original idea is to propose not just one binding mode, but rather a collection of energetically favorable interactions between ligand and site atoms. The set of all pairs of contacts that are compatible with the triangle inequality can be represented by a graph, whose adjacent vertices correspond to compatible pairs. The maximal cliques (i.e., sets of mutually adjacent vertices) in this docking graph provide one with a large number of possible binding modes, which can be filtered versus several simple screens. The distance bounds implied by these cliques, together with the covalent structure of the ligand and known spatial structure of the receptor protein, are used as input for the EMBED algorithm. This further screens the cliques for geometric feasiblity, and at the same time generates structures for additional refinement. The viability of this largely automatic approach to discovering binding modes has been demonstrated using complexes of dihydrofolate reductase with Baker's triazines.[44]

A recent variation on this approach represents the binding site by a collection of overlapping spheres, and attempts to dock the ligand into the site such that each atom of the ligand is contained in at least one site sphere. This involves a novel disjunctive form of constraint, which cannot yet be incorporated into the EMBED algorithm and must be handled entirely by the challenging minimization of an appropriate error function.[45] In principle, these distance geometry approaches to ligand docking can also be used to screen entire three-dimensional databases for ligands that fit a known binding site, in an effort to find new lead compounds. Because it allows full conformational flexibility, the distance geometry approach is less efficient than more specialized methods which assume that the ligand is rigid or allow it at most only limited flexibility.[46] As computers continue to become faster, however, such applications may prove easier to carry out.

4.2.2 Pharmacophore Identification and Site Mapping

A more challenging class of problems arises when the structure of the receptor protein is not known, and one wants to identify the common structural features in a series of analogous active ligands, with the goal of finding new ligands that bind better. With the assumption that all the ligands bind to the same site and with the same binding mode, this leads to the concept of the pharmacophore. The first distance geometry method for pharmacophore identification was developed by Scott Dixon and co-workers.[47] In this elegant method, one proposes a common set of atoms that are present in all the ligands, and which are supposed to be involved in strong interactions with the receptor protein. The assumption of a common binding mode then implies that each ligand can assume a conformation such that corresponding atoms, one from each ligand's set, can all be simultaneously located in the same small region of space.

The search for such a set of conformations can be formulated as a distance geometry problem as follows. First, one imposes small upper bounds on the distances between pairs of corresponding atoms, one from each ligand, of ca. 1 Å. Next, one allows the ligands to pass through one another, by resetting all the hard-sphere lower bounds between different ligands to zero. Then any set of ligand conformations that satisfies these (physically unrealistic!) constraints is likely to contain the active conformations of the ligands when they are bound to the receptor. Moreover, if such a family of conformers cannot be found, then either the proposed pharmacophore was incorrect, or else one or more of the ligands binds with a different mode. This method was tested on a series of four flexible nicotinic receptor agonists using a three-atom pharamacophore, with results that are compatible with earlier results derived using only rigid ligands.[47]

A much more ambitious goal attempts to find not merely a pharmacophore, but also a model for the binding site together with a set of binding modes and interaction energies, that can at least qualitatively account for the binding affinities of a potentially diverse set of ligands.[48] In the simplest version of this approach, one models the binding site geometry by a collection of spheres, representing the domains of influence of functional groups within the site. The binding modes may then be enumerated in a fashion analogous to that described above for the case when the true site geometry is known, and interaction energies can be assigned to the occurrence of each ligand atom in each site sphere, such that the lowest energy binding modes parallel the observed binding affinities. The site model/binding modes can then be tested by trying to construct conformations for the ligands that place their atoms in the correct spheres, using the EMBED algorithm. Recent work along these lines is directed towards the automatic construction of more sophisticated site models, involving an interesting generalization of distance bounds to 'intervals of bounds'.[49]

4.3 NMR Spectroscopy

The best-known application of distance geometry is as a means of determining the solution conformations of small biological macromolecules from NMR data. The only other method of determining these structures at atomic resolution, X-ray crystallography, must necessarily observe these molecules in an unnatural crystalline environment, and cannot be used in many cases because the molecule cannot be crystallized. A more detailed account of the application of distance geometry to NMR data may be found in the author's article in the companion 'Encyclopedia of Nuclear Magnetic Resonance'.[50] For completeness' sake, we nevertheless include a brief summary of the main ideas here. Because the number of applications of distance geometry to NMR data now numbers in the hundreds, no attempt will be made to include even representative literature citations.

4.3.1 The Nature of the Data

The most important geometric information that is available from NMR spectroscopy comes in the form of contacts between pairs of hydrogen atoms, i.e., upper bounds on their distances of 5 Åor less. This information is obtained from a two-dimensional spectrum called NOESY, whose diagonal corresponds to the usual one-dimensional spectrum, and whose cross-peaks occur at the frequency coordinates of spatially proximal pairs of protons. A less important, but still very useful, type of information consists of bounds on the torsion angles about single bonds. This information may be obtained from a variety of NMR spectra, most notably two-dimensional COSY spectra, and can be represented by a combination of constraints on the vicinal distances and corresponding signed volumes as previously described.

There are several potentially serious problems involved in the interpretation of NMR data in terms of distance constraints. First, a phenomenon called spin diffusion may result in spurious NOESY cross-peaks between protons that share a common neighboring proton, but which themselves are greater than 5 Å apart. Second, biological macromolecules are always flexible to at least some degree, and NOESY cross-peak intensities are expected to reflect the inverse average sixth root of the interproton distances. As a result, it is entirely possible for a proton to appear to be adjacent to two other protons simultaneously, when the other two protons are nevertheless always greater than 10 Å apart. Finally, NOESY cross-peak intensities are affected by numerous spectral artefacts and well as by other sources of relaxation, which may cause many cross-peaks to be missing.

The fact that high-resolution protein structures are routinely calculated from NMR data with no significant residual constraint violations in them is a strong indication that these problems are not serious in most cases. One should nevertheless always evalulate the convergence of the calculations, and check the residual violations carefully against the data, in order to be sure that such a problem has not occurred. Although distance geometry is capable in principle of representing the actual range of conformations present in solution, the possibility of missing data means that one cannot automatically assume that any region of the structure that is not well-defined in the final conformational ensemble is also disordered in solution, at least without further evidence. These problems are obviously not problems in distance geometry per se, but rather in the interpretation of the data. Fortunately, NMR is also capable of providing direct information on conformational flexibility, in the form of relaxation or chemical exchange studies. For further details, the reader is referred to Ref. 51.

4.3.2 Computational Issues

The first problem that must be solved in order to determine a conformation from NMR data is to assign the individual peaks in the NMR spectrum to the corresponding atoms in the molecule. This assignment problem is usually solved by a combination of spectroscopic techniques, which need not concern us here. It is interesting to note, however, that several attempts have been made to solve some or all of the assignment problem using distance geometry.[52,53] The approach we favor takes two copies of the molecule, and applies only the covalent constraints to the first, and only the NOESY constraints to the second (assuming some arbitrary assignment compatable with whatever is known at the time). One then defines an error function that includes terms that are zero whenever any compatible pair of atoms, one from each copy, have the same spatial positions (e.g., the methyl groups in any pair of alanines). These disjunctive constraints are similar to those mentioned above for ligand docking, and finding spatial structures consistent with such constraints likewise poses a challenging minimization problem. Analogous disjunctive constraints have also been used to resolve ambiguities between intra- and intermolecular distances in C_2-symmetric complexes.[54]

Even when all the assignments can be unambiguously determined, the problem of computing the conformation of biological macromolecules from NMR data remains difficult for two reasons. One is the sheer size of the molecules involved, which often exceeds 1000 atoms even with suitable united atom approximations. The second lies in the sparsity of the distance information, usually covering less than 1% of the million or more different distances in such large molecules; this is a consequence of the fact that one seldom has any information at all on distances much greater than 5 Å. Since triangle inequality bound smoothing is not very effective at propagating these sparse constraints to all the distances, particularly their lower limits, the distance matrices obtained from metrization are also not very good approximations to the distance matrices of conformations compatible with all the data. As described above, this problem can be alleviated by biasing the distances chosen during metrization towards their upper limits, thereby correcting for the lack of large lower limits. Nevertheless, a strong reliance on simulated annealing is unavoidable, if good conformations are to be found reliably.[29,55,56]

Because of this lack of information on the longer distances, many researchers once doubted that the distance constraints available from the NOESY experiment would be sufficient to determine a reasonably precise conformation for a globular protein. As a result, extensive simulations had to be performed to demonstrate the surprising power that a relatively small number of contacts has, when combined with the covalent geometry and the ubiquitous packing constraints.[57,58] These expectations were validated when the first solution conformation of a complete protein was determined from NMR data shortly thereafter (see Figure 6).[59] With current NMR methods for elucidating protein structure, upwards of 20 interproton contacts may be obtained per residue on the average, and the uncertainties in the relative positions of the atoms in the resulting structures can average less than 1 Å.[51,60] The situation in the determination of nucleic acid structures is quite different, since these structures tend to be more extended and to exhibit relatively few long-range interproton contacts. With such molecules it is important to collect and quantitate as many torsion angle constraints as possible, and the reliance on simulated annealing is likely to be even greater.

4.4 Homology Modeling

A more recently demonstrated biological application of distance geometry is as a means of automatically constructing model protein structures from sequence alignments with homologues of known structure, a task known more generally as homology modeling. The usual way of doing this, which we call the 'cut and paste' method, involves aligning the structurally conserved regions of the homologues in space, and then copying the coordinates of the backbone atoms from selected structures, along with any conserved sidechain coordinates, and using these directly as the coordinates of those atoms in the unknown structure. Thereafter, the loops and mutated sidechains are added, usually from a database of reference conformations, and steric clashes, broken bonds, and other energetically unfavorable interactions alleviated by energy minimization. In contrast, the distance geometry approach builds an entire family of possible conformations at once, including all the atoms together and free of any energetically disastrous defects. These may then be analyzed in order to determine which geometric features of the structure have been determined by the evolutionary and geometric hypotheses, and can also serve as diverse starting points for subsequent energy refinement.

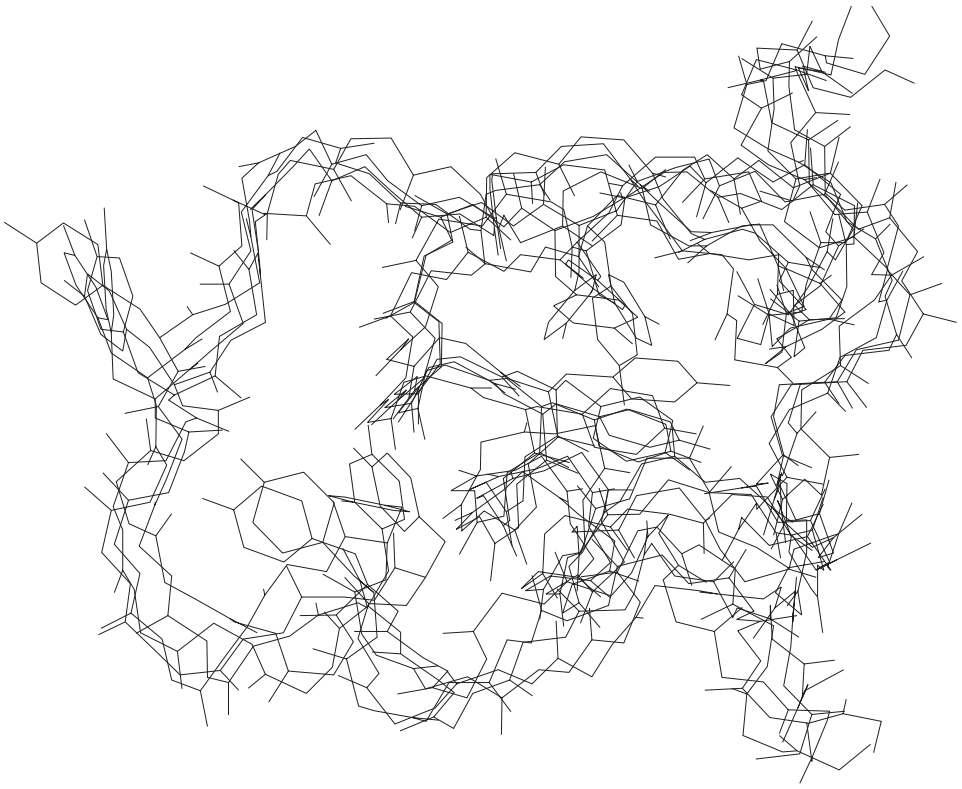

Figure 6 Illustration of a five-structure conformational ensemble for the protein BUSI. Only the backbone and aromatic sidechains are shown. This was the first complete protein structure to be determined in solution from NMR data[59]

4.4.1 The Computational Procedure

The computational procedure we developed for homology modeling consists of the following steps:[61]

(i) The sequences of the homologues of known structure are aligned with one another, usually so as to make it possible to superimpose the corresponding backbone atoms in their structures as closely as possible in space, rather than by the usual amino acid dissimilarlity measures.
(ii) The sequence of the unknown is added to this multiple sequence alignment, using amino acid dissimilarity measures, but permitting no insertions or deletions in the structurally conserved regions.
(iii) Distance constraints among the atoms of the unknown protein are derived by finding the minimum and maximum values of the corresponding distances in the homologues (where the correspondence is determined from the sequence alignment), expanding these ranges somewhat to be safe, and imposing them as bounds on the unknown.
(iv) Supplementary constraints, including hydrogen bonds and ranges of torsion angles, are derived from similar considerations.
(v) These constraints are combined with those that follow from the covalent structure of the unknown protein, and used as input for computing a conformational ensemble via the EMBED algorithm.

Steps (i) and (ii) are probably the most time-consuming steps, and require considerable subjective judgement. Serious errors in the alignment obtained in step (ii), which result in an inserted region that is not long enough to span the space between its adjacent structurally conserved regions, can be discovered during bound smoothing or from a convergence analysis of the structure calculations. In such cases one must go back and further adjust the alignments in light of this new knowledge.

Even when all the protons are deleted, the number of atoms in many proteins is too large to allow all the distances to be constrained. The majority of the distances chosen in step (iii) are therefore restricted to pairs of alpha carbons, since it is most important to get the backbone right. The bounds are obtained by expanding the minimum and maximum values of each distance over all the homologues, according to the formulae

$$\begin{aligned} \ell &= (d_{\max} + d_{\min})/2 - \rho(d_{\max} - d_{\min})/2 - \delta/n \\ u &= (d_{\max} + d_{\min})/2 + \rho(d_{\max} - d_{\min})/2 + \delta/n \end{aligned} \tag{63}$$

where $\rho > 0$ is a precision, $\delta > 0$ is a tolerance, and n is the number of homologues in which corresponding pairs of atoms occur. Additionally, constraints on the distances between the heavy atoms, beyond those determined by the covalent structure, in the same or adjacent amino acids in the sequence are derived by the same recipe (possibly with a different precision and tolerance). In the case that mutations have occurred between the unknown and a homologue, a fixed set of rules is used to decide how far out in the sidechain the constraints should go.

The torsion angle constraints are likewise derived by finding the range of torsion angle values that occur in the homologues, and expanding it by the same formulae. Once again, a fixed

set of rules is used to decide how far out in the sidechain to go in mutated residues, but in this case no constraint is imposed if the expanded range covers more than 180°. Finally, the homologues are energy minimized, their hydrogen bonds identified, and any that are common to all the homologues with a corresponding donor–acceptor pair in the unknown are imposed on the unknown by suitable distance constraints.

One key feature of the alpha carbon constraints is that they are uniformly spread across the entire structure, and include many large lower bounds. As a result, bound smoothing produces relatively good limits on all the distances, and hence the EMBED algorithm produces good initial coordinates. Although simulated annealing is still needed in order to satisfactorily fulfill all the constraints, a fairly rapid annealing schedule can be used with good final results. Even with a generous precision and tolerance, the above constraints are sufficient to determine the positions of the backbone atoms to within 1 Å, or about as well as a high-resolution NMR structure determination. The only exception occurs in large insertions, which do not occur in any of the known homologous structures. There the procedure effectively does a random search of the possibilities that are sterically and covalently consistent with the well-defined regions of the structure. Of course, the accuracy of the final structures can only be as good as the structural conservation of the unknown relative to its homologues!

4.4.2 A Case Study with E. coli Flavodoxin

A simplified version of the above procedure was originally evaluated on the Kazal family of trypsin inhibitors.[62] The first full-scale test was carried out on the Flavodoxin from *E. coli*.[61] This protein had no crystal structure at the time the calculations were performed, but structures were available for four homologous Flavodoxins, namely *A. nidulans*, *C. beijerinckii*, *C. crispus*, and *D. vulgaris*. The sequence alignment used to generate the constraints is shown in Figure 7. It may be observed that two of the homologues contained large deletions with respect to the other three Flavodoxins, and that the *E. coli* form has an additional six amino acids on its C-terminus, which are not found in any of the homologues. The percent identities between the *E. coli* sequence and those of its homologues in this alignment were 44%, 16%, 33%, and 23% for *A. nidulans*, *C. beijerinckii*, *C. crispus*, and *D. vulgaris*, respectively.

A total of 13 043 alpha carbon constraints, 8893 local heavy atom constraints, 452 torsion angle constraints and 116 hydrogen bond constraints were derived from this alignment. The ten structure conformational ensemble computed from these constraints had an average alpha carbon RMSD between all pairs of structures of 0.85 Å (excluding the first three and last six residues); the corresponding average for all heavy atoms was 1.60 Å. After a careful restrained energy minization using solvated molecular dynamics, these numbers increased to 1.00 and 1.66 Å, respectively. A preliminary crystal structure of the *E. coli* form at 2.5 Å resolution was subsequently obtained courtesy of David Hoover and Martha Ludwig at the University of Michigan. The average RMSDs between this structure and the computed structures were 1.13 and 2.75 Å for the alpha and heavy atoms, respectively (see Figure 8); these numbers increased to 1.24 and 2.84 Å in the energy minimized ensemble.

```
A: AKIGLFYGTQTGVTQTIAESIQQEFGG-ESIVD-LNDIANADA-SDLNA
B:   MKIVYWSGTGNTEKMAELIAKGIIESGKDVN-TINVSDVNI-DELLN
C:  KIGIFFSTSTGNTTEVADFIGKTLG-A--KADAPIDVDDVTDPQALKD
D: AKALIVYGSTTGNTEYTAETIARELADAGYEVD-SRDAASVEAGGLFEG
E:MAITGIFFGSDTGNTENIAKMIQKQLGK--DVAD-VHDIAKSSK-EDLEA
          10        20          30         40
A:YDYLIIGCPTWN---VGELQ-SLWEGIYDD-LDSVNFQGKKVAYFGAGDQ
B:EDILILGCSAMGDE-V--LEESEFEPFIEE-ISTK-ISGKKVALFGSYG-
C:YDLLFLGAPTWNTAGDTERSGTSWDEFLYDKLPEVDMKDLPVAIFGLGDA
D:FDLVLLGCSTWGDDSI-ELQ-DDFIPLFDS-LEETGAQGRKVACFGCGDS
E:YDILLLGIPTWY---YGEAQ-CDWDDFFPT-LEEIDFNGKLVALFGCGDQ
    50          60         70         80         90
A:VGYSDNFQDAMGILEEKISSLGSQTVGYWPIEGYDFNESKAVRNNQ-FVG
B:--WGD--GKWMRDFEERMNGYGCVVVET--------------------P
C:EGYPDNFCDAIEEIHDCFAKQGAKPVGFSNPDDYDYEESKSVRDGK-FLG
D:-SY-EYFCGAVDAIEEKLKNLGAEIVQD--------------------G
E:EDYAEYFCDALGTIRDIIEPRGATIVGHWPTAGYHFEASKGLADDDHFVG
        100        110        120        130        140
A:LAIDEDNQPDLTKNRIKTWVSQLKSEFGL
B:LIVQNE--PDEAEQDCIEFGKKIANI
C:LPLDMVNDQIPMEKRVAGWVEAVVSETGV
D:LRIDGD--PRAARDDIVGWAHDVRGAI
E:LAIDEDRQPELTAERVEKWVKQISEELHLDEIHNA
        150        160        170
```

Figure 7 The multiple sequence alignment used for predicting the structure of *E. coli* Flavodoxin. The letters before the colon on each line are A for *A. nidulans*, B for *C. beijerinckii*, C for *C. crispus*, D for *D. vulgaris*, and E for *E. coli*. The remaining letters on each line are the standard one-letter amino acid codes, while dashes indicate deletions with respect to one or more of the other sequences. The numbering is with respect to the *E. coli* sequence. Letters of the sequences A–D with a line through them were not used for generating constraints; letters of the sequence E with a line through them had no corresponding residues in the homologues, and hence were completely unconstrained

We conclude that the backbone conformations of our models were nearly as accurate as they were precise, but that there were significant errors in many of the sidechain conformations. Given the low sequence identities between the *E. coli* form and its homologues, and the inevitable changes of sidechain conformation upon mutation, this is not surprising. It remains to be seen if a constraint generation protocol can be developed that performs better on the sidechains even in such cases of low sequence identities.

5 OUTLOOK

We have shown that distance geometry provides molecular modelers with a powerful set of tools for solving a wide variety of conformational problems. While there are certainly limits on what one can do with a purely geometric model of molecular structure, there are also substantial advantages, both computational and conceptual, to using such a model whenever possible. Distance geometry has the additional benefit of a solid mathematical foundation, which provides a route to deriving global insights into the ‘structure’ of conformation space as a whole. Indeed, distance geometry can be regarded as a means of defining and working with infinite sets of conformations, thereby achieving a degree of mathematical parallelism that potentially dwarfs that of any conceivable computer. This theory is actually in a rather primitive state of development at this time, and there is the potential for major

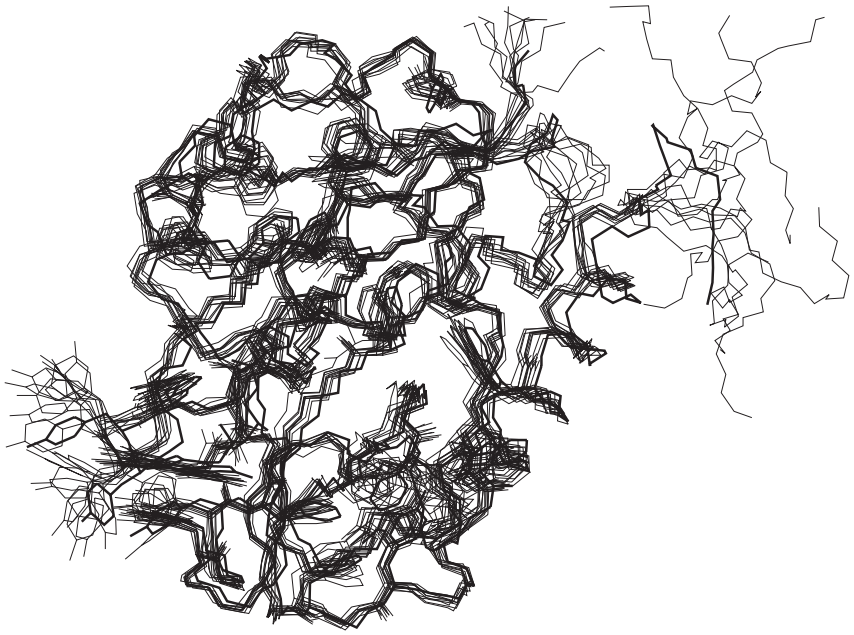

Figure 8 Illustration of the ensemble of 10 *E. coli* Flavodoxin structures obtained from homology modeling using distance geometry, superimposed on the crystal structure (heavy line) so as to minimize the coordinate differences to the alpha carbons in residues 4–170. Only the heavy backbone and aromatic sidechain atoms are shown, together with those of the flavin mononucleotide cofactor (lower left)

advances with far-reaching implications, perhaps even in the analysis of more complete molecular models based on either classical or quantum mechanics (cf. Ref. 63).

6 RELATED ARTICLES

Conformational Flexibility in 3D Structure Searching; Conformational Sampling; Conformational Search: Proteins; De Novo Ligand Design; Drug Design; Macromolecular Structure Calculation and Refinement by Simulated Annealing: Methods and Applications; Macromolecular Structures Determined using NMR Data; Molecular Docking and Structure-based Design; NMR Refinement; Protein Folding and Optimization Algorithms; Stereochemistry: Representation and Manipulation; Structural Chemistry: Application of Mathematics; Structure Representation; Superfamily Analysis: Understanding Protein Function from Structure and Sequence.

7 REFERENCES

1. G. M. Crippen and T. F. Havel, 'Distance Geometry and Molecular Conformation', Research Studies Press, Taunton, 1988.
2. D. Hestenes, 'New Foundations for Classical Mechanics', Kluwer, Dordrecht, 1986.
3. M. J. Crowe, 'A History of Vector Analysis', University of Notre Dame Press, Notre Dame, IL, 1967.
4. K. Menger, *Math. Ann.*, 1928, **100**, 75–163.
5. L. M. Blumenthal, 'Theory and Applications of Distance Geometry', Oxford University Press, Oxford, 1953 (reprinted by Chelsea, Bronx, 1970).
6. H. Kraft, 'Geometrische Methoden in der Invariantentheorie', Friedr. Vieweg & Sohn, Braunschweig, 1984.
7. J. P. Dalbec, *Ann. Math. Artif. Intell.*, 1995, **13**, 97–108.
8. J. J. Seidel, *Indag Math.*, 1955, **17**, 329–340, 535–541.
9. A. W. M. Dress and T. F. Havel, *Found. Phys.*, 1993, **23**, 1357–1374.
10. D. Hestenes, *Acta Appl. Math.*, 1991, **23**, 65–93.
11. A. W. M. Dress and T. F. Havel, *Discrete Appl. Math.*, 1988, **19**, 129–141.
12. T. F. Havel and K. Wüthrich, *Bull. Math. Biol.*, 1984, **46**, 673–698.
13. P. L. Easthope and T. F. Havel, *Bull. Math. Biol.*, 1989, **51**, 173–194.
14. R. Connelly, *Invent. Math.*, 1982, **66**, 11–33.
15. T. F. Havel, *Biopolymers*, 1990, **29**, 1565–1585.
16. G. M. J. Crippen, *Comput. Chem.*, 1989, **10**, 896–902.
17. W. Glunt, T. L. Hayden, and M. Raydan, *J. Comput. Chem.*, 1993, **14**, 114–120.
18. G. M. Crippen and T. F. Havel, *Acta Crystallogr., Sect. A*, 1978, **34**, 282–284.
19. J. Kuszewski, M. Nilges, and A. T. Brünger, *J. Biomol. NMR*, 1992, **2**, 33–56.
20. J. de Leeuw, *J. Classification*, 1988, **5**, 163–180.
21. J. C. Gower, *Lin. Alg. Appl.*, 1985, **67**, 81–97.
22. F. Critchley, *Lin. Alg. Appl.*, 1988, **105**, 91–107.
23. T. F. Havel, I. D. Kuntz, and G. M. Crippen, *Bull. Math. Biol.*, 1983, **45**, 665–720.
24. D. F. Shanno, *Math. Oper. Res.*, 1978, **3**, 244–256.
25. W. Braun and N. Gõ, *J. Mol. Biol.*, 1985, **186**, 611–626.
26. P. Güntert, W. Braun, and K. Wüthrich, *J. Mol. Biol.*, 1991, **217**, 517–530.
27. L. M. Rice and A. T. Brünger, *Proteins*, 1994, **19**, 277–290.
28. P. Luginbühl, P. Güntert, M. Billeter, and K. Wüthrich, *J. Biomol. NMR*, 1996, **8**, 136–146.
29. G. M. Clore, M. Nilges, A. T. Brünger, M. Karplus, and A. M. Gronenborn, *FEBS Lett.*, 1987, **213**, 269–277.
30. R. Kaptein, R. Boellens, R. M. Scheek, and W. F. van Gunsteren, *Biochemistry*, 1988, **27**, 5389–5395.
31. W. Nerdal, D. R. Hare, and B. R. Reid, *J. Mol. Biol.*, 1988, **201**, 717–739.
32. T. F. Havel, *Prog. Biophys. Mol. Biol.*, 1991, **56**, 43–78.

33. M. Nilges, G. M. Clore, and A. M. Gronenborn, *FEBS Lett.*, 1988, **239**, 129–136.
34. G. M. Crippen, *J. Comput. Chem.*, 1984, **5**, 548–554.
35. P. Weiner, S. Profeta, G. Wipff, T. F. Havel, I. D. Kuntz, R. Langeridge, and P. A. Kollman, *Tetrahedron*, 1983, **39**, 1113–1121.
36. J. M. Blaney and J. S. Dixon, in 'Reviews in Computational Chemistry', eds. K. B. Lipkowitz and D. B. Boyd, VCH, New York, 1994, Vol. V, pp. 299–335.
37. T. F. Havel, I. D. Kuntz, and G. M. Crippen, *J. Theor. Biol.*, 1983, **104**, 359–381.
38. R. H. Lathrop and T. F. Smith, in 'Proc. 27th Annual Hawaii Conf. Sys. Sci.', IEEE, Washington, DC, 1994, pp. 365–374.
39. G. M. Crippen, *J. Comput. Chem.*, 1992, **13**, 351–361.
40. N. Gõ and H. A. Scheraga, *Macromolecules*, 1970, **3**, 178–187.
41. T. F. Havel and I. Najfeld, *J. Mol. Struct. (Theochem)*, 1995, **336**, 175–189.
42. M. Billeter, T. F. Havel, and I. D. Kuntz, *Biopolymers*, 1987, **26**, 777–793.
43. I. Najfeld and T. F. Havel, *J. Math. Chem.*, 1997, **21**, 223.
44. A. S. Smellie, G. M. Crippen, and W. G. Richards, *J. Chem. Inf. Comput. Sci.*, 1991, **31**, 386–392.
45. J. M. Blaney, personal communication, 1996.
46. I. D. Kuntz, *Science*, 1992, **257**, 1078–1082.
47. R. P. Sheridan, R. Nilakantan, J. S. Dixon, and R. Venkataraghavan, *J. Med. Chem.*, 1986, **29**, 899–906.
48. A. K. Ghose and G. M. Crippen, in 'Comprehensive Medicinal Chemistry', eds. C. Hansch, P. G. Sammes, J. B. Taylor, and C. A. Ramsden, Pergamon, Oxford, 1990, Vol. 4, pp. 715–734.
49. G. M. Crippen, *J. Comput. Chem.*, 1995, **16**, 486–500.
50. T. F. Havel, in 'Encyclopedia of Nuclear Magnetic Resonance', eds. R. K. Harris and D. M. Grant, Wiley, Chichester, 1995, Vol. 4, pp. 1701–1710.
51. G. Wagner, S. G. Hyberts, and T. F. Havel, *Annu. Rev. Biophys. Biomol. Struct.*, 1992, **21**, 167–198.
52. T. E. Malliavin, A. Rouh, M. A. Delsuc, and J. Y. Lallemand, *C. R. Acad. Sci., Series II*, 1992, **315**, 653–659.
53. C. M. Oshiro and I. D. Kuntz, *Biopolymers*, 1993, **33**, 107–115.
54. M. Nilges, *Proteins: Struct. Func. Genet.*, 1993, **17**, 297–309.
55. A. T. Brünger and M. Karplus, *Acc. Chem. Res.*, 1991, **24**, 54–61.
56. A. T. Brünger and M. Nilges, *Q. Rev. Biophys.*, 1993, **26**, 49–125.
57. T. F. Havel, G. M. Crippen, and I. D. Kuntz, *Biopolymers*, 1979, **18**, 73–82.
58. T. F. Havel and K. Wüthrich, *J. Mol. Biol.*, 1985, **182**, 281–294.
59. M. P. Williamson, T. F. Havel, and K. Wüthrich, *J. Mol. Biol.*, 1985, **182**, 295–315.
60. S. G. Hyberts, M. S. Goldberg, T. F. Havel, and G. Wagner, *Protein Sci.*, 1992, **1**, 736–751.
61. T. F. Havel, *Mol. Simul.*, 1993, **10**, 175–210.
62. T. F. Havel and M. P. Snow, *J. Mol. Biol.*, 1991, **217**, 1–7.
63. J. Paldus and B. Jeziorski, *Theor. Chim. Acta*, 1988, **73**, 81–103.