THE SOLUTION OF THE DISTANCE GEOMETRY PROBLEM IN PROTEIN MODELING VIA GEOMETRIC BUILDUP

By

Di Wu

Zhijun Wu

and

Yaxiang Yuan

IMA Preprint Series # 2184

(December 2007)

INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS

UNIVERSITY OF MINNESOTA
400 Lind Hall
207 Church Street S.E.
Minneapolis, Minnesota 55455–0436

Phone: 612/624-6066   Fax: 612/626-7370
URL: http://www.ima.umn.edu

THE SOLUTION OF THE DISTANCE GEOMETRY PROBLEM IN PROTEIN MODELING VIA GEOMETRIC BUILDUP*

DI WU

Department of Mathematics, Western Kentucky University, USA

ZHIJUN WU

Department of Mathematics, Iowa State University, USA

YAXIANG YUAN

Institute of Computational Mathematics, Chinese Academy of Science, China

Abstract. A well-known problem in protein modeling is the determination of the structure of a protein with a given set of inter-atomic or inter-residue distances obtained from either physical experiments or theoretical estimates. A general form of the problem is known as the distance geometry problem in mathematics, the graph embedding problem in computer science, and the multidimensional scaling problem in statistics. The problem has applications in many other scientific and engineering fields as well such as sensor network localization, image recognition, and protein classification. We describe the formulations and complexities of the problem in its various forms, and introduce a geometric buildup approach to the problem. Central to this approach is the idea that the coordinates of the atoms in a protein can be determined one atom at a time, with the distances from the determined atoms to the undetermined ones. It can determine a structure more efficiently than other conventional approaches, yet without requiring more distance constraints than necessary. We present the general algorithm and its theory and review the recent development of the algorithm for controlling the propagation of the numerical errors in the buildup process, for determining rigid vs. unique structures, and for handling problems with inexact distances (distances with errors). We show the results from applying the algorithm to some of the model problems and justify the potential use of the algorithm in protein modeling.

Key words Biomolecular modeling, protein structure determination, distance geometry, graph embedding, linear and nonlinear systems of equations, linear and nonlinear optimization

* Work supported by the NIH/NIGMS grant R01GM081680, and the NSF grant of China.

1. Distance Based Protein Modeling

Proteins are an important class of biological molecules. They are encoded in genes and produced in cells through genetic translation. They are life supporting (or sometimes, destructing) ingredients and are indispensable for almost all biological processes. For example, humans have hundreds of thousands of different proteins and would not be able to maintain normal life even if short of a singe type of protein (Figure 1a). On the other hand, with the help of some proteins, viruses are able to grow, translate, integrate, and replicate, causing diseases (Figure 1b). Some proteins themselves are toxic and even infectious such as the proteins in poisonous plants and in beef causing the Mad Cow Disease (Figure 1c). [1]

A protein consists of a linear chain of amino acids connected with strong chemical bonds. The amino acids and their order in the chain are fixed for each different protein, and they are specified by the gene (a sequence of DNA molecules) from which the protein is generated. Once the chain of amino acids for a protein is produced, it immediately folds into a unique and stable 3D structure, which is crucial for the protein to function. Since the function of the protein depends on its structure, the determination of the structure becomes a necessary step for the understanding of the biological properties of every protein. [1]

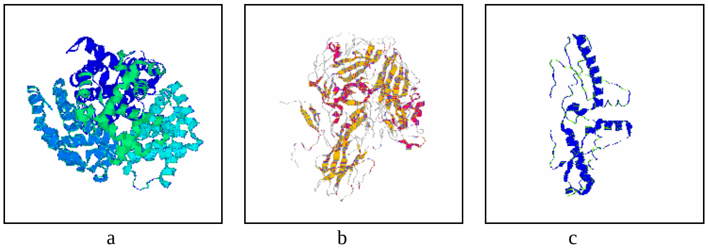


Figure 1 **Example proteins** a. hemoglobin protein, 1BUW, in blood; b. protein 2PLV, supporting poliovirus; c. prion protein 1I4M-D, causing the Mad Cow Disease in human.

Unfortunately, there is no direct physical means to observe a protein structure at an atomic level. There are only techniques that can be used to measure certain physical properties of the protein upon which the structure can be deduced. X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR) are major experimental techniques of such in practice. They are responsible for the determination of 80% and 15% of the protein structures (total about 30,000) so far deposited in the Protein Data Bank (PDB), respectively [2]. The experimental techniques have many limitations, though. X-ray crystallography requires purifying and crystallizing proteins, which may take months or years to finish, if not failed. The results often vary with varying experiments for reasons not fully understood [3]. NMR can only be applied to small proteins for otherwise the spectral data would become too difficult to clarify [4]. The structures determined by NMR are not as accurate and detailed as well [1]. Theoretical or computational approaches such as homology modeling, structural alignment, threading, energy minimization, dynamic simulation, etc., have been developed [5][6], but they are more successful in building theoretical models or refining experimental structures than determining the structures completely independently, although recent progress as shown in the CASP competitions [7] and in utilizing more powerful computing resources is indeed exciting and encouraging [8].

In this paper, we discuss a well-known problem in protein modeling, for the determination of the structure of a protein with a given set of inter-atomic or inter-residue distances obtained from either physical experiments or theoretical estimates (Figure 2). A more general and abstract form of the problem is known as the distance geometry problem in mathematics [9], the graph embedding problem in computer science [10], and the multidimensional scaling problem in statistics [11]. In general, the problem can be stated as to find the coordinates for a set of points in some topological space given the distances for certain pairs of points. Therefore, in addition to protein modeling where everything is discussed only in three-dimensional Euclidean space, the problem has applications in many other scientific and engineering fields as well, such as sensor network localization [12], image recognition [13], and protein classification [14], to name a few. In any case, the problem may or may not have a solution in a given topological space, and even if it does have a solution, the solution may not be easy to find, depending on the given distances. For example, in any k-dimensional Euclidean space, the problem is polynomial time solvable if the distances for all the pairs of points are provided, and is NP-complete otherwise in general [10].

In protein modeling, the distances or their ranges for certain pairs of atoms or residues in a given protein may be obtained from either physical experiments such as NOE (Nuclear Overhauser Effects), J-coupling, and dipolar coupling in NMR [4][15][16], or theoretical estimates such as the bond lengths and bond angles known from general organic chemistry [1], or statistical estimates on certain inter-atomic or inter-residue distances based on their distributions in databases of known protein structures [17][18][19]. Then, a structure may be determined for the protein by using the available distances. However, the given distances may not necessarily be sufficient for determining the structure uniquely, or even just rigidly. Here, by uniquely we mean that the structure is unique under translation and rotation, and by rigidly we mean that any part of the structure cannot be changed continuously without violating the given distance restraints. Sometimes, the distances may contain errors and may be inconsistent in the sense that they may have violated some basic geometric conditions such as the triangle inequality for the distances among any three atoms. In that case, a structure that fits the given distances will not even exist. After all, even if a structure does exist, it is still not trivial to determine based on the given distances. A distance geometry problem needs to be solved, which is computationally intractable in general [10].

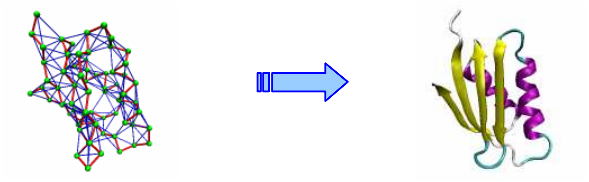


Figure 2 **Distance based protein modeling** Given a set of inter-atomic distances or their ranges, find the coordinates of the atoms in the protein.

Crippen and Havel and several other research groups [20][21] pioneered the work on using the solution of a distance geometry problem for protein structure determination, especially for NMR structure modeling, where the distances for certain pairs of atoms and in particular, the pairs of hydrogen atoms that are within say, 5 Ǻ distance, can be estimated through J-couplings and NOE, with additional ones that can be derived from known bond lengths and bond angles. However, in NMR modeling, the distances obtained are restricted to a small subset of all pairs of atoms in the protein. Otherwise, if the distances for all pairs of atoms are available, a structure would be much easier to build upon. The NMR distances also contain experimental errors and are not necessarily always consistent. A structure that can fit the distances approximately rather than exactly may be the best we can hope for in practice. Moreover, in NMR, instead of exact distances, the ranges or lower and upper bounds of the distances are usually provided, due to the fact that the structures are flexible in solution and the distances are not fixed. An ensemble of structures rather than a single one that can fit in the distance ranges are therefore sought in real practice to show the dynamic nature of the structure [22][23]. For these various reasons, the focus on NMR modeling has been more on developing methods for extracting the bounds on the missing distances (bound smoothing), removing the inconsistencies in the distances (distance metrication), and fitting the structures in the distance ranges (optimization), as described in the embed algorithm [20][21] and implemented in NMR modeling software such as the CNS [24][25]. Therefore, the solution of an exact distance geometry problem has not been improved much since the embed algorithm was first developed, and its impact in NMR modeling has been rather limited. On the other hand, important theoretical and algorithmic issues related to the solution of the problem still remain to be resolved, while its applications in more general areas of distance-based protein modeling are expanding [26][27][28][29].

Existing approaches to the solution of the distance geometry problem include, for example, the embedding algorithm by Crippen and Havel [20][21], the alternating projection method by Glunt and Hayden [32][33], the graph reduction approach by Hendrickson [30][31], the global optimization method by Moré and Wu [34][35], the stochastic/perturbation method by Zou, Byrd, and Schnabel [36], the multidimensional scaling method by Kearsly, Tapia, and Trosset [37][38], the dc programming method by Le Thi Hoai and Pham Dinh [39], the semi-definite programming approach by Biswas, Liang, Toh, and Ye [40], and the stochastic search method by Grosso, Locatelli, and Schoen [41].

We investigate the solution of the distance geometry problem within a so-called geometric buildup framework. Dong and Wu [42][43] first implemented a geometric buildup algorithm for the solution of the distance geometry problem with exact distances and justified the linear computation time for the case when the distances required in every buildup step are always available. Central to the geometric buildup approach is the idea to determine only a small group of atoms at the beginning and then complete the whole molecule by repeatedly determining one or more atoms every time using the available distances between the determined and undetermined atoms. The advantage of using a geometric buildup approach is that it works directly on the given distances and exploits the special structure of a given problem, and hence may be able to solve the problem more efficiently than a general approach. We present the general algorithm of this approach, and discuss related computational issues including control of numerical errors, determination of rigid vs. unique structures, and tolerance of distance errors, based on the recent development of the algorithm [44][45][46]. The theoretical basis of the approach is established based on the theory of distance geometry. A group of necessary and sufficient conditions for the determination of a structure with a given set of distances using a geometric buildup algorithm are justified. The applications of the algorithm to model protein problems are demonstrated.

2. The Distance Geometry Problem

Let n be the number of atoms in a given protein and $x_1, \ldots, x_n$ be the coordinate vectors for the atoms, where $x_i = (x_{i,1}, x_{i,2}, x_{i,3})^T$ and $x_{i,1}$, $x_{i,2}$, and $x_{i,3}$ are the first, second, and third coordinates of atom i. If the coordinates $x_1, \ldots, x_n$ are known, the distances $d_{i,j}$ between atoms i and j can be computed with $d_{i,j} = ||x_i - x_j||$, where $||\cdot||$ is the Euclidean norm. Conversely, if the distances $d_{i,j}$ are given, the coordinates $x_1, \ldots, x_n$ for the atoms can also be obtained based on the distances $d_{i,j}$, but the computation is not as straightforward. The solution of a system of equations as can be stated in the following for $x_1, \ldots, x_n$ is required.

$$\| x_i - x_j \| = d_{i,j}, \quad (i, j) \in S, \tag{2.1}$$

where S is a subset of all atom pairs. The latter problem is known as a distance geometry problem in mathematics [9], a graph embedding problem in computer science [10], and a multidimensional scaling problem in statistics [11]. In practice, the distances may have errors, and therefore, a more general yet practical form of the problem would be to find the coordinates of the atoms $x_1, \ldots, x_n$, given only a set of lower and upper bounds, $l_{i,j}$ and $u_{i,j}$, of the distances $d_{i,j}$ such that

$$l_{i,j} \le \| x_i - x_j \| \le u_{i,j}, \quad (i, j) \in S. \tag{2.2}$$

The distance geometry problem is polynomial time solvable if the distances for all pairs of atoms are available. However, it has been proved to be NP-hard in general. Even if errors are allowed for the distances, the problem is still hard, if only small errors are allowed.

2.1 Problems with Exact Distances

We first consider the simple case when a complete set of exact distances is given. By exact distances we mean the distances are given in exact values, not in ranges, and by a complete set of distances we mean the distances for all pairs of atoms are included. A solution to the distance geometry problem with such a set of distance data can be obtained efficiently by using for example an algorithm that requires the singular value decomposition (SVD) of an induced distance matrix.

Assume that a set of coordinates $x_1, \ldots, x_n$ can be found for a given set of distances $d_{i,j}$, where $i, j = 1, \ldots, n$. Then, $\|x_i - x_j\| = d_{i,j}$ for all $i, j = 1, \ldots, n$, and

$$\| x_i \|^2 - 2x_i^T x_j + \| x_j \|^2 = d_{i,j}^2, \quad i, j = 1,\ldots,n. \tag{2.3}$$

Since the molecular structure is invariant under any translation or rotation, we set a reference system so that the origin is located at the last atom or in other words, $x_n = (0, 0, 0)^T$. It follows that

$$d_{i,n}^2 - 2x_i^T x_j + d_{j,n}^2 = d_{i,j}^2, \quad i, j = 1,\ldots,n-1. \tag{2.4}$$

Define a coordinate matrix X and an induced distance matrix D,

$$\begin{aligned} &X = \{x_{i,j} : \ i = 1,\ldots,n-1, \ j = 1, 2, 3\} \quad \text{and} \\ &D = \{(d_{i,n}^2 - d_{i,j}^2 + d_{j,n}^2)/2 : \ i, j = 1,\ldots,n-1\}. \end{aligned} \tag{2.5}$$

Then, $XX^T = D$ and D must be of maximum rank 3.

The distance geometry problem can be defined in a general space R^k with $x_1, \ldots, x_n$ in R^k and $d_{i,j}$ the Euclidean distances between atoms i and j. Then, the equation $XX^T = D$ still holds, and D must be of maximum rank k, where $X = \{x_{i,j} : i = 1, \ldots, n, j = 1, \ldots, k\}$.

Theorem 2.1.1 [9] Let $\{d_{i,j} : i, j = 1, \ldots, n\}$ be a set of distances in R^k, for some $k \leq n$. Then, the induced matrix D as defined in (2.5) is of maximum rank k.

Proof It follows from the fact that $D = XX^T$ for a coordinate matrix X in $R^{n-1} \times R^k$ and X is of maximum rank k. □

The equation $XX^T = D$ can be solved using the singular value decomposition of D. Let $D = U\Sigma U^T$ be the singular value decomposition of D, where U is an orthogonal matrix and Σ a diagonal matrix with the singular values of D along the diagonal. If D is a matrix of rank less than or equal to k, the decomposition can be obtained with U being $(n\text{-}1) \times k$ and Σ being $k \times k$. Then, $X = U\Sigma^{1/2}$ solves the equation $XX^T = D$. Here the singular value decomposition of D requires $O(kn^2)$ floating-point operations [47], and therefore, the distance geometry problem with a complete set of exact distances can be solved in polynomial time.

Note that although in practice, the distances may not be available for all the pairs of atoms, the solution of the problem with all exact distances can still be important for the solution of the general problem with a sparse set of distances. For example, in the embed algorithm, a complete set of distances among all the atoms is generated after bound smoothing, and the solution of a distance geometry problem with all exact distances is always required afterwards

[20][21]. Also, if a subset of atoms has all the distances among the atoms, but the whole set of atoms does not, the coordinates of the subset of atoms can still be determined efficiently by solving a distance geometry problem with all exact distances for the subset of atoms. The procedure may also be applied repeatedly as some of the atoms are determined and the availability of the distances among them is changed, until no such subsets of atoms can be found [48][49].

2.2 Problems with Sparse Distances

We now consider the problem with an incomplete set of exact distances. Let S be a subset of all pairs of atoms such that (i,j) is in S if the distance $d_{i,j}$ between atoms i and j is given. Then, the problem is to find the coordinates $x_1, \ldots, x_n$ for the atoms so that

$$\| x_i - x_j \| = d_{i,j}, \quad (i,j) \in S \cdot \tag{2.6}$$

Let $G = (V, E, W)$ be a weighted graph, where $V = \{v_1, \ldots, v_n\}$ is the set of vertices, $E = \{e_{i,j} : (i,j) \text{ in S}\}$ the set of edges, and $W = \{w_{i,j} = d_{i,j} : (i,j) \text{ in } S\}$ the weights on the edges. Then, the distance geometry problem for molecular structure determination can be considered as a graph embedding problem for G in R^3, i.e., to find a mapping from the vertices $v_1, \ldots, v_n$ in V to a set of points $x_1, \ldots, x_n$ in R^3 so that the distances between points i and j for all (i,j) in S are equal to the weights $d_{i,j}$ on the corresponding edges $e_{i,j}$.

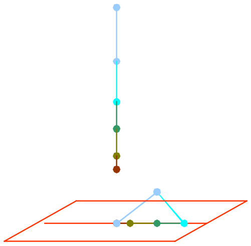


Figure 3 **Folding a closed chain** The integer set partition problem can be reduced to the problem of folding a closed chain in a line, a one-dimensional graph embedding problem, thereby proving that the one-dimensional graph embedding problem must be NP-hard, for the integer set partition problem has proved to be.

The graph embedding problem can be considered in a Euclidean space of any dimension. In any case, it has been proved that the graph embedding problem is an NP-hard problem even for the one-dimensional case [10]. The proof can be demonstrated via the solution of a special class of one-dimensional graph embedding problem, the problem of folding a closed chain in a line (in one-dimensional space, Figure 3). Let $G = (V, E, W)$, with $V = \{v_1, \ldots, v_{n+1}\}$, $E =$

$\{e_{i,i+1} : i = 1, \ldots, n\} \cup \{e_{1,n+1}\}$, and $W = \{w_{i,i+1} = l_i : i = 1, \ldots, n\} \cup \{w_{1,n+1} = 0\}$, where l_i is the length of the link between node i and node i+1 in the chain. Then, the problem can be stated formally as to find a mapping from the nodes $\{v_1, \ldots, v_{n+1}\}$ of G to a set of points $\{x_1, \ldots, x_{n+1}\}$ in R so that

$$| x_{i+1} - x_i | = l_i, \quad i = 1, \ldots, n, \quad | x_{n+1} - x_1 | = 0 \cdot \tag{2.7}$$

Theorem 2.2.1 The integer set partition problem can be reduced to the problem of folding a closed chain in a line.

Proof Let $A = \{a_1, \ldots, a_n\}$ be a given set of positive integers. Define a graph $G = (V, E, W)$, with $V = \{v_1, \ldots, v_{n+1}\}$, $E = \{e_{i,i+1} : i = 1, \ldots, n\} \cup \{e_{1,n+1}\}$, and $W = \{w_{i,i+1} = a_i : i = 1, \ldots, n\} \cup \{w_{1,n+1} = 0\}$. The graph defines a closed chain. Suppose that the chain can be folded in a line or in other words, the graph can be embedded in R. Then, v_i is placed at x_i in R for $i = 1, \ldots, n$+1, and

$$| x_{i+1} - x_i | = a_i, \quad i = 1, \ldots, n, \quad | x_{n+1} - x_1 | = 0 \cdot$$

Let $A_1 = \{a_i = |x_{i+1} - x_i| = x_{i+1} - x_i\}$ and $A_2 = \{a_i = |x_{i+1} - x_i| = x_i - x_{i+1}\}$. Then,

$$\sum_{i=1}^{n} (x_{i+1} - x_i) = \sum_{a_i \in A_1} (x_{i+1} - x_i) - \sum_{a_i \in A_2} (x_i - x_{i+1}) \cdot$$

However,

$$\sum_{i=1}^{n} (x_{i+1} - x_i) = x_{n+1} - x_1 = 0 \cdot$$

It follows that

$$\sum_{a_i \in A_1} (x_{i+1} - x_i) - \sum_{a_i \in A_2} (x_i - x_{i+1}) = \sum_{a_i \in A_1} a_i - \sum_{a_i \in A_2} a_i = 0 ,$$

and A_1 and A_2 solves the set partition problem for A. □

It follows from the above theorem that the problem of folding a closed chain in a line cannot be in P, for otherwise, the set partition problem would be solvable in P via the solution of an equivalent chain folding problem, which is contradictory to the fact that the set partition problem is in NP [50].

2.3 Problems with Inexact Distances

In protein modeling practice, the distances are often provided with estimated ranges only. The related distance geometry problem then becomes to find the coordinates $x_1, \ldots, x_n$ of the atoms, so that the distances between atoms i and j, for all (i,j) in a subset S of all pairs of atoms, are within their estimated ranges, i.e.,

$$l_{i,j} \le \| x_i - x_j \| \le u_{i,j}, \quad (i,j) \in S \cdot \tag{2.8}$$

where $l_{i,j}$ and $u_{i,j}$ are the lower and upper bounds of the distances between atoms i and j. Let $d_{i,j} = (l_{i,j} + u_{i,j}) / 2$ and $\varepsilon_{i,j} = (u_{i,j} - l_{i,j}) / 2$. The above problem can be written as

$$|\| x_i - x_j \| - d_{i,j} | \le \varepsilon_{i,j}, \quad (i,j) \in S \text{,} \tag{2.9}$$

and be viewed as to find an approximate solution to the distance geometry problem for a set of exact distances $d_{i,j}$ with each distance $\|x_i - x_j\|$ allowed to have an error $\varepsilon_{i,j}$ from $d_{i,j}$. We call such a solution an ε-approximate solution.

If large errors are allowed, an approximate solution is certainly easier to obtain than an exact solution. However, if only small errors are allowed, the problem for finding an approximate solution can be as hard as for finding an exact solution. To see this, again, we can consider the simple case of folding a closed chain in a line, but this time, we allow the links to be connected loosely. Let $G = (V, E, W)$, with $V = \{v_1, \ldots, v_{n+1}\}$, $E = \{e_{i,i+1} : i = 1, \ldots, n\} \cup \{e_{1,n+1}\}$, and $W = \{w_{i,i+1} = l_i : i = 1, \ldots, n\} \cup \{w_{1,n+1} = 0\}$, where l_i is the length of the link between node i and node i+1 in the chain. Then, the problem can be stated formally as to find a mapping from the nodes $\{v_1, \ldots, v_{n+1}\}$ of G to a set of points $\{x_1, \ldots, x_{n+1}\}$ in R so that

$$\| x_{i+1} - x_i | - l_i | \le \varepsilon_i, \quad i = 1, \ldots, n, \quad | x_{n+1} - x_1 | \le \varepsilon_{n+1}, \tag{2.10}$$

for a set of errors $\{\varepsilon_1, \ldots, \varepsilon_{n+1}\}$.

Moré and Wu [51] showed that the above problem is also NP-hard when the allowed errors are small. In fact, the set partition problem can again be reduced to this problem with $\varepsilon_i < 1/(2n)$ for $i = 1, \ldots, n+1$. Here, we give another proof that requires only $\Sigma_i \varepsilon_i < 1$, removing the dependence of the required bound of the errors on the problem size n explicitly.

Theorem 2.3.1 The integer set partition problem can be reduced to the problem of folding a closed chain with total allowed error $\Sigma_i \varepsilon_i < 1$.

Proof Let $A = \{a_1, \ldots, a_n\}$ be a given set of positive integers. Define a graph $G = (V, E, W)$, with $V = \{v_1, \ldots, v_{n+1}\}$, $E = \{e_{i,i+1} : i = 1, \ldots, n\} \cup \{e_{1,n+1}\}$, and $W = \{w_{i,i+1} = a_i : i = 1, \ldots, n\} \cup \{w_{1,n+1} = 0\}$. The graph defines a closed chain. Suppose that the chain can be folded in a line with an error ε_i allowed on each length a_i and $\Sigma_i \varepsilon_i < 1$. Then, v_i is placed at x_i in R for $i = 1, \ldots, n+1$, and

$$|| x_{i+1} - x_i | - a_i | \le \varepsilon_i, \quad i = 1,\ldots,n, \quad | x_{n+1} - x_1 | \le \varepsilon_{n+1} .$$

Let $A_1 = \{a_i = |x_{i+1} - x_i| = x_{i+1} - x_i\}$ and $A_2 = \{a_i = |x_{i+1} - x_i| = x_i - x_{i+1}\}$. Then,

$$\begin{aligned}&\sum_{i=1}^{n}(x_{i+1} - x_i) = \sum_{a_i \in A_1}(x_{i+1} - x_i) - \sum_{a_i \in A_2}(x_i - x_{i+1})\\ &\ge \sum_{a_i \in A_1}(a_i - \varepsilon_i) - \sum_{a_i \in A_2}(a_i + \varepsilon_i) = \sum_{a_i \in A_1} a_i - \sum_{a_i \in A_2} a_i - \sum\nolimits_{i=1}^{n} \varepsilon_i ,\end{aligned}$$

and

$$\begin{aligned}&\sum_{i=1}^{n}(x_{i+1} - x_i) = \sum_{a_i \in A_1}(x_{i+1} - x_i) - \sum_{a_i \in A_2}(x_i - x_{i+1})\\ &\le \sum_{a_i \in A_1}(a_i + \varepsilon_i) - \sum_{a_i \in A_2}(a_i - \varepsilon_i) = \sum_{a_i \in A_1} a_i - \sum_{a_i \in A_2} a_i + \sum\nolimits_{i=1}^{n} \varepsilon_i .\end{aligned}$$

Therefore,

$$\sum_{i=1}^{n}(x_{i+1} - x_i) - \sum\nolimits_{i=1}^{n} \varepsilon_i \le \sum_{a_i \in A_1} a_i - \sum_{a_i \in A_2} a_i \le \sum_{i=1}^{n}(x_{i+1} - x_i) + \sum\nolimits_{i=1}^{n} \varepsilon_i .$$

However,

$$-\varepsilon_{n+1} \le \sum_{i=1}^{n}(x_{i+1} - x_i) = x_{n+1} - x_1 \le \varepsilon_{n+1} \cdot$$

It follows that

$$-1 < -\sum\nolimits_{i=1}^{n+1} \varepsilon_i \le \sum_{a_i \in A_1} a_i - \sum_{a_i \in A_2} a_i \le \sum\nolimits_{i=1}^{n+1} \varepsilon_i < 1 \cdot$$

Note that the two sums in the middle are over the integers and their difference cannot be a fraction. Therefore,

$$\sum_{a_i \in A_1} a_i - \sum_{a_i \in A_2} a_i = 0 ,$$

and A_1 and A_2 solves the set partition problem for A. □

3. The Geometric Buildup Approach

Central to the geometric buildup approach to the distance geometry problem is the idea to determine only a small group of atoms at the beginning and then complete the whole molecule by repeatedly determining one or more atoms every time using the available distances between the determined and undetermined atoms. The advantage of using a geometric buildup approach is that it works directly on the given distances and exploits the special structure of a given problem, and hence may be able to solve the problem more efficiently than a general approach. Dong and Wu [42] first applied a geometric buildup algorithm to the solution of the distance geometry problem, and showed that the algorithm can find a solution to the problem in $O(n)$ floating-point operations if the distances for all the pairs of atoms are available. The work was later extended to sparse distances [43] with an updating scheme to control the propagation of numerical errors in the buildup process [44]. The recent development on the algorithm includes the enhancement of the algorithm on rigid vs. unique structure determination [45] and the extension of the algorithm to handling inexact or inconsistent distance data [46].

3.1 The General Algorithm

Given an arbitrary set of distances, the algorithm first finds four atoms that are not in the same plane and determines the coordinates for the four atoms, using for example the singular value decomposition algorithm as described in Section 2.1, with all the distances among them (assuming available). Then, for any undetermined atom j, the algorithm repeatedly performs a procedure as follows: Find four determined atoms that are not in the same plane and have distances available to atom j, and determine the coordinates for atom j. Let $x_i = (x_{i,1}, x_{i,2}, x_{i,3})^T$, $i = 1, 2, 3, 4$, be the coordinate vectors of the four atoms. Then, the coordinates $x_j = (x_{j,1}, x_{j,2}, x_{j,3})^T$ for atom j can be determined by using the distances $d_{i,j}$ from atoms $i = 1, 2, 3, 4$ to atom j (Figure 4). Indeed, x_j can be obtained from the solution of the following system of equations,

$$\| x_i \|^2 - 2x_i^T x_j + \|x_j\|^2 = d_{i,j}^2, \quad i = 1, 2, 3, 4 \cdot \tag{3.1}$$

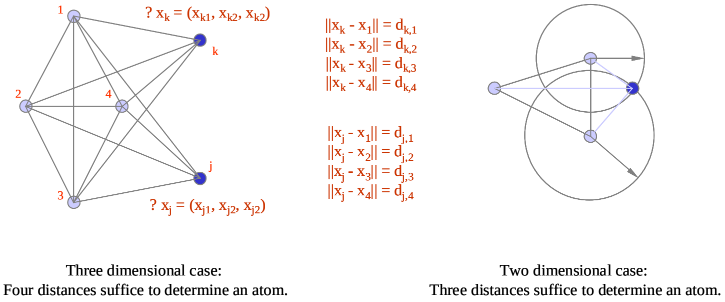
By subtracting equation i from equation i+1 for $i = 1, 2, 3$, we can eliminate the quadratic terms for x_j to obtain

$$\begin{aligned} & -2(x_{i+1} - x_i)^T x_j \\ & = (d_{i+1,j}^2 - d_{i,j}^2) - (\| x_{i+1} \|^2 - \|x_i\|^2), \quad i = 1, 2, 3. \end{aligned} \tag{3.2}$$

Let A be a matrix and b a vector, and

$$A = -2\begin{bmatrix} (x_2 - x_1)^T \\ (x_3 - x_2)^T \\ (x_4 - x_3)^T \end{bmatrix}, \quad b = \begin{bmatrix} (d_{2,j}^2 - d_{1,j}^2) - (\| x_2 \|^2 - \|x_1\|^2) \\ (d_{3,j}^2 - d_{2,j}^2) - (\| x_3 \|^2 - \|x_2\|^2) \\ (d_{4,j}^2 - d_{3,j}^2) - (\| x_4 \|^2 - \|x_3\|^2) \end{bmatrix}. \tag{3.3}$$

We then have $Ax_j = b$. Since x_1, x_2, x_3, x_4 are not in the same plane, A must be nonsingular, and we can therefore solve the linear system to obtain a unique solution for x_j. Here, solving the linear system requires only constant time. Since we only need to solve n–4 such systems for n–4 coordinate vectors x_j, the total computation time is proportional to n, if in every step, the required coordinates x_i and distances $d_{i,j}$, i = 1, 2, 3, 4 are always available.


1
2
3
4
k
j
? $x_k = (x_{k1}, x_{k2}, x_{k2})$
? $x_j = (x_{j1}, x_{j2}, x_{j2})$
$||x_k - x_1|| = d_{k,1}$
$||x_k - x_2|| = d_{k,2}$
$||x_k - x_3|| = d_{k,3}$
$||x_k - x_4|| = d_{k,4}$
$||x_j - x_1|| = d_{j,1}$
$||x_j - x_2|| = d_{j,2}$
$||x_j - x_3|| = d_{j,3}$
$||x_j - x_4|| = d_{j,4}$

Three dimensional case:
Four distances suffice to determine an atom.

Two dimensional case:
Three distances suffice to determine an atom.

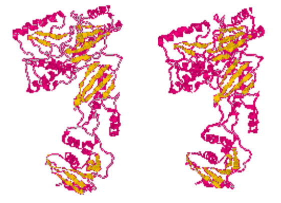
Figure 4 **Geometric buildup** In two-dimensional space, if there are three determined atoms that are not in the same line and there are distances from these atoms to an undetermined atom, the undetermined atom can be determined uniquely using the three distances. In three-dimensional space, if there are four determined atoms that are not in the same plane and there are distances from these atoms to an undetermined atom, the undetermined atom can be determined uniquely using the four distances.

Figure 5 shows an example protein structure determined by using the general geometric buildup algorithm, with the distances for all the pairs of atoms in the protein, as demonstrated in Dong and Wu [42]. The structure is determined accurately and uniquely. The RMSD value of the structure compared with its X-ray reference structure is 1.0e-04 Å. The computation time is much more efficient than the conventional singular value decomposition algorithm as described in Section 2.1.

The theoretical basis of the general geometric buildup algorithm can be traced back in the theory of distance geometry [9]. Several authors had discussions on the theoretical issues related to such an approach as well, including Saxe [10], Sippl and Scheraga [48][49], and Huang, Liang, and Pardalos [52]. Based on the distance geometry theory, any point in a Euclidean space can be determined in terms of the distances from this point to a special set of points.

The General Geometric Buildup Algorithm

```
1. Determine an initial set of atoms.
2. Repeat:
     For each undetermined atom j,
       If atom j has distances to four independent and determined atoms,
     Determine atom j with these distances.
         End
     End
     If no atoms are determined in the loop, unsuccessfully stop.
3. All atoms are successfully determined.
```



RMSD = 1.0e-04 Å

Figure 5 **Geometric buildup** The X-ray crystal structure (left) of the HIV-1 RT p66 protein (4200 atoms) and the structure (right) determined by the geometric buildup algorithm using the distances for all pairs of atoms in the protein. The algorithm took only 188,859 floating-point operations, while a conventional singular-value decomposition algorithm required 1,268,200,000 floating-point operations.

Definition 3.1.1 A set of points *B* in a space *S* is a metric basis of *S* provided any point in *S* can be uniquely determined by its distances to the points in *B*.

Definition 3.1.2 A set of k+1 points in R^k is called an independent set of points if it is not a set of points in $R^{k\text{-}1}$.

Theorem 3.1.1 A set of k+1 independent points in R^k form a metric basis for R^k.

Proof It follows directly by generalizing the basic geometric buildup step to the k-dimensional Euclidean space. Let $x_i = (x_{i,1}, \ldots, x_{i,k})^T$ be the coordinate vectors of an independent set of points $i = 1, \ldots, k+1$ in R^k. Let $x_j = (x_{j,1}, \ldots, x_{j,k})^T$ be the coordinate vector for any point j in R^k with distances $d_{i,j}$ from points $i = 1, \ldots, k+1$ to point j. Then,

$$\| x_i \|^2 - 2x_i^T x_j + \|x_j\|^2 = d_{i,j}^2, \quad i = 1, \ldots, k+1, \tag{3.4}$$

and $Ax_j = b$, where

$$A = -2\begin{bmatrix} (x_2 - x_1)^T \\ (x_3 - x_2)^T \\ \cdots \\ (x_{k+1} - x_k)^T \end{bmatrix}, \quad b = \begin{bmatrix} (d_{2,j}^2 - d_{1,j}^2) - (\| x_2 \|^2 - \|x_1\|^2) \\ (d_{3,j}^2 - d_{2,j}^2) - (\| x_3 \|^2 - \|x_2\|^2) \\ \cdots \\ (d_{k+1,j}^2 - d_{k,j}^2) - (\| x_{k+1} \|^2 - \|x_k\|^2) \end{bmatrix}. \tag{3.5}$$

Since the points i = 1, …, k+1 are not in R^{k-1}, the matrix A must be nonsingular and x_j is determined uniquely. □

Given the above properties, we can easily see that a necessary condition for uniquely determining the coordinates of the atoms with a given set of distances is that each atom must have at least four distances to other atoms, and a sufficient condition is that in every step of the geometric buildup algorithm, there is an undetermined atom and the atom has four distances from four determined atoms who are not in the same plane. In general, we have

Theorem 3.1.2 A necessary condition for the unique determination of the coordinates of a group of points x_1, …, x_n in R^k with a given set of distances among the points is that each point must have at least k+1 distances from other k+1 points, assuming that this point is not in R^{k-1} with any k of the k+1 points.

Proof It follows immediately from the fact that in R^k, a point can be defined uniquely only if it has k+1 distances from k+1 independent points, assuming it is not in R^{k-1} with any k of the k+1 points. If it has only k distances from k points, the point will have at least two reflective positions. □

Theorem 3.1.3 A sufficient condition for the unique determination of the coordinates of a group of points x_1, …, x_n in R^k with a given set of distances among the points is that in every step of the geometric buildup algorithm, there is an undetermined point with k+1 distances from k+1 independent and determined points.

Proof The geometric buildup algorithm gives a constructive proof for the theorem, because if the condition holds in every step of the algorithm, the algorithm will be able to determine the coordinates of all the points uniquely. □

3.2 Control of Numerical Errors

The general geometric buildup algorithm can be sensitive to the numerical errors generated during the calculation of the coordinates of the atoms. With this algorithm, the coordinates of many atoms are determined by using the coordinates of previously determined atoms, and therefore, the errors in the previously determined atoms are passed to and accumulated in later determined atoms. As a result, the coordinates for later determined atoms may become completely incorrect, especially if there is a long sequence of atoms to be determined.

Wu and Wu [44] proposed an updating scheme to prevent the accumulation of the numerical errors. The idea of the scheme is based on the fact that the coordinates of any four atoms can be determined without any other information if all the distances among them are given. Therefore, the coordinates of any four determined atoms should be recalculated whenever possible using the distances among them, before they are used as a basis set of atoms for the determination of other atoms. The recalculated coordinates do not depend on the coordinates of previously determined atoms and therefore do not inherit any errors from them. They are determined from "scratch" and will not pass previous errors to later atoms as well. In this way, the coordinates of many atoms can be "corrected", and the errors in the calculated coordinates can be prevented from growing into incorrect structural results.

The recalculation of the coordinates of the four atoms in the above algorithm usually is done in an independent coordinate system, which is not related to the overall structure already constructed by the algorithm. However, they can be moved back to the original structure by aligning them to their original locations with an appropriate translation and rotation (Figure 6). In other words, the new coordinates of the four atoms can be translated and rotated so that the root-mean-square-deviation (RMSD) between the new coordinates and the old ones is minimized.

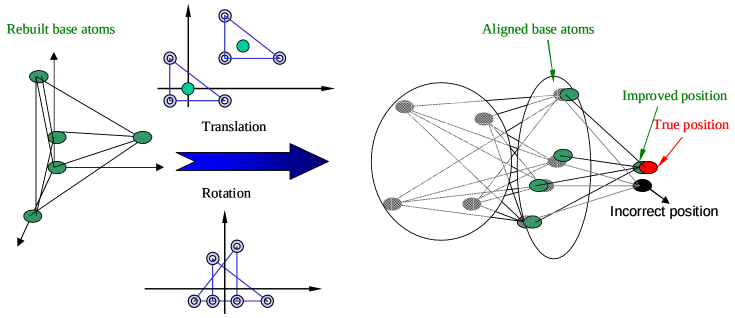


Figure 6 **Re-determination of base atoms** The four base atoms are re-determined if the distances among them are given. The atoms are then moved to and aligned with their original positions, and used to determine other atoms.

Let y_1, …, y_4 be the coordinate vectors of the four atoms calculated in the regular geometric buildup process, and x_1, …, x_4 the recalculated coordinate vectors. Let Y and X be the corresponding coordinate matrices. If the distances among all the four atoms are available, X can be obtained for example using the singular value decomposition algorithm described in Section 2.1. In order to move X to the position where Y is located in the molecule, the geometric centers of X and Y are calculated first:

$$x_c^T = \sum_{i=1}^{4} X(i,:)/4, \quad y_c^T = \sum_{i=1}^{4} Y(i,:)/4 \cdot \tag{3.6}$$

Then, X is translated so that the geometric centers of X and Y are at the same location,

$$X <= X + e(y_c - x_c)^T, \tag{3.7}$$

where $e = (1, 1, 1, 1)^T$. After the translation, a rotation for X is selected so that the root-mean-square-deviation of X and Y is minimized. In fact, the calculation of such a deviation can be done by solving an optimization problem,

$$\min_{Q} \| Y - XQ \|_F, \quad QQ^T = I, \tag{3.8}$$

where $\| \|_F$ is the matrix Frobenius norm and Q the rotation matrix. Let $C = X^TY$, and let $C = U\Sigma V^T$ be the singular-value decomposition of C. Then, it is not difficult to verify that $Q = UV^T$ solves the above optimization problem [47].

The Updated Geometric Buildup Algorithm

```
1. Determine an initial set of atoms.
2. Repeat:
   For each undetermined atom j,
       If atom j has distances to four independent and determined atoms,
          If the distances among the determined atoms are given in the original data,
          Recalculate their coordinates with these distances.
          End
          Determine atom j with these distances.
       End
   End
   If no atoms are determined in the loop, unsuccessfully stop.
3. All atoms are successfully determined.
```

Figure 7 demonstrates in some scenarios for how the structure determined by a geometric buildup algorithm can be affected by the accumulated numerical errors and how they can be corrected by using the updating scheme, as given in Wu and Wu [44]. The figure shows the structures (red lines) of protein 4MBA (1086 atoms) determined using ≤ 5 Å distances, first by the general geometric buildup algorithm (Figure 7a) and then by the updating algorithm (Figure 7b). The graphs show that the general algorithm results in a structure that disagrees with the X-ray reference structure (blue lines) in many regions, while the updating algorithm generates a structure that agrees with the X-ray reference structure (blue lines) almost completely.

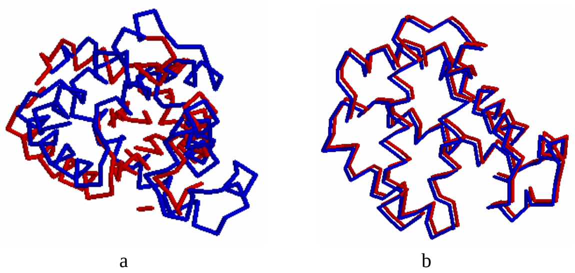


Figure 7 **Control of rounding errors** a. The structure (red lines) of 4MBA determined by using a general geometric buildup algorithm and compared with the original structure of 4MBA (blue lines). b. The structure (red lines) of 4MBA determined by using an updating geometric buildup algorithm and compared with the original structure of 4MBA (blue lines).

3.3 Rigid vs. Unique Buildup

For the unique determination of a structure, it is necessary that every atom has at least four distances from other atoms. Further, the general geometric buildup algorithm requires four distances from four determined atoms to the atom to be determined in every buildup step. These conditions may not be satisfied by a given set of distances in practice. If the first condition is not satisfied, the structure will not be guaranteed unique. If the second condition is not satisfied, the general geometric buildup algorithm will not be able to determine the structure, even if the first condition is satisfied and the structure is unique.

In order to handle more sparse distance data, we can consider determining the structures only rigidly instead of uniquely. The necessary condition to have a rigid structure requires only three distances for each atom. Therefore, in every buildup step, the geometric buildup algorithm can be modified to require only three distances from three determined atoms to the atom to be determined. The atom can then be determined rigidly, although with two possible positions. In the end, the algorithm may produce multiple structures, due to the multiple choices of the positions of the atoms, but the structures are rigid and in finite number.

More formally, in any buildup step, let $x_i = (x_{i,1}, x_{i,2}, x_{i,3})^T$, $i = 1, 2, 3$, be the coordinate vectors of three determined atoms that are not in a line. Let $x_j = (x_{j,1}, x_{j,2}, x_{j,3})^T$ be the coordinate vector for an undetermined atom j and $d_{i,j}$ the distances from atoms $i = 1, 2, 3$ to atom j. Then, x_j can be obtained from the solution of the following system of equations,

$$\| x_i \|^2 - 2x_i^T x_j + \|x_j\|^2 = d_{i,j}^2, \quad i = 1, 2, 3 \cdot \tag{3.9}$$

By subtracting equation i from equation i+1 for $i = 1, 2$, we can eliminate the quadratic terms for x_j to obtain

$$\begin{aligned} & -2(x_{i+1} - x_i)^T x_j \\ & = (d_{i+1,j}^2 - d_{i,j}^2) - (\| x_{i+1} \|^2 - \|x_i\|^2), \quad i = 1, 2. \end{aligned} \tag{3.10}$$

Let A be a matrix and b a vector, and

$$A = -2\begin{bmatrix} (x_2 - x_1)^T \\ (x_3 - x_2)^T \end{bmatrix}, \quad b = \begin{bmatrix} (d_{2,j}^2 - d_{1,j}^2) - (\| x_2 \|^2 - \|x_1\|^2) \\ (d_{3,j}^2 - d_{2,j}^2) - (\| x_3 \|^2 - \|x_2\|^2) \end{bmatrix}. \tag{3.11}$$

We then have $Ax_j = b$. Let $x_j = A^T y_j$, where $y_j = (y_{j,1}, y_{j,2})^T$. Then, $AA^T y_j = b$. Since x_1, x_2, x_3 are not in the same line, A must be full rank and AA^T be nonsingular. We can therefore solve the linear system $AA^T y_j = b$ to obtain a unique solution for y_j. Let $x_j' = (x_{j,1}, x_{j,2})^T$ and $A' = A(1{:}2,1{:}2)$. Then, $x_j' = [A']^{\mathrm{T}} y_j$. By using one of the equations in (3.9), we can obtain two possible values for $x_{j,3}$, assuming that the equation has real solutions. In the end, we obtain two solutions for (3.9).

The advantage of using the modified buildup algorithm is that the algorithm requires fewer distance constraints than the general buildup algorithm. It can handle even more sparse distance data, yet determine meaningful structures. The modified algorithm may find multiple structures, but they all are rigid, and in some cases, it can find a unique structure as well, because the requirement by the general buildup algorithm on the availability of the special four distances in every buildup step is sufficient for the determination of a unique structure, but not necessary.

However, a problem with the modified buildup algorithm is that it may produce too many possible structures: Since in every step, an atom is only determined rigidly, there may be at least two possible positions for it. We have to keep both positions unless later on we find that one of them can be excluded with other distance constraints. Moreover, the three determined atoms may also have multiple positions. Let the ith determined atom have l_i possible positions, $i = 1, 2, 3$. Then, in the worst case, there can be $2 \times l_1 \times l_2 \times l_3$ possible positions for the atom to be determined. Therefore, as the algorithm proceeds, the total number of possible positions for an atom to be determined may grow into exponentially many.

To reduce the number of possible positions for an atom, we can allow the algorithm to determine the atom uniquely first if there are more than three required distances available, and determine it rigidly otherwise. Also, in every buildup step, after the atom is determined, either rigidly or uniquely, we can examine all given distances from this atom to other determined atoms for their possible positions. If some positions have violated their distance constraints, they can be removed for further consideration. In this way, the structures generated in the end are guaranteed to satisfy all available distance constraints among the atoms, and they may be reduced to a unique structure after all infeasible structures are identified and removed.

The Rigid Geometric Buildup Algorithm

```
1. Determine an initial set of atoms.
2. Repeat:
   For each undetermined atom $j$,
       If atom $j$ has distances to four independent and determined atoms,
          Determine atom $j$ with these distances.
          Check multiple structures with additional available distances.
       End
       If atom $j$ has distances to three independent and determined atoms,
          Determine atom $j$ with these distances.
          Record multiple structures generated from reflections.
       End
   End
   If no atoms are determined in the loop, unsuccessfully stop.
3. All atoms are successfully determined.
```

Figure 8 shows how a structure can be determined rigidly and how multiple structures can be generated and also reduced. Figure 8a shows that atom i is first determined with three available distances. There are two positions for atom i due to reflection, which makes two possible structures. Figure 8b shows that atom j again is determined with three available distances, with two positions for each of the possible structures. Total four possible structures are made. In Figure 8c, atom k is determined uniquely with four distances, and therefore, the number of possible structures is not increased. However, there is an additional distance between atoms *i* and *k*. By examining all the structures, we find that two of them do not satisfy this distance constraint, and they can be removed from the structure pool, as shown in Figure 8d.

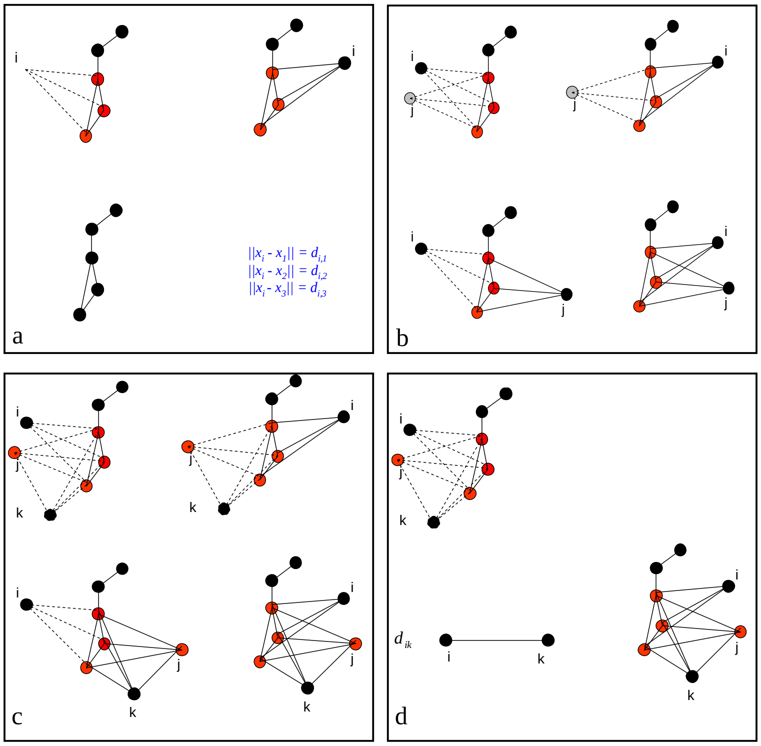

i
i
$||x_i - x_1|| = d_{i,1}$
$||x_i - x_2|| = d_{i,2}$
$||x_i - x_3|| = d_{i,3}$
a
i
j
i
j
i
j
i
j
b
i
j
k
i
j
k
i
j
k
i
j
k
c
i
j
k
d_{ik}
i
k
i
j
k
d

Figure 8 **Multiple rigid structures** a. Atom *i* is determined. The number of structures is two. b. Atom *j* is determined. The number of structures is increased to four. c. Atom *k* is determined. d. Two structures are removed because they do not satisfy the distance constraint for atom *i* and *k*.

Figure 9 further demonstrates the application of the rigid geometric buildup algorithm to a small protein, 1AKG, and the nature of the multiple structures it can generate, as given along with other examples in [45]. The protein 1AKG is a small polypeptide with 16 amino acids and 110 atoms. The general geometric buildup algorithm is able to determine to the structure for this protein completely, with distances ≤ 4.5 Å, and the RMSD value of the structure is 8.3e-07 Å against the original structure. Here, the number of distances used is 1638, which is about 14% of all the distances. However, with distances ≤ 3.5 Å, the general geometric buildup algorithm fails, but the rigid algorithm is still able to find a reasonable number of rigid structures. Here, the number of distances used is 898, which is only 7.5% of all the distances. There are total 8192 multiple conformations found by the rigid algorithm. The one closest to the original structure has the RMSD value equal to 4.3e-07 Å. Note that 8192 = 2^{13}, and therefore, the multiple structures are perhaps generated just from a sequence of 13 reflections of the atomic positions. In fact, as can be observed in the figure, most of the reflections happen for the side-chain atoms when they are in the surface of the protein, and they only affect the determination of a small part of the structure. On the other hand, the major parts of the protein with the backbone atoms and the atoms in the interior of the protein are all uniquely determined.

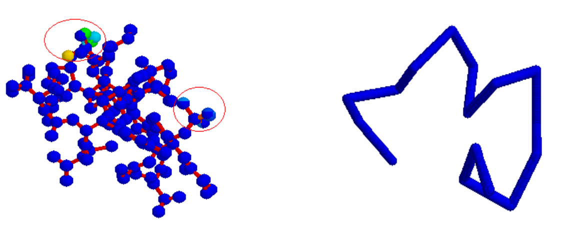


Figure 9 **Rigid structure determination** Shown is the structure of protein 1AKG, with 16 residues, 110 atoms. The distances < 3.5 Å were used. Total 8192 rigid structures were determined. They all were almost identical except for the circled small regions.

Similar to the general geometric buildup algorithm, the theoretical basis for the rigid geometric buildup algorithm can be established and generalized to any k-dimensional Euclidean space. For this purpose, we define a reduced metric basis for a space and k independent points in R^k.

Definition 3.3.1 A set of points B in a space S is a reduced metric basis of S provided any point in S can be determined rigidly by its distances to the points in B.

Definition 3.3.2 A set of k points in R^k is said to be an independent set of points if it is not a set of points in R^{k-2}.

Theorem 3.3.1 A set of k independent points in R^k form a reduced metric basis for R^k.

Proof It follows directly by generalizing the modified geometric buildup step to the *k*-dimensional Euclidean space. Let $x_i = (x_{i,1}, \ldots, x_{i,k})^T$ be the coordinate vectors of an independent set of points $i = 1, \ldots, k$ in R^k. Let $x_j = (x_{j,1}, \ldots, x_{j,k})^T$ be the coordinate vector for any point *j* in R^k with distances $d_{i,j}$ from points $i = 1, \ldots, k$ to point *j*. Then

$$\| x_i \|^2 - 2x_i^T x_j + \|x_j\|^2 = d_{i,j}^2, \quad i = 1, \ldots, k, \tag{3.12}$$

and $Ax_j = b$, where

$$A = -2\begin{bmatrix} (x_2 - x_1)^T \\ (x_3 - x_2)^T \\ \cdots \\ (x_k - x_{k-1})^T \end{bmatrix}, \quad b = \begin{bmatrix} (d_{2,j}^2 - d_{1,j}^2) - (\| x_2 \|^2 - \|x_1\|^2) \\ (d_{3,j}^2 - d_{2,j}^2) - (\| x_3 \|^2 - \|x_2\|^2) \\ \cdots \\ (d_{k,j}^2 - d_{k-1,j}^2) - (\| x_k \|^2 - \|x_{k-1}\|^2) \end{bmatrix}. \tag{3.13}$$

Let $x_j = A^T y_j$, where $y_j = (y_{j,1}, \ldots, y_{j,k-1})^T$. Then, $AA^T y_j = b$. Since $x_1, \ldots, x_k$ are not in R^{k-2}, *A* must be full rank and AA^T be nonsingular. We can therefore solve the linear system $AA^T y_j = b$ to obtain a unique solution for y_j. Let $x_j' = (x_{j,1}, \ldots, x_{j,k-1})^T$ and $A' = A(1{:}k\text{-}1, 1{:}k\text{-}1)$. Then, $x_j' = [A']^T y_j$. By using one of the equations in (3.12), we can obtain two possible values for $x_{j,k}$, assuming that the equation has real solutions. In the end, we obtain two solutions for (3.12), and the positions for point *j* are determined rigidly. □

Given the above properties, we can easily see that a necessary condition for rigidly determining the coordinates of the atoms with a given set of distances is that each atom must have at least three distances to other atoms, and a sufficient condition is that in every step of the geometric buildup algorithm, there is an undetermined atom and the atom has three distances from three determined atoms who are not in the same line. In general, we have

Theorem 3.3.2 A necessary condition for the rigid determination of the coordinates of a group of points $x_1, \ldots, x_n$ in R^k with a given set of distances among the points is that each point must have at least *k* distances from other *k* points, assuming that this point is not in R^{k-2} with any *k*-1 of the *k* points.

Proof It follows immediately from the fact that in R^k, a point can be defined rigidly only if it has *k* distances to *k* independent points, assuming it is not in R^{k-2} with any *k*-1 of the *k* points. If it has only *k*-1 distances from *k*-1 points, the position of the point will be flexible. □

Theorem 3.3.3 A sufficient condition for the rigid determination of the coordinates of a group of points $x_1, \ldots, x_n$ in R^k with a given set of distances among the points is that in every step of the geometric buildup algorithm, there is an undetermined point with k distances from k independent and determined points.

Proof The modified geometric buildup algorithm gives a constructive proof for the theorem, because if the condition holds in every step of the algorithm, the algorithm will be able to determine the coordinates of all the points rigidly. □

3.4 Tolerance of Inexact Distances

In practice, the distance data often contains errors. As a result, the distances may become inconsistent or in other words, may have violated some basic geometric rules such as the triangle inequality for the distances among any three atoms. The general geometric buildup algorithm usually assumes that the distances are consistent and therefore, in every step, only four (or three) distances are required for the determination of the coordinates of an atom uniquely (or rigidly), although there may be more available. However, this will not be the case if the distances are not consistent. In order for the algorithm to handle inexact distances (distances with errors), the general buildup procedure has to be modified. First, in every buildup step, if l distances are found from an undetermined atom to l determined atoms, $l \geq 4$, all l distances should be used for the determination of the unknown atom. Second, if $l \geq 4$, an over-determined system of equations is obtained for the determination of the position of the unknown atom. If the distances have errors, the system may not be consistent. Therefore, we can only solve the system approximately by using for example a least-squares method. Third, a new updating scheme may be necessary to prevent the accumulation of the rounding errors. The updating scheme described in Section 3.2 may not be practical any more for $l >> 4$ because it requires all the distances available among l determined atoms.

A simple way to extend the geometric buildup algorithm to handle the possible errors from the distance data is as follows. In every buildup step, in addition to the four required distances, we can include all the available distances, say l distances, from the determined atoms to the one to be determined (see Figure 10). Let $x_i = (x_{i,1}, x_{i,2}, x_{i,3})^T$, $i = 1, \ldots, l$, be the coordinate vectors of the l determined atoms and $d_{i,j}$ the distances from atoms $i = 1, \ldots, l$ to the undetermined atom j. Then, the coordinates $x_j = (x_{j,1}, x_{j,2}, x_{j,3})^T$ for atom j can be obtained from the solution of the following system of equations,

$$\| x_i \|^2 - 2x_i^T x_j + \|x_j\|^2 = d_{i,j}^2, \quad i = 1, \ldots, l \, . \tag{3.14}$$

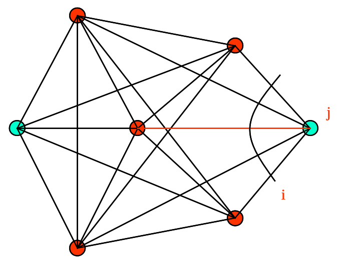
By subtracting equation i from equation i+1 for i = 1, …, l-1, we can eliminate the quadratic terms for x_j to obtain

$$\begin{aligned} &-2(x_{i+1}-x_i)^T x_j \\ &=(d_{i+1,j}^2-d_{i,j}^2)-(\|x_{i+1}\|^2-\|x_i\|^2), \quad i=1,\ldots,l-1. \end{aligned} \tag{3.15}$$

Let A be a matrix and b a vector, and

$$A=-2\begin{bmatrix}(x_2-x_1)^T\\(x_3-x_2)^T\\\cdots\\(x_l-x_{l-1})^T\end{bmatrix}, \quad b=\begin{bmatrix}(d_{2,j}^2-d_{1,j}^2)-(\|x_2\|^2-\|x_1\|^2)\\(d_{3,j}^2-d_{2,j}^2)-(\|x_3\|^2-\|x_2\|^2)\\\cdots\\(d_{l,j}^2-d_{l-1,j}^2)-(\|x_l\|^2-\|x_{l-1}\|^2)\end{bmatrix}. \tag{3.16}$$

We then have $Ax_j = b$. This system is certainly over-determined if $l > k+1$. However, it can be solved by using a standard linear least-squares method. For example, we can compute the QR factorization of A to obtain an equation $QRx_j = b$, where Q is $(l\text{-}1)\times 3$ and R is 3×3. If at least four of the l determined atoms are not in the same plane, A must be full rank and R be nonsingular. We can then solve the linear system $QRx_j = b$ to obtain a unique solution $x_j = R^{-1}Q^Tb$, which minimizes $\|b - Ax_j\|$. Here, solving the linear system $QRx_j = b$ requires $O(l)$ computing time, but QR factorization may take $O(l^2)$ time. Since we only need to solve $\sim n$ such linear least-squares problems for $\sim n$ coordinate vectors x_j, the total computation time must be in order of $l_m^2\ n$, if in every step, the required coordinates x_i and distances $d_{i,j}$ are always available, where $l_m = \max_j \{|S_j|\}$, $S_j = \{i : (i,j) \text{ in } S\}$.



x_i are determined.

$\| x_i - x_j \|^2 = d_{i,j}^2, \quad (i, j) \in S_j \subset S$

x_j is to be determined.

Figure 10 **Tolerance of distance errors** The extended algorithm tries to determine the coordinates of each atom by taking all available distance constraints into account and by minimizing the errors for all the constraints. In this way, all the constraints are intended to be satisfied, and the algorithm is also more stable with possible errors in the distance data.

Again, the theory for the extended geometric buildup algorithm can be established and generalized to any k-dimensional Euclidean space in a similar fashion as that for the general geometric buildup algorithm. For this purpose, we define an extended metric basis for a space and an extended set of independent points in R^k.

Definition 3.4.1 A set of points B in a space S is an extended metric basis of S provided any point in S can be determined uniquely by its distances from the points in B.

Definition 3.4.2 A set of l points is said to be an extended set of independent points in R^k if it contains k+1 independent points.

Theorem 3.4.1 An extended set of l independent points in R^k form an extended metric basis for R^k.

Proof It follows directly by generalizing the extended geometric buildup step to the k-dimensional Euclidean space. Let $x_i = (x_{i,1}, \ldots, x_{i,k})^T$ be the coordinate vectors for an extended set of independent points $i = 1, \ldots, l$ in R^k. Let $x_j = (x_{j,1}, \ldots, x_{j,k})^T$ be the coordinate vector for any point j in R^k with distances $d_{i,j}$ from points $i = 1, \ldots, l$ to point j. Then

$$\| x_i \|^2 - 2x_i^T x_j + \|x_j\|^2 = d_{i,j}^2, \quad i = 1, \ldots, l, \tag{3.17}$$

and $Ax_j = b$, where

$$A = -2\begin{bmatrix} (x_2 - x_1)^T \\ (x_3 - x_2)^T \\ \cdots \\ (x_l - x_{l-1})^T \end{bmatrix}, \quad b = \begin{bmatrix} (d_{2,j}^2 - d_{1,j}^2) - (\| x_2 \|^2 - \|x_1\|^2) \\ (d_{3,j}^2 - d_{2,j}^2) - (\| x_3 \|^2 - \|x_2\|^2) \\ \cdots \\ (d_{l,j}^2 - d_{l-1,j}^2) - (\| x_l \|^2 - \|x_{l-1}\|^2) \end{bmatrix}. \tag{3.18}$$

Multiply the equation by A^T to obtain $A^TAx_j = A^Tb$. Since k+1 of the l determined points are independent, A must be full rank and A^TA be nonsingular. We can then solve the linear system $A^TAx_j = A^Tb$ to obtain a unique solution $x_j = [A^TA]^{-1}A^Tb$. □

The above algorithm may not necessarily be stable for preventing rounding errors from growing, because in every step, the coordinates of the unknown atom must have rounding errors, which can still be propagated and accumulated into later calculations. On the other hand, different from the general geometric buildup algorithm, it is difficult to employ an updating scheme as described in Section 3.2 for the extended algorithm, because the scheme requires the availability of the distances among all l determined atoms, which is not so realistic when l is large. In order to control the rounding errors as well as tolerate the distance errors, a nonlinear instead of linear least-squares approximation can in fact be used in the buildup procedure instead. The idea is to determine the unknown atom in each buildup step by using not only the l distances from l determined atoms to the unknown atom, but also the distances among all the l determined atoms. The l distances from l determined atoms to the unknown atom must be given. The distances among the l determined atoms may not necessarily be provided, but they can be calculated. In any case, once all these distances become available, the coordinates for the unknown atom and the l known atoms can all be calculated (or recalculated) using these distances.

Let x_1, …, x_l and x_{l+1} be the coordinate vectors of atoms 1, …, l+1. If the distances among all these atoms, $d_{i,j}$, i, j = 1, …, l+1, are available, then, $||x_i - x_j|| = d_{i,j}$ for all i, j = 1, …, l+1, and

$$\| x_i \|^2 - 2x_i^T x_j + \| x_j \|^2 = d_{i,j}^2, \quad i, j = 1, ..., l+1. \tag{3.19}$$

Since the structure formed by these atoms is invariant under any translation or rotation, we can set a reference system so that the origin is located at the last atom or in other words, $x_{l+1} = (0, 0, 0)^T$. It follows that $||x_i|| = d_{i,l+1}$, $||x_j|| = d_{j,l+1}$, and

$$d_{i,l+1}^2 - 2x_i^T x_j + d_{j,l+1}^2 = d_{i,j}^2, \quad i, j = 1, \ldots, l. \tag{3.20}$$

We now have a system of equations similar to the one discussed in Section 2.1. Define a coordinate matrix X and an induced distance matrix D,

$$\begin{aligned} &X = \{x_{i,k} : \; i = 1, \ldots, l, \quad k = 1, 2, 3\} \quad \text{and} \\ &D = \{(d_{i,l+1}^2 - d_{i,j}^2 + d_{j,l+1}^2)/2 : \; i, j = 1, \ldots, l\}. \end{aligned} \tag{3.21}$$

Then, $XX^T = D$. Let $D = U\Sigma U^T$ be the singular value decomposition of D, where U is an orthogonal matrix and Σ a diagonal matrix with the singular values of D along the diagonal. If D is a matrix of rank less than or equal to 3, $X = V\Lambda^{1/2}$ solves the equation $XX^T = D$, where $V = U(:,l:3)$ and $\Lambda = \Sigma(1:3,1:3)$. In other words, if the distances $d_{i,j}$ are available for all i, j = 1, …, l+1, we can always construct an induced matrix D for the distances and then, based on the singular value decomposition of D, obtain the coordinates for all the atoms 1, …, l as given in X with atom l+1 fixed at $(0,0,0)^T$.

Note that the distances may have errors. Then, the matrix D may in fact have a higher rank than k or in other words, the equation $XX^T = D$ may not have an exact solution. However, $X = V\Lambda^{1/2}$ as defined above is still a good approximation to the solution of the equation (3.20) in the following least-squares sense.

Theorem 3.4.2 Let $D = U\Sigma U^T$ be the singular value decomposition of D. Let $V = U(:,1:3)$ and $\Lambda = \Sigma(1:3,1:3)$. Then, $X = V\Lambda^{1/2}$ minimizes $||D - XX^T||_F$, where $||\ ||_F$ is the matrix Frobenius norm.

Proof [[21] Let $f(X) = ||D - XX^T||^2$. Then $(D - XX^T)X = 0$ for any stationary point X of f. It follows that $(D - XX^T)X = (D - XX^T)XX^T = 0$ and

$$f(X) = \text{trace}(D^2) - \text{trace}(2DXX^T - XX^TXX^T) = \text{trace}(D^2) - \text{trace}(XX^TXX^T).$$

Let $\sigma_1 \geq \ldots \geq \sigma_l \geq 0$ be the singular values of D and $\lambda_1 \geq \lambda_2 \geq \lambda_3 > 0$ be the singular values of XX^T. Then,

$$f(X) = \text{trace}(D^2) - \text{trace}(XX^TXX^T) = \sum_{j=1}^{l} \sigma_j^2 - \sum_{j=1}^{k} \lambda_j^2.$$

Let $XX^T = V\Lambda V^T$ be the singular value decomposition of XX^T, where V is an l×3 orthogonal matrix and $\Lambda = \text{diag}\ \{\lambda_1, \lambda_2, \lambda_3.\}$. Since $DXX^T = XX^TXX^T$, $V^TDV = \Lambda$ and therefore, $\{\lambda_j : j = 1, 2, 3\} \subset \{\sigma_j : j = 1, \ldots, n\}$. It follows that $f(X)$ is minimized when $\lambda_j = \sigma_j$ for $j = 1, 2, 3$. □

Geometric Buildup with Linear Least-Squares

```
1. Find four atoms that are not in the same plane.
2. Determine the coordinates of the atoms with the distances among them.
3. Repeat:
   For each of the undetermined atoms,
       If the atom has $l$ distances to $l$ determined atoms that are not in the same plane,
          Determine the atom with the least-squares fit to the distances.
       End
   End
4. If no atom can be determined in the loop, stop.
5. All atoms are determined.
```

The extended buildup procedure has the following properties. First, the coordinates of the unknown atom are determined by using l previously determined atoms, to which the unknown atom has distances given. Second, the coordinates are determined by solving a system of distance equations approximately. They are the best possible estimations in a nonlinear least-squares sense as stated in Theorem 3.4.2, and can therefore be evaluated even if the distances have errors. Third, the calculations not only determine the coordinates of the unknown atom, but also recalculate the coordinates of all the involved atoms including the determined ones. Most importantly, these coordinates do not depend completely on the results from previous calculations. Rather, they are determined by using the provided distances among the atoms (determined and undetermined) as much as possible, thereby reducing the risk of large error propagation and accumulation. In this sense, the method should be more stable numerically than the one using linear linear-squares approximation [46].

Geometric Buildup with Nonlinear Least-Squares

```
1. Find four atoms that are not in the same plane.
2. Determine the coordinates of the atoms with the distances among them.
3. Repeat:
   For each of the undetermined atoms,
       If the atom has $l$ distances to $l$ determined atoms that are not in the same plane,
         Determine the $l$+1 atoms with the distances among them.
         Put the atoms back to their original positions by proper translation and rotation
      End
   End
4. If no atom can be determined in the loop, stop.
5. All atoms are determined.
```

Of course, the calculations of the coordinates are conducted in an independent reference system with its origin at the position of the atom to be determined. In order to recover the coordinates of the atoms in their original structure, we need to make a proper translation and rotation for the coordinates just like we need to do in the updating scheme for the general geometric buildup algorithm. More specifically, let Y be an l×3 matrix having the original coordinates of the l determined atoms. Let X be an l×3 matrix with the recalculated coordinates of the determined atoms. First, we translate X to Y with a translation vector $y_c – x_c$, where y_c and x_c are the geometric centers of X and Y, respectively. Then, we can rotate the coordinates of all the atoms by using a rotation matrix $Q = UV^T$, where U and V are obtained from the singular value decomposition, $X^TY = U\Sigma V^T$. That is, if x_i is the coordinate vector of atom i, i = 1, …, l+1, then, we set x_i to Qx_i.

Table 1 and 2 show some results from applying the extended geometric buildup algorithm with either linear or nonlinear least-squares approximation to the determination of the structures of a set of proteins using the distance data generated from the experimental structures of the proteins with 5 Å and 6 Å cutoff values. Table 1 contains the RMSD (root-mean-square deviation) values of the structures (compared with their original structures) obtained by using the extended buildup algorithm with linear least-squares on the generated data sets. The RMSD values show that the algorithm solved almost all the problems with cutoff distances equal to 6 Å, but failed for those with cutoff distance equal to 5 Å. The last cutoff value is critical because in NMR modeling, usually only less than or equal 5 Å distances can be estimated. In any case, the results show that with linear least-squares, the new buildup algorithm performed well in general if the distance data was not too sparse. The reason that it did not work well for very sparse data was that a long sequence of buildup steps had to be carried out and a large amount of rounding errors was accumulated.

Table 1 **RMSD Values of Structures Computed with Linear Least-Squares**

ID	TA	≤5 Å		≤6 Å	
		DA	RMSD	DA	RMSD
1PTQ	402	402	1.4e-00	402	2.6e-09
1HOE	558	558	5.8e-02	558	3.1e-09
1LFB	641	641	2.0e-02	641	2.1e-10
1PHT	814	809	1.2e+01	814	8.2e-09
1POA	914	914	6.6e-00	914	1.9e-09
1AX8	1003	1003	5.2e-00	1003	1.8e-05
4MBA	1086	1083	4.9e-00	1086	3.8e-06
1F39	1534	1534	1.4e+01	1534	6.3e-08
1RGS	2015	2010	2.0e+01	2015	1.1e-01
1BPM	3672	3669	6.4e+04	3672	3.6e-02
1HMV	7398	7389	1.2e+03	7398	3.5e+01

*ID – Protein ID, TA – Total number of atoms, DA – Total number of determined atoms, RMSD – RMSD values of the computed structure against the original structures.

Table 2 **RMSD Values of Structures Computed with Nonlinear Least-Squares**

ID	TA	≤5 Å		≤6 Å	
		DA	RMSD	DA	RMSD
1PTQ	402	402	5.5e-14	402	5.0e-14
1HOE	558	558	1.6e-13	558	2.7e-13
1LFB	641	641	9.5e-14	641	5.5e-14
1PHT	814	809	1.1e-13	814	1.8e-13
1POA	914	914	3.2e-13	914	1.5e-13
1AX8	1003	976	4.0e-13	1003	4.6e-12
4MBA	1086	1083	1.8e-13	1086	2.6e-13
1F39	1534	1534	7.9e-13	1534	1.9e-13
1RGS	2015	2010	8.3e-12	2015	2.4e-12
1BPM	3672	3669	8.1e-11	3672	1.0e-11
1HMV	7398	7389	1.1e-08	7398	5.5e-07

*ID – Protein ID, TA – Total number of atoms, DA – Total number of determined atoms, RMSD – RMSD values of the computed structure against the original structures.

Table 2 contains the RMSD (root-mean-square deviation) values of the structures (compared with their original structures) obtained by using the new buildup algorithm with nonlinear least-squares on the data sets. The RMSD values show that the algorithm solved almost all the problems with cutoff distances equal to 5 Å and 6 Å. Therefore, the results indicated that with nonlinear least-squares, the new buildup algorithm performed well in general. The reason it worked well for very sparse data was that it calculated the coordinates of the undetermined as well as determined atoms in every buildup step using the distances among them (most presumably given in the original distance data) and therefore, stopped the propagation of the rounding errors.

4. Concluding Remarks

In this paper, we have discussed a well-known problem in protein modeling, for the determination of the structure of a protein with a given set of inter-atomic or inter-residue distances obtained from either physical experiments or theoretical estimates. A more general and abstract form of the problem is known as the distance geometry problem in mathematics, the graph embedding problem in computer science, and the multidimensional scaling problem in statistics. In general, the problem can be stated as to find the coordinates for a set of points in some topological space given the distances for certain pairs of points. Therefore, in addition to protein modeling where everything is discussed only in three-dimensional Euclidean space, the problem has applications in many other scientific and engineering fields as well, such as sensor network localization, image recognition, and protein classification, to name a few. In any case, the problem may or may not have a solution in a given topological space, and even if it does have a solution, the solution may not be easy to find, depending on the given distances. For example, in any *k*-dimensional Euclidean space, the problem is polynomial time solvable if the distances for all the pairs of points are provided, and is NP-complete otherwise in general.

We have investigated the solution of the distance geometry problem within a so-called geometric buildup framework. Central to the geometric buildup approach is the idea to determine only a small group of atoms at the beginning and then complete the whole molecule by repeatedly determining one or more atoms every time using the available distances between the determined and undetermined atoms. The advantage of using a geometric buildup approach is that it works directly on the given distances and exploits the special structure of a given problem, and hence may be able to solve the problem more efficiently than a general approach. We have discussed the formulations and complexities of the distance geometry problem in its various forms, and described the general geometric buildup algorithm and its theoretical basis. We have also discussed the issues of the general algorithm for controlling rounding errors, determining rigid vs. unique structures, and handing inexact distances, and reviewed various versions of the algorithm that can address these issues and showed their test results.

A basic principle for the general geometric buildup algorithm is that whenever there are four determined atoms that are not in the same plane and there are distances from these atoms to an undetermined atom, the undetermined atom can immediately be determined uniquely by solving a system of four distance equations using the available distances. If for every atom, the required atoms and the distances can be found, the whole structure can be determined uniquely. The distance equations can in fact be reduced to a set of linear equations and hence solved in constant time. Therefore, as we have detailed in the paper, in ideal cases, a geometric buildup algorithm can solve a distance geometry problem with only $4n$ distances in $O(n)$ computing time, while the conventional singular value decomposition algorithm requires all $n(n\text{-}1)/2$ distances and $\mathrm{O}(n^2)$ computing time, where n is the number of atoms to be determined**.**

However, the requirement for four determined atoms and hence four corresponding distances in every step of the buildup procedure is sufficient but not necessary for the unique determination of a structure. Therefore, the general geometric buildup algorithm can in fact be modified so that in every buildup step, only three determined atoms and hence three corresponding distances are required. There may be multiple structures that can be determined in this way, but they are still rigid and can possibly end up unique as well. Indeed, as we have reviewed in the paper, a modified geometric buildup algorithm has been developed and tested successfully on a set of proteins. The results showed that the modified algorithm was able to produce meaningful structures rigidly with very sparse distance data, although they may be multiple in many cases.

The geometric buildup algorithm, either rigid or unique, can be sensitive to the numerical errors though, for the coordinates of the atoms are determined using the coordinates of previously determined atoms and the rounding errors in the previously determined atoms can be passed to and accumulated in later determined atoms, resulting in incorrect structural results. An updating scheme has been developed to prevent the accumulation of the numerical errors, as we have described in the paper. The idea of the scheme is based on the fact that the coordinates of any four atoms can be determined without any other information if all the distances among them are given. Therefore, the coordinates of any four determined atoms can be recalculated whenever possible using the distances among them, before they are used as a basis set of atoms for the determination of other atoms. The recalculated coordinates do not depend on the coordinates of previously determined atoms and therefore do not inherit any errors from them.

The general geometric buildup algorithm cannot tolerate errors in given distances either, for the distances then may not be consistent and the systems of distance equations may not be solvable. However, in practice, the distances must have errors because they come from either experimental measures or theoretical estimates. We have demonstrated how an extended geometric buildup algorithm can be developed to prevent the accumulation of the rounding errors in the buildup calculations successfully and also tolerate the errors in the given distances. In this algorithm, in every buildup step, all (instead of a subset of) the distances available for each unknown atom are taken into account for the determination of the position of the atom by using a least-squares approximation (instead of solving a system of equations exactly). We have shown that the least-squares approximation could actually be obtained by using a special singular value decomposition method, which could not only provide an approximate solution to the original system of distance equations, but also prevent the accumulation of the rounding errors in the buildup procedure effectively.

As we have discussed in the introduction section of the paper, a further complicated yet practical case of the distance geometry problem is when the distances are given with only their lower and upper bounds. The problem then becomes to find the coordinates $x_1, \ldots, x_n$ for the atoms for a given set of lower and upper bounds, $l_{i,j}$ and $u_{i,j}$, of the distances $d_{i,j}$ such that

$$l_{i,j} \leq \| x_i - x_j \| \leq u_{i,j}, \quad (i, j) \in S \cdot$$

The general geometric buildup algorithm and its modifications or extensions presented in this paper have not been developed to deal with distance bounds yet. However, the general buildup procedure should be extendable for the solution of such a problem as well. Here, different from other implementations, in every buildup step, an atom should be determined by satisfying a set of distance bounds instead of exact distances. The computation will certainly be more involved and subject to even more arbitrary errors. The solution to such a problem will not be unique, either. In fact, there can be an ensemble of solutions all satisfying the given distance inequalities. On the other hand, in practice, it is actually preferred to obtain the entire ensemble of solutions instead of a few samples. How to implement a buildup algorithm to achieve that can be challenging and will be the topic of our future investigation.

Acknowledgements

The authors would like to acknowledge the support from the Ogden College of Science and Engineering, Western Kentucky University, the Department of Mathematics and the Baker Center for Bioinformatics and Biological Statistics, Iowa State University, and the Laboratory for Scientific and Engineering Computing, Chinese Academy of Sciences. The work has been funded partially by the NIH/ /NIGMS grant R01GM081680, the NSF of China, and the Institute of Mathematics and Its Applications with funds provided by the NSF of America.

References

[1] T. E. Creighton, Proteins*: Structures and Molecular Properties, 2nd Edition*, Freeman and Company, 1993.

[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, L. N. Shindyalov, and P. E. Bourne, The Protein Data Bank, *Nuc. Acid. Res.*, **28**, 2000, 235-242.

[3] J. Drenth, Principles of Protein X-ray Crystallography, Springer-Verlag, 1994.

[4] H. Gunther, NMR Spectroscopy: Basic Principles, Concepts, and Applications in Chemistry, John Wiley & Sons, 1995.

[5] T. Schlick, Molecular Modeling and Simulation: An Interdisciplinary Guide, Springer, 2003.

[6] P. E. Bourne and H. Weissig, *Structural Bioinformatics*. John Wiley & Sons, Inc., 2003.

[7] J. Moult, K. Fidelis, B. Rost, T. Hubbard, and A. Tramontano, Critical assessment of methods for protein structure prediction (CASP), *Proteins: Structure, Function, Bioinformatics*, **61**, 2005, 3-7.

[8] V. S. Pande, I Baker, J. Chapman, S. P. Elmer, S. Khaliq, S. M. Larson, Y. M. Rhee, M. R. Shirts, C. D. Snow, E. J. Sorin, and B. Zagrovic, Atomistic protein folding simulations on submillisecond time scale using worldwide distributed computing, *Biopolymers*, **68**, 2003, 91-109.

[9] L. M. Blumenthal, *Theory and Applications of Distance Geometry*, Oxford Clarendon Press, 1953.

[10] J. B. Saxe, Embeddability of weighted graphs in k-space is strongly NP-hard, in *Proc. 17th Allerton Conference in Communications, Control and Computing*, 1979, 480-489.

[11] W. S. Torgerson, *Theory and Method of Scaling*, John Wiley & Sons, 1958.

[12] P. Biswas, T. Liang, T. Wang, and Y. Ye, Semidefinite programming based algorithms for sensor network localization. *ACM J on Transactions on Sensor Networks*, **2**, 2006, 188-220.

[13] H. Klock and J. M. Buhmann, Multidimensional scaling with deterministic annealing, in *Lecture Notes in Computer Science 1223: Energy Minimization Methods in Computer Vision and Patter Recognition*, M Pilillo and E. R. Hancock, eds., Springer-Verlag, 1997, 246-260.

[14] J. T. Hou, G. E. Sims, C. Zhang, and S. H. Kim, A global representation of the protein fold space, *Proc. Natl. Acad. Sci. USA*, **100,** 2003, 2386–2390.

[15] K. Wuthrich, *NMR of Proteins and Nucleic Acids*, John Wiley & Sons, 1986.

[16] G. M. Clore and A. M. Gronenborn, New methods of structure refinement for macromolecular structure determination by NMR, *Proc. Natl. Acad. Sci. USA*, **95**, 1998, 5891-5898.

[17] F. Cui, R. Jernigan, and Z. Wu, Refinement of NMR-determined protein structures with database derived distance constraints, *J Bioinformatics and Computational Biology*, **3**, 2005, 1315-1329.

[18] D. Wu, F. Cui, R. Jernigan, and Z. Wu, PIDD: A database for protein inter-atomic distance distributions, *Nucleic Acids Research*, **35**, 2007, D202-D207.

[19] D. Wu, R. Jernigan, and Z. Wu, Refinement of NMR-determined protein structures with database derived mean-force potentials, *Proteins: Structure, Function, Bioinformatics*, **68**, 232-242, 2007.

[20] G. M. Crippen and T. F. Havel, *Distance Geometry and Molecular Conformation*, John Wiley & Sons, 1988.

[21] T. F. Havel, Distance geometry, in *Encyclopedia of Nuclear Magnetic Resonance*, D. M. Grant and R. K. Harris, eds., John Wiley & Sons, 1995, 1701-1710.

[22] T. Havel, An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance, *Prog. Biophys. Molec. Biol.*, 56, 1991, 43-78.

[23] A. T. Brünger and M. Niles, Computational challenges for macromolecular modeling, in *Reviews in Computational Chemistry*, **5**, K. B. Lipkowitz and D. B. Boyd, eds., VCH Publishers, 1993, 299-335.

[24] Kuszewski, M. Niles, and A. T. Brünger, Sampling and efficiency of metric matrix distance geometry: A novel partial metrization algorithm, *J. Biomolecular NMR*, **2**, 1992, 33-56.

[25] A. T. Brünger, P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, . S. Jiang, J. Kuszewski, N. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T. Simonson, and G. L. Warren, Crystallography and NMR System: A new software suite for macromolecular structure determination, *Acta Cryst.*, **D54**, 1998, 905-921.

[26] T. F. Havel and M. E. Snow, A new method for building protein conformations from sequence alignments with homologues of known structure, *J. Mol. Biol*., **217,** 1991, 1-7.

[27] S. Srinivasan, C. J. March, and S. Sudarsanam, An automated method for modeling proteins on known templates using distance geometry, *Protein Science*, **2**, 1993, 277-289.

[28] E. S. Huang, R. Samudrala, and J. W., Ponder, Distance geometry generates native-like folds for small helical proteins using the consensus distances of predicted protein structures, *Protein Science*, **7**, 1998.

[29] G. A. Williams, J. M. Dugan, and R. B. Altman, Constrained global optimization for estimating molecular structure from atomic distances, *J. Comput. Biol.*, **8**, 2001, 523-547.

[30] B. Hendrickson, Conditions for unique graph realizations, *SIAM J. Comput*., **21**, 1992, 65-84.

[31] B. Hendrickson, The molecule problem: Exploiting structure in global optimization, *SIAM J. Optim*., **5**, 1995, 835-857.

[32] W. Glunt, T. L. Hayden, S. Hong, and J. Wells, An alternating projection algorithm for computing the nearest Euclidean distance matrix, SIAM *J. Mat. Anal. Appl*., **11**, 1990, 589-600.

[33] W. Glunt and T. L. Hayden and M. Raydan, Molecular conformations from distance matrices, *J. Comput. Chem.*, **14**, 1993, 114-120.

[34] J. Moré and Z. Wu, Global continuation for distance geometry problems, *SIAM J. Optim*., **7**, 1997, 814-836.

[35] J. Moré and Z. Wu, Distance geometry optimization for protein structures, *J. Global Optim*. **15**, 1999, 219-234.

[36] Z. Zou, R. H. Byrd, and R. B. Schnabel, A stochastic/perturbation global optimization algorithm for distance geometry problems", *J. Global Optim*., **11**, 1997, 91-105.

[37] A. Kearsly, R. Tapia, and M. Trosset, Solution of the metric STRESS and SSTRESS problems in multidimensional scaling by Newton’s method, *Computational Statistics* **13**, 1998, 369-396

[38] M. Trosset, Applications of multidimensional scaling to molecular conformation, *Computing Sciences and Statistics* **29**, 1998, 148-152.

[39] A. Le Thi Hoai and T. Pham Dinh, Large scale molecular optimization from distance matrices by a d.c. optimization approach, *SIAM J. Optim*., **4**, 2003, 77-116

[40] P. Biswas, T. Liang, K. Toh, and Y. Ye, A SDP based approach to anchor-free 3D graph realization, Department of Management Science and Engineering, Electrical Engineering, Stanford University, Stanford, California, 2007.

[41] A. Grosso, M. Locatelli, and F. Schoen, Solving molecular distance geometry problems by global optimization algorithms, J. Comput. Opt. and Appl., 2007, to appear.

[42] Q. Dong and Z. Wu, A linear-time algorithm for solving the molecular distance geometry problem with exact inter-atomic distances, *J. Global Optim.*, **22**, 2002, 365-375.

[43] Q. Dong and Z. Wu, A geometric buildup algorithm for solving the molecular distance geometry problem with sparse distance data, *J. Global Optim.*, **26**, 2003, 321-333.

[44] D. Wu and Z. Wu, An updated geometric buildup algorithm for solving the molecular distance geometry problem with sparse distance data, *J. Global Optim.*, **37**, 2007, 661-673.

[45] D. Wu, Z. Wu, and Y. Yuan, Rigid vs. unique determination of protein structures, *Optimization Letters*, (published online, DOI: 10.1007/s11590-007-0060-7), 2007.

[46] A. Sit, Z. Wu, and Y. Yuan, A stable geometric buildup algorithm for the solution of the distance geometry problem using east-squares approximation, 2007, submitted.

[47] G. H. Golub and C. F. van Loan, *Matrix Computations*, Johns Hopkins University Press, 1989.

[48] M. Sippl and H. Scheraga, Solution of the embedding problem and decomposition of symmetric matrices, *Proc. Natl. Acad. Sci. USA*, **82**, 1985, 2197-2201.

[49] M. Sippl and H. Scheraga, Cayley-Menger coordinates, *Proc. Natl. Acad. Sci. USA* **83**, 1986, 2283-2287.

[50] M. R. Garey and D. S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, W. H. Freeman & Co., 1979.

[51] J. Moré and Z. Wu, ε-Optimal solutions to distance geometry problems via global continuation, in *Global Minimization of Non-Convex Energy Functions: Molecular Conformation and Protein Folding*, P. M. Pardalos, D. Shalloway, and G. Xue, eds., American Mathematical Society, 1996, 151-168.

[52] H. X. Huang and Z. A. Liang, and P. Pardalos, Some properties for the Euclidean distance matrix and positive semi-definite matrix completion problems, *J. Global Optim.*, **25**, 2003, 3-21.