J Glob Optim (2007) 37:661–673
DOI 10.1007/s10898-006-9080-6

ORIGINAL ARTICLE

An updated geometric build-up algorithm for solving the molecular distance geometry problems with sparse distance data

Di Wu · Zhijun Wu

Received: 20 March 2006 / Accepted: 5 August 2006 / Published online: 7 September 2006

Abstract An updated geometric build-up algorithm is developed for solving the molecular distance geometry problem with a sparse set of inter-atomic distances. Different from the general geometric build-up algorithm, the updated algorithm re-computes the coordinates of the base atoms whenever necessary and possible. In this way, the errors introduced in solving the algebraic equations for the determination of the coordinates of the atoms are controlled in the intermediate computational steps. The method for re-computing the coordinates of the base atoms based on the estimation on the root-mean-square deviation (RMSD) is described. The results of applying the updated algorithm to a set of protein structure problems are presented. In many cases, the updated algorithm solves the problems with high accuracy when the results of the general algorithm are inadequate.

Keywords Protein structure determination · Distance geometry · Geometric build-up · Root-mean-square deviation

1 Introduction

The molecular distance geometry problem arises in the study of the structure of a molecule based on a given set of inter-atomic distances for the molecule. This problem has an important application in molecular biology and biochemistry and in particular, in protein structure prediction and determination (see Havel 1995; Yoon et al. 2002 for a general review). The distances between certain pairs of atoms in protein can often be determined based on our knowledge of various types of bond-lengths and bond-angles (Brooks III et al. 1988; Creighton 1993), or from nuclear magnetic resonance

D. Wu (✉) · Z. Wu
Program on Bioinformatics and Computational Biology, Department of Mathematics, Iowa State University, Ames, IA, USA
e-mail: diwu@iastate.edu

Z. Wu
e-mail: zhijun@iastate.edu

(NMR) experiments (Brüger and Niles 1993; Kuntz et al. 1993), or sometimes, through homology modeling (Havel and Snow 1991). Therefore, a natural approach for the determination of the structure of a protein is to solve a molecular distance geometry problem if a set of distance data for the protein is given. However, the molecular distance geometry problem is difficult to solve in general, especially since often in practice, only sparse and inexact distance data is available. Several algorithms have been developed to solve the problem, including for example the embed algorithm by Crippen and Havel (1988), the alternating projection algorithm by Glunt et al. (1990, 1993), the graph reduction algorithm by Hendrickson (1991, 1995), the multi-scaling algorithm by Trosset (1998) and Kearsly et al. (1998), the global smoothing algorithm by Moré and Wu (1996a,b, 1997a,b, 1999), etc. Most of these algorithms can provide an approximate solution to the problem, but often not to a desired accuracy. They are costly requiring intensive computation as well.

In their recent work, Dong and Wu (2002) proposed a new approach to the molecular distance geometry problem. This approach, called the geometric build-up approach, determines the coordinates of the atoms in the molecule one atom at a time repeatedly using a simple geometric relationship between determined and undetermined atoms, i.e., if an undetermined atom has known distances to four previously determined atoms and if the four atoms are not in the same plane, then it is a simple geometric fact that the coordinates of the undetermined atom can immediately be determined by using the four known distances (see also Huang et al. 2002 for more general discussions on these properties). If the exact distances between all pairs of atoms are given, this approach can determine the coordinates of n atoms in n steps or in other words, in order of n floating point operations, while a conventional singular-value decomposition algorithm (as used in the embed algorithm) requires at least order of n^2 floating point operations.

In this paper, we consider the solution of a molecular distance geometry problem with sparse but exact distance data by using a geometric build-up algorithm. For such a problem, since the data is sparse, the required distances may not be available when an atom is to be determined. The atom is then put aside until the distances become available after more atoms are determined. For this purpose, the algorithm is applied repeatedly to the undetermined atoms until all remaining ones are determined. Dong and Wu (2003) implemented such an algorithm, but they found that the algorithm is very sensitive to the numerical errors introduced in calculating the coordinates of the atoms. The reason is that the coordinates of the atoms are all determined using the coordinates of previously determined atoms, and 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 become incorrect, especially when the molecule is large, say with more than a thousand atoms. Note that this problem does not exist for the problem with all exact distances since in that case we can just use one set of determined atoms to determine all other atoms and there will not be a chance for the errors to get propagated.

In this paper, we describe a so-called updated geometric build-up algorithm for solving the molecular distance geometry problem with sparse but exact distance data. We show that using this algorithm the accumulation of the errors in calculating the coordinates of the atoms can be controlled and prevented. The idea for the algorithm is based on the fact that the coordinates of any four atoms can be determined without any other information as long as all distances among them are given. For this reason, the coordinates of any four determined atoms can be re-calculated whenever

possible using the distances among them if the distances are given. The re-calculated coordinates do not depend on the coordinates of previously determined atoms and therefore do not inherit any errors from them. In this way, the coordinates for many of the atoms can be "corrected", and the errors in the calculated coordinates can be prevented from growing into incorrect structural results. The re-calculated coordinates for the four atoms are independent of their original coordinates and are not related to the overall structure already built-up by the algorithm. However, they can be put back to the original structure by aligning them to their original locations with an appropriate translation and rotation.

The paper is organized as follows. In Sect. 2, we describe the general geometric build-up algorithm and discuss the related numerical issues and, in particular, the error accumulation issue in coordinate calculations. In Sect. 3, we present the updated algorithm in detail including the method to transform the independently calculated coordinates of the atoms to the coordinate system of the rest of the structure through a simple root-mean-squares deviation (RMSD) calculation. In Sect. 4, we present some numerical results from applying the updated algorithm to ten protein structure problems and compare them with a general geometric build-up algorithm. We conclude the paper and make some remarks in Sect. 5.

2 The general geometric build-up algorithm

A geometric build-up algorithm for solving the molecular distance geometry problem given the exact distances between all pairs of atoms in the molecule is outlined in Fig. 1. There are two parts in the algorithm. The first one is to select four initial atoms that are not in the same plane and find a set of coordinates for the atoms using the distances among them. Let us call the atoms the base atoms. After a set of base atoms is selected and allocated, the second part of the algorithm is to find the coordinates for each of the remaining atoms using the distances from the atoms to the four base ones. The first part of the algorithm is based on the fact that the coordinates of four atoms can be determined if all distances among them are given, while the second part is that the coordinates of an atom can be determined if the distances from the atom to four determined atoms are given. In both cases, the coordinates can be determined through simple algebraic calculations and in particular, for the latter case, through the solution of a small system of algebraic equations. We state these facts in a more rigorous form in the following theorems.

Theorem 2.1 *If the distances among four atoms are given, the coordinates of the atoms can then be determined with the given distances, subject to translation, rotation, and reflection.*

The Geometric Build-up Algorithm for Problems with All Exact Distances

1. Find four base atoms that are not in the same plane; determine the coordinates of the base atoms with the distances among them.
2. For each of the remaining atoms determine the coordinates of the atom with its distances to the base atoms.
3. All atoms are determined.

Fig. 1 The outline of the general geometric build-up algorithm for solving the molecular distance geometry problem with all exact distances (Dong and Wu 2002)

Proof Let $x_i = (u_i,\ v_i,\ w_i)^T, i = 1,2,3,4$, be the coordinate vectors of the four atoms. Let $d_{i,j}$ be the given distances between atoms i and j for $i,j = 1,2,3,4$. The coordinates can then be determined as follows, based on the given distances.

First, since the atoms can be allocated in an arbitrary coordinate system, without loss of generality, we set a system with the first atom at its origin, the second on its x-axis, and the third on its xy-plane. Then, we have in this system that $u_1 = 0$, $v_1 = 0$, $w_1 = 0$, $v_2 = 0$, $w_2 = 0$, and $w_3 = 0$. Since the distance from the second atom to the first atom is equal to $d_{2,1}$, we have also that $u_2 = d_{2,1}$, and the first two atoms are then determined.

Since the distances from the third atom to the first and second atoms are equal to $d_{3,1}$ and $d_{3,2}$, respectively, then

$$\begin{aligned} u_3^2 + v_3^2 &= d_{3,1}^2, \\ (u_3 - u_2)^2 + v_3^2 &= d_{3,2}^2. \end{aligned}$$

Solve the equations for u_3 and v_3. We obtain

$$\begin{aligned} u_3 &= (d_{3,1}^2 - d_{3,2}^2)/(2u_2) + u_2/2 \\ v_3 &= \pm(d_{3,1}^2 - u_3^2)^{1/2} \end{aligned}$$

and the third atom is then determined by choosing v_3 either positive or negative.

Finally, with the distances, $d_{4,1}$, $d_{4,2}$, $d_{4,3}$, from the fourth atom to the first three atoms, we can form three equations,

$$\begin{aligned} u_4^2 + v_4^2 + w_4^2 &= d_{4,1}^2, \\ (u_4 - u_2)^2 + v_4^2 + w_4^2 &= d_{4,2}^2, \\ (u_4 - u_3)^2 + (v_4 - v_3)^2 + w_4^2 &= d_{4,3}^2. \end{aligned}$$

The coordinates u_4, v_4, w_4 for the fourth atom can then be determined by solving the equations, and

$$\begin{aligned} u_4 &= (d_{4,1}^2 - d_{4,2}^2)/(2u_2) + u_2/2, \\ v_4 &= (d_{4,2}^2 - d_{4,3}^2 - (u_4 - u_2)^2 + (u_4 - u_3)^2)/(2v_3) + v_3/2, \\ w_4 &= \pm(d_{4,1}^2 - u_4^2 - v_4^2)^{1/2}. \end{aligned}$$

This completes the proof for Theorem 2.1. □

Theorem 2.2 *If the coordinates of four atoms that are not in the same plane and the distances from the fifth atom to the four atoms are given, the coordinates of the fifth atom can be determined uniquely.*

Proof Let $x_i = (u_i, v_i, w_i)^T, i = 1,2,3,4$, be the coordinate vectors of the first four atoms and $x_j = (u_j, v_j, w_j)^T$ the coordinate vector of the fifth atom with an arbitrary index j. Let $d_{i,j}$ be the given distances from any of the first four atoms i to the fifth atom j for $i = 1,2,3,4$. We then have a set of equations,

$$||x_i - x_j|| = d_{i,j}, \quad i = 1,\ 2,\ 3,\ 4.$$

Square the equations and expand their left-hand-sides to obtain

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

Subtract the first equation from the rest to reduce the equations to the following three,

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

Let A be a matrix and b a vector, and

$$A = -2\begin{bmatrix} (x_2 - x_1)^T \\ (x_3 - x_1)^T \\ (x_4 - x_1)^T \end{bmatrix},$$

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

We can then write the above equations in the following matrix form.

$$A\, x_j = b.$$

Since x_1, x_2, x_3, x_4 are not in the same plane, the matrix A is nonsingular and therefore, the linear system of equations can be solved to obtain a unique solution for x_j. □

Note that Theorems 2.1 and 2.2 both assume that the given distances are accurate and consistent, and are true distances among a set of points. Given such distances, the coordinates of the atoms can obviously be determined by using the algorithm described in Fig. 1 based on the two theorems. Moreover, it can be proved that the coordinates of the atoms for a molecule of n atoms can be determined in n steps, each for one atom, as stated in the following theorem.

Theorem 2.3 *The general geometric build-up algorithm solves a molecular distance geometry problem with all exact distances for a molecule of n atoms in order of n floating point operations or in other words, in linear time in n.*

Proof As shown in Fig. 1, once the base atoms are determined, the remaining atoms are determined using the distances from the atoms to the base atoms, each requiring the solution of a small linear system of equations based on Theorem 2.2. Solving the linear system can be done in constant time, so for all remaining atoms, the time for determining them all is proportional to the number of atoms, $n-4$. The determination of the coordinates of the base atoms does not cost more than constant time, but to make sure the base atoms are not in the same plane may take longer time. In the worst case, the latter may take order of n computing time to examine through the entire atom list to find the third atom that is not in the line formed by the first two atoms ($v_3 \neq 0$) and then the fourth atom that is not in the plane formed by the first three atoms ($w_4 \neq 0$). In any case, the algorithm requires order of n floating-point operations or in other words, linear time in n to find the coordinates of all n atoms.

□

We now consider the case when only a subset of all distances among the atoms is available. The problem can be called one with sparse exact distances. In this case, the algorithm in Fig. 1 will not work since the required distances from the base atoms to the atom to be determined may not be available. However, the distances from other determined atoms to the atom may be available and may suffice for the determination of the atom. Therefore, the algorithm can be modified to cover the sparse case by determining the coordinates of an atom using any determined atoms as long as they can serve as its base atoms. Such a modified algorithm is outlined in Fig. 2.

The Geometric Build-Up Algorithm for Problems with Sparse Exact Distances

1. Find four base atoms that are not in the same plane;
   determine the coordinates of the base atoms with the distances among them.
2. Repeat:
   For each of the remaining atoms,
      find four determined atoms that can serve as its base atoms;
      determine the coordinates of the atom with its distances to the base atoms.
   End
   If no atom is determined in the whole loop, stop.
3. All atoms are determined.

Fig. 2 The outline of the general geometric build-up algorithm for solving the molecular distance geometry problem with sparse exact distances (Dong and Wu 2003)

Note that when only a sparse set of distances is given, the molecular distance geometry problem becomes difficult to solve in general. We do not expect to have a polynomial time algorithm for the problem, since Saxe (1979) has proved that the problem actually becomes *NP*-complete. Also, in the algorithm outlined in Fig. 2, the four qualified base atoms may not be available in the first step anyway; the for-loop in the second step may be repeated many times until all remaining atoms can be determined. However, Dong and Wu (2003) demonstrated that for protein structure determination, the algorithm seemed to be a reasonable one. When the distances less than 8 Å were used, reasonable structures for a set of tested proteins with up to 4,200 atoms were obtained by using such an algorithm. A numerical problem in this algorithm, as pointed out in Dong and Wu (2003), is that the base atoms that are used to determine an atom are determined themselves by some other base atoms in previous steps. The errors introduced in previous steps are thus passed to the current atom, and to the atoms in later steps as well. This may cause a completely incorrect result in the coordinates of the atoms. The errors in calculating the coordinates of an atom usually come from solving the linear system of equations, especially if the coefficient matrix A is ill formed. The matrix A is determined by the coordinates of the base atoms as shown in the proof for Theorem 2.2. Therefore, in Dong and Wu (2003), if the determinant of A is found small, a different set of base atoms would be used to avoid possible errors due to this matrix A, which resolved the problem for some of the test cases, but not for all.

3 The updated geometric build-up algorithm

In this section we describe the updated geometric build-up algorithm. The algorithm is a modified version of the general algorithm for problems with sparse exact distances. Two new strategies are used to minimize the errors introduced in the coordinate calculations. First, the condition number instead of the determinant of matrix A is examined when solving each of the linear systems in the algorithm. When the condition number is too big, a different set of base atoms is sought to avoid the possible errors due to an ill-conditioned matrix A. This is better than evaluating the determinant since a matrix can still be ill conditioned even if its determinant is large. Second, the coordinates of four determined atoms are re-calculated or re-initialized by the procedure described in Theorem 2.1, whenever the four atoms are found that they have all distances

available among them. Since they are independent of the coordinates of previously determined atoms, the re-calculated coordinates do not have the errors accumulated from previous calculations and hence re-calculation of coordinates reduces the chance of error accumulation. As described in the proof for Theorem 2.1, the re-calculated coordinates are represented in a new coordinate system with one atom located in the origin, another along the x-axis, etc. However, the atoms can be put back to the original structure by aligning their new coordinates with the old ones, using an appropriate translation and rotation for the new coordinates, so that the RMSD between the new coordinates and the old ones is minimized. The translation vector and the rotation matrix can be obtained exactly in the same way as in regular RMSD calculations.

Figure 3 is an outline of the updated algorithm. We call it updated since the coordinates are updated repeatedly in the algorithm to prevent errors. The way we calculate the RMSD of two structures (defined in terms of their Cartesian coordinates) is the following. Let X and Y be the coordinate matrices of two structures after they are translated so that their centers of geometry coincide. The RMSD of the two structures is then defined as

$$\text{RMSD}\,(X, Y) = \min_Q \, ||X - YQ||_F/\sqrt{n},$$

where Q is a rotation matrix and $QQ^T = I$. Let $C = Y^T X$, 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 minimization problem (Golub and Van Loan, 1989). Therefore, computationally, we can first compute the geometric centers of the two structures,

$$xc = \frac{1}{n}\sum_{i=1}^{n} X(i,:), \quad yc = \frac{1}{n}\sum_{i=1}^{n} Y(i,:).$$

We then update matrix Y,

$$\begin{aligned} Y(:,1) &= Y(:,1) - [yc(1) - xc(1)], \\ Y(:,2) &= Y(:,2) - [yc(2) - xc(2)], \\ Y(:,3) &= Y(:,3) - [yc(3) - xc(3)]. \end{aligned}$$

The two structures now have the same geometric center. We then compute the matrix $C = Y^T X$ and its singular-value decomposition $C = U\Sigma V^T$. Let $Q = UV^T$.

The Updated Geometric Build-Up Algorithm for Problems with Sparse Exact Distances

1. Find four base atoms that are not in the same plane;
   determine the coordinates of the base atoms with the distances among them.
2. Repeat:
   For each of the remaining atoms,
   find four determined atoms that can serve as its base atoms;
   determine the coordinates of the atom with its distances to the base atoms.
   If four determined atoms are found having all distances among them,
   re-initialize the coordinates of the four atoms;
   put the atoms back to the original structure.
   End
   End
   If no atom is determined in the whole loop, stop.
3. All atoms are determined.

Fig. 3 The outline of the updated geometric build-up algorithm for solving the molecular distance geometry problem with sparse exact distances

The RMSD of the two structures can then be calculated as

$$\text{RMSD}\,(X, Y) = ||X - YQ||_F/\sqrt{n}.$$

In the updated algorithm, every time the coordinates of four atoms are re-calculated, if X contains the old coordinates and Y the new ones, YQ in the above formula gives the coordinates best aligned with the old ones.

4 Numerical results

We have implemented an updated geometric build-up algorithm in Matlab (Version 5.3). The matrix-vector calculations required in the algorithm including linear system solves, estimations of condition numbers, and singular-value decompositions are all done through the Matlab build-in functions. We tested the algorithm with a set of problems generated using the known structures of ten proteins downloaded from the PDB data bank (Berman et al. 2000). Each of the structures is used to obtain two sets of distances, one including all distances ≤ 5 Å and another ≤ 8 Å. We then solve a molecular distance geometry problem for each set of distances using the updated algorithm, to obtain the coordinates of the atoms for the corresponding protein. The result is compared with the original structure of the protein in terms of RMSD. The choice of 5 Å as the cut-off distance is made to simulate the distance data in NMR experiments since in most cases, NMR can only detect the distances between atoms in that range. The choice of 8 Å is to make a relaxation on the cut-off to observe the performance difference of the algorithm under a different condition. Note that in practice, NMR actually can provide only lower and upper bounds of the distances. However, in this work, we only consider problems with exact distances. The extension of the algorithm to problems with distance bounds is possible and under another line of investigation.

Table 1 contains the results of using the updated geometric build-up algorithm for solving the generated test problems. They are also compared with the results of using the general geometric build-up algorithm for the same set of problems obtained by Dong and Wu (2003). The first column of the table contains the names of the proteins in the PDB Data Bank. The second column contains the numbers of atoms in the proteins. The remaining columns list the results of using the updated and general algorithms for problems with 5 and 8 Å distance cut-offs. The results for each problem include the number of fixed atoms and the RMSD for the fixed structure compared with the original one. For the ten tested structures, five of them were determined with the general algorithm with the distances less than 8 Å, but none with less than 5 Å. However, nine of the ten structures were determined with the updated algorithm with the distances less than 8 Å, and five of them were determined with the distances less than 5 Å. For the updated algorithm, we also list the results for problems that were not completely resolved by the algorithm with the distances less than 5 Å. They include 1PHT, 1AX8, 1RGS, 1BPM, and 1HMV. These problems are relatively large, but for four of them, the algorithm actually was able to determine the coordinates for almost all the atoms. For 1PHT only 5 out of 814 atoms were not fixed, and for 1RGS only 5 out of 2,015, for 1BPM only 3 out of 3,674, and for 1HMV only 13 out of 4,201. We have examined the atoms that were not fixed by the algorithm and found that in many cases, the atoms are in the side chains of the proteins and do not have enough neighboring atoms within 5 Å distance. For example, in 1PHT, the unfixed atoms are

Table 1 Results of using the updated and general geometric build-up algorithms for solving a set of molecular distance geometry problems generated from ten known protein structures downloaded from the PDB Data Bank

Protein	No. of atom	5 Å				8 Å			
		Updated		General		Updated		General	
		No. of fixed atom	RMSD	No. of fixed atom	RMSD	No. of fixed atom	RMSD	No. of fixed atom	RMSD
1PTQ	404	404	2.7e-12	–	–	404	3.5e-13	402	2.8e-8
1HOE	558	558	8.2e-13	–	–	558	1.0e-11	558	9.4e-6
1LFB	641	641	9.5e-12	–	–	641	3.9e-12	–	–
1F39A	767	767	3.5e-11	–	–	767	2.4e-12	767	2.3e-6
1PHT	814	809	7.9e-9	–	–	814	1.8e-12	814	4.4e-5
1POA	914	914	6.8e-10	–	–	914	1.7e-11	–	–
1AX8	1,003	–	–	–	–	998	3.5e-12	1,003	1.5e-6
1RGS	2,015	2,010	7.4e-8	–	–	2,015	1.1e-9	–	–
1BPM	3,674	3,671	1.8e-9	–	–	3,674	3.2e-7	–	–
1HMV	4,201	4,188	6.8e-11	–	–	4,201	2.5e-5	–	–

located in the side chain of LYS, where there are not enough distances to determine the atoms. The structure 1AX8 seems difficult to determine probably because it is a double helix and is lack of enough distance information among the atoms. We include this instance in the table to show the possible difficult case for the algorithm. There are two odds in the table requiring some explanations as well. First, for 1PTQ, with an 8 Å cut-off, the number of fixed atoms for the general algorithm is 402 instead of 404. This is because that there are two het atoms in the structure that were not considered in the experiment with the general algorithm. Second, for 1AX8, with an 8 Å cut-off, 988 out of 1,003 atoms were determined using the updated algorithm, but all atoms were determined using the general algorithm. This may be because of the specific structure of the molecule or the specific implementation of the two algorithms, but it does not reflect the general behaviors of the algorithms.

Figures 4 and 5 further demonstrate in some worst-case scenarios how the structure determined by a geometric build-up algorithm can be affected by the accumulated numerical errors. The figures show the structures (red lines) of protein 4MBA (1,086 atoms) determined using ≤ 5 Å distances, first by a general algorithm and then by the updated algorithm. The pictures show clearly that the general algorithm results in a structure (red lines in Fig. 4) that disagrees with the original structure (blue lines) in many regions, while the updated algorithm determines one (red lines in Fig. 5) that agrees with the original structure (blue lines) almost completely.

Finally, Fig. 6 further shows how the numerical error grows as the geometric build-up algorithm proceeds. Shown in the figure is the RMSD of the computed structure for 4MBA compared with its original structure as a function of th e size (the number of atoms) of the computed structure. For a general geometric build-up algorithm, from around 300 atoms, the RMSD (the green line) starts increasing rapidly, and in the end, the RMSD for the entire structure (with 1,086 atoms) becomes bigger than 10 Å. On the other hand, for the updated algorithm, the RMSD (the blue line) is bounded in around $5.0e-04$ Å in the whole build-up procedure.

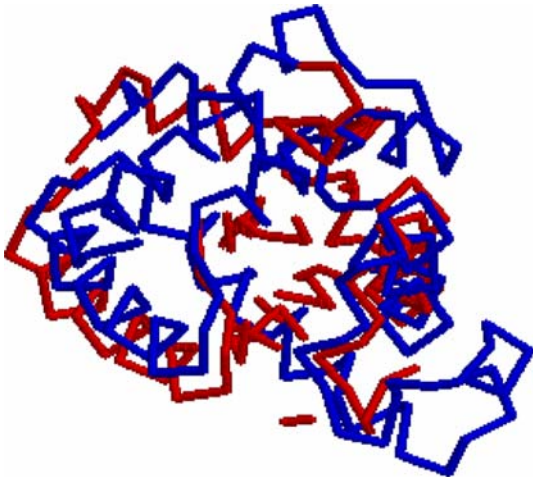

Fig. 4 The structure (*red lines*) of 4MBA determined by using a general geometric build-up algorithm and compared with the original structure of 4MBA (*blue lines*). Here, 4MBA is the PDB entry for the crystal structure of the ferric form of myoglobin from the mollusc Aplysia limacina refined at 1.6 Å resolution, by restrained crystallographic refinement methods. The crystallographic *R*-factor is 0.19. The tertiary structure of the molecule conforms to the common globin fold, consisting of eight alpha-helices. The *N*-terminal helix *A* and helix *G* deviate significantly from linearity. See Bolognesi et al. (1989) for more details.

5 Summary and remarks

The molecular distance geometry problem has an important application in macromolecular modeling and in particular, in protein structure determination. The problem is difficult to solve especially in practice when only sparse and inexact distances are given. In this paper, we consider the solution of a molecular distance geometry problem with sparse but exact distance data by using a geometric build-up algorithm. For such a problem, since the data is sparse, the coordinates of the atoms cannot be determined with only one set of base atoms since the required distances between the base atoms and the atom to be determined may not be available. Therefore, in most cases, the atoms are determined using a set of base atoms that are determined in previous steps. Dong and Wu (2003) implemented such an algorithm, but they found that the algorithm is very sensitive to the numerical errors introduced in calculating the coordinates of the atoms. The reason is that the coordinates of the atoms depend on the coordinates of previously determined atoms, and 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 become incorrect, especially when the molecule is large, say with more than a thousand atoms.

In this paper, we have introduced an updated geometric build-up algorithm for solving the molecular distance geometry problem with sparse but exact distance data. We have shown that using this algorithm the accumulation of the errors in calculating the coordinates of the atoms could be controlled and prevented. The idea for the updated algorithm is based on the fact that the coordinates of any four atoms can be determined without any other information as long as all distances among them are given. Therefore, the coordinates of any four determined atoms can be re-calculated

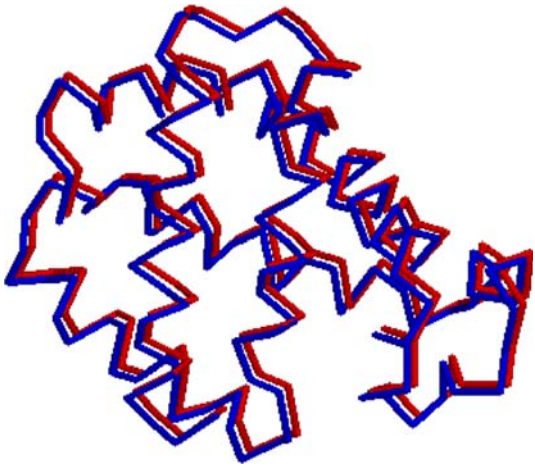

Fig. 5 The structure (*red lines*) of 4MBA determined by using an updated geometric build-up algorithm and compared with the original structure of 4MBA (*blue lines*)

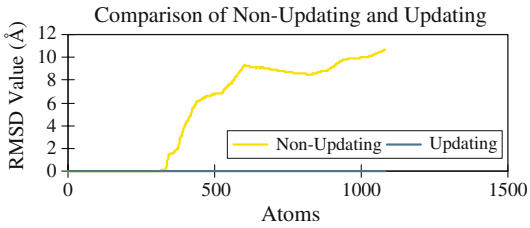

Fig. 6 The RMSD (Å) of the computed structure of 4MBA compared with its original structure as a function of the size (number of atoms) of the computed structure

whenever possible using the distances among them if the distances are given. The re-calculated coordinates do not depend on the coordinates of previously determined atoms and therefore do not inherit any errors from them. In this way, the coordinates for many of the atoms can be "corrected", and the errors in the calculated coordinates can be prevented from growing into incorrect structural results.

We have described the general geometric build-up algorithm with a presentation that is more formal than that of other papers. Several important properties related to the algorithm are stated as theorems and formal proofs are also given. Some of them are the foundations for the development of the general as well as updated geometric build-up algorithms. We have discussed the numerical issues associated with the general geometric build-up algorithm and presented the updated algorithm including the procedure for re-evaluating the coordinates and the method for updating the old coordinates with the new ones through RMSD calculation. We have presented numerical results of using the updated algorithm for a set of test problems generated with known protein structures. The results for two sets of problems have been obtained, one with distances less than or equal to 5 Å and another 8 Å. The results showed that the updated algorithm determined the structures for most of the problems while the general algorithm failed.

The algorithm discussed in this paper may be of only theoretical value in a certain sense since in practice the given distances usually are inexact and the algorithm may only be used for solving a sub-problem. However, the algorithm represents a significant advance in solving a general molecular distance geometry problem. It can

certainly be modified and extended to problems with inexact distances. Work in this direction is being pursued and will be reported later elsewhere.

Acknowledgements We would like to thank Peter Vedell for reading the paper and offering helpful suggestions. The support for the first author from the ISU Graduate Program on Bioinformatics and Computational Biology is also gratefully acknowledged.

References

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, L.N., Bourne, P.E.: The protein data bank. Nuc. Acid. Res. **28**, 235–242 (2000)

Bolognesi, M., Onesti, S., Gatti, G., Coda, A., Ascenzi, P., Brunori, M.: Aplysia limacina myoglobin: crystallographic analysis at 1.6 Å resolution. J. Mol. Biol. **205**, 529–544 (1989)

Brooks III, C.L., Karplus, M., Pettitt, B.M.: Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics. Wiley, New York (1988)

Brüger, A.T., Niles, M.: Computational challenges for macromolecular modeling. In: Lipkowitz, K.B., Boyd, D. B. (eds.), Reviews in Computational Chemistry, vol. 5, pp. 299–335. VCH Publishers, Wcinheim (1993)

Creighton, T.E.: Proteins: Structures and Molecular Properties, 2nd edn. Freeman and Company, San Franscisco, CA, New York (1993)

Crippen, G.M., Havel, T.F.: Distance Geometry and Molecular Conformation. Wiley, New York (1988).

Dong, Q., Wu, Z.: A linear-time algorithm for solving the molecular distance geometry problem with exact inter-atomic distances. J. Global Optim. **22**, 365–375 (2002)

Dong, Q., Wu, Z.: A geometric build-up algorithm for solving the molecular distance geometry problem with sparse distance data. J. Global. Optim. **26**, 321–333 (2003)

Glunt, W., Hayden, T.L., Hong, S., Wells, J., An alternating projection algorithm for computing the nearest euclidean distance matrix. SIAM J. Mat. Anal. Appl. **11**(4), 589–600 (1990)

Glunt, W., Hayden, T.L., Raydan, M.: Molecular conformations from distance matrices. J. Comput. Chem. **14**(1), 114–120 (1993)

Golub, G.H., van Loan, C.F.: Matrix Computations. Johns Hopkins University Press, Baltimore, MD (1989)

Havel, T.F.: Distance geometry. In: Grant, D.M., Harris, R.K., (eds.), Encyclopedia of Nuclear Magnetic Resonance, pp. 1701–1710. Wiley, New York (1995)

Havel, T.F., Snow, M.E.: A new method for building protein conformations from sequence alignments with homologues of known structure. J. Mol. Biol. **217**, 1–7 (1991)

Hendrickson, B.A.: The molecular problem: determining conformation from pairwise distances. Ph.D. thesis, Cornell University, Ithaca, NY (1991)

Hendrickson, B.A.: The molecule problem: exploiting structure in global optimization. SIAM J. Optim. **5**(4), 835–857 (1995)

Huang, H.X., Liang, Z.A., Pardalos, P.: Some Properties for the Euclidean Distance Matrix and Positive Semi-Definite Matrix Completion Problems. Department of Industrial and Systems Engineering, University of Florida (2002)

Kearsly, A., Tapia, R., Trosset, M.: Solution of the metric STRESS and SSTRESS problems in multidimensional scaling by Newton's method. Comput. Stat. **13**, 369–396 (1998)

Kuntz, I.D., Thomason, J.F., Oshiro, C.M.: Distance geometry. In: Oppenheimer, N.J., James, T.L. (eds.), Methods in Enzymology, vol. 177, pp. 159–204. Academic Press, New York (1993)

Moré, J., Wu, Z.: ε-Optimal solutions to distance geometry problems via global continuation. In: Pardalos, P.M., Shalloway, D., Xue, G. (eds.), Global Minimization of Non-Convex Energy Functions: Molecular Conformation and Protein Folding, pp. 151–168. American Mathematical Society, Providence, RI (1996a)

Moré, J. Wu, Z.: Smoothing techniques for macromolecular global optimization. In: Di Pillo G., Gianessi, F. (eds.), Nonlinear Optimization and Applications, pp. 297–312. Plenum Press, New York (1996b)

Moré, J., Wu, Z.: Global continuation for distance geometry problems. SIAM J. Optim. **7**(3), 814–836 (1997a)

Moré, J. Wu, Z.: Issues in large scale global molecular optimization. In: Biegler, L., Coleman, T., Conn, A., Santosa, F. (eds.), Large Scale Optimization with Applications, pp. 99–122. Springer-Verlag, Berlin (1997b)

Moré, J., Wu, Z.: Distance geometry optimization for protein structures. J. Global Optim. **15**, 219–234 (1999)
Saxe, J. B.: Embeddability of weighted graphs in k-space is strongly NP-hard. In Proc. 17th Allerton Conference in Communications, Control and Computing, pp. 480–489 (1979)
Trosset, M.: Applications of multidimensional scaling to molecular conformation. Comput. Sci. Stat. **29**, 148–152 (1998)
Yoon, J., Gad, Y., Wu, Z., Mathematical modeling of protein structure with distance geometry, to appear. In: Yuan, Y., et al. (eds), Numerical Linear Algebra and Optimization, Scientific Press, (2002)