RIGID VS. UNIQUE DETERMINATION OF PROTEIN STRUCTURES WITH GEOMETRIC BUILDUP

By

Di Wu

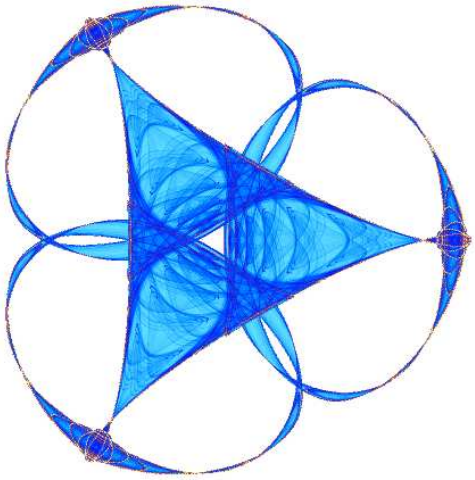
Zhijun Wu

and

Yaxiang Yuan

IMA Preprint Series # 2175

(October 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

Rigid vs. Unique Determination of Protein Structures with Geometric Buildup*

Di Wu[1], Zhijun Wu[2], and Yaxiang Yuan[3]

[1]Department of Mathematics
West Kentucky University
Bowling Green, Kentucky, USA

[2]Department of Mathematics
Program on Bioinformatics and Computational Biology
Iowa State University
Ames, Iowa, USA

[2]Laboratory of Scientific and Engineering Computing
Chinese Academy of Science
Beijing, China

Abstract. We introduce a geometric buildup approach to the distance geometry problem in protein modeling, and discuss the necessary and sufficient conditions on the distances for rigid or unique determination of a protein structure. We describe a new buildup algorithm for determining protein structures rigidly instead of uniquely. The algorithm requires even fewer distance constraints than the general buildup algorithm. We present the test results from applying the algorithm to determining the protein structures with varying degrees of availability of the distances, and show that the new development increases the modeling ability of the geometric buildup method even more while retaining much of the computational feasibility of the method.

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

*

1. Introduction 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), and dipolar coupling in NMR, or theoretical estimates such as the bond lengths and bond angles known from general organic chemistry [1][2], or statistical estimates on certain inter-atomic or inter-residue distances based on their distributions in databases of known protein structures [3][4][5]. Then, a structure may be determined for the protein by using the available distances [6]. 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 [7]. 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 points. In that case, a structure that fits the given distances will not even exist [8]. After all, even if a structure does exist, it is still not trivial to determine it based on the given distances. A distance geometry problem needs to be solved, which is computationally intractable in general [9].

We investigate the problem of determining a protein structure with a given set of inter-atomic or inter-residue distances within a so-called geometric build-up framework. Dong and Wu [10][11] first applied a geometric build-up algorithm for the solution of the distance geometry problem with exact distances and justified the linear computation time for the case when the distances for all pairs of atoms are given. Wu and Wu [12] later proposed an updating scheme to control the rounding errors accumulated in the buildup procedure and guaranteed the numerical stability of the algorithm. Central to the algorithm is the idea 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 using the distances. If for every atom, the required atoms and the distances can be found, the whole structure can be determined uniquely [10][12]. Here the condition is sufficient but not necessary, for there are cases we will elaborate later that a structure can still be determined uniquely even if the condition does not hold for some of the atoms. In fact, the structure does not have to be unique, as long as it is rigid. For this reason, we can consider a weaker condition under which a structure can be determined rigidly and under certain circumstance, even uniquely. With this condition, the minimal requirement on the availability of the distances in every buildup step can be dropped from four to three, which implies that a structure can still be determined even with a much sparser set of distances. Along the line, we develop a new buildup algorithm, which determines the atom just rigidly in every step, if there are only three required distances available. The position of the atom may have multiple reflections, but can be fixed uniquely if later on some distance constraints are found to be violated by its multiple positions. With such an algorithm, a rigid structure can be guaranteed in the end. It may or may not be unique. In any case, different from Crippen and Havel [6], we only use the available distances to determine the structure, but do not consider extrapolating the missing distances. Also, different from Hendrickson [7], we use all given distances as many as necessary, and do not remove any distances even if they may be redundant for defining a rigid structure since they may still be useful for the determination of some of the atoms in the buildup process. We develop algorithms for generation of multiple rigid structures, for identification of unique structures, and for combination of partially determined structures. We apply the algorithms

to determining the protein structures with varying degrees of availability of the distances and justify the increase in the modeling ability of the geometric buildup method with the new development.

2. General Geometric Buildup Approach Dong and Wu [10] applied a geometric build-up 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 [11] with an updating scheme to control the propagation of numerical errors in the buildup process [12]. Central to the algorithm is the idea 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 using the distances. If for every atom, the required atoms and distances can be found, the whole structure can be determined uniquely (Figure 1).

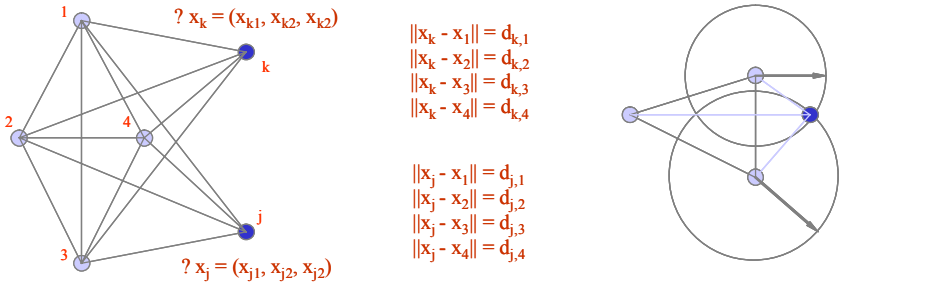


Figure 1 **Geometric buildup** Central to the algorithm is the idea that whenever there are four determined atoms that are not in the same plane and there are distances from these atoms to an undetermined atom, the undetermined atom can immediately be determined uniquely using the distances. If for every atom, the required atoms and the distances can be found, the whole structure can be determined uniquely.

The General Geometric Buildup Algorithm

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 4 distances to the determined atoms,
         Determine the atom uniquely with the distances.
      End
   End
4. If no atom can be determined in the loop, stop.
5. All atoms are determined.

More specifically, 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 the singular value decomposition algorithm [6] 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. Indeed, x_j can be obtained from the solution of the following system of equations,

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

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

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

Let A be a matrix and b a vector, and

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

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.

The theoretical basis of the geometric buildup algorithm can be traced back in distance geometry [13]. Several authors had discussions on theoretical issues related to such an approach, including Sippl and Scheraga [14][15] and Huang, Liang, and Pardalos [16]. Based on 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.

Definition 2.1 A set of points B in a space S is a metric basis of S provided each point of S is uniquely determined by its distances from the points in B.

Definition 2.2 A set of k+1 points in R^k is called independent if it is not a set of points in R^{k-1}.

Theorem 2.1 Any k+1 independent points in R^k form a metric basis for R^k.

Proof It can be proved by generalizing the basic geometric buildup step to the k-dimensional Euclidean space. □

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 2.2 A necessary condition for the unique determination of the coordinates of a set of atoms, $x_1, \ldots, x_n$, with a given set of distances among the atoms is that each atom must have at least four distances from other four atoms, assuming that this atom is not in the same plane with any three of them.

Proof It follows from the fact that the position of an atom can have a reflection if it has only three distances from other three atoms unless it is in the same plane with the three atoms. □

Theorem 2.3 A sufficient condition for the unique determination of the coordinates of a set of atoms, $x_1, \ldots, x_n$, with a given set of distances among the atoms is that in every step of the geometric buildup algorithm, there is an undetermined atom with four distances from four determined atoms that are not in the same plane.

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 atoms uniquely. □

3. Rigid Structure Determination 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 specifically, 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 available from atoms $i = 1, 2, 3$ to atom j. Then, x_j can be obtained from the solution of the following system of equations,

$$\| x_i \|^2 - 2x_i^T x_j + \|x_j\|^2 = d_{i,j}^2, \quad i = 1, 2, 3. \tag{5}$$

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

$$\begin{aligned} &-2(x_{i+1} - x_i)^T x_j \\ &= (d_{i+1,j}^2 - d_{i,j}^2) - (\| x_{i+1} \|^2 - \|x_i\|^2), \quad i = 1, 2. \end{aligned} \tag{6}$$

Let A be a matrix and b a vector, and

$$A = -2\begin{bmatrix} (x_2 - x_1)^T \\ (x_3 - x_2)^T \end{bmatrix}, \quad b = \begin{bmatrix} (d_{2,j}^2 - d_{1,j}^2) - (\| x_2 \|^2 - \|x_1\|^2) \\ (d_{3,j}^2 - d_{2,j}^2) - (\| x_3 \|^2 - \|x_2\|^2) \end{bmatrix}. \tag{7}$$

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, $x_j' = [A']^T y_j$. By using one of the equations in (5), we can obtain two possible values for $x_{j,3}$. If the values are complex, the distance data in (5) must be inconsistent. If the two values are equal, atom j must be in the same plane formed by the atoms 1, 2, 3. Otherwise, we obtain two solutions for (5).

The Rigid Geometric Buildup Algorithm

1. Find at least three atoms that are not in the same line.
2. Determine the coordinates of the atoms with the distances among them.
3. Repeat:
   For each of the undetermined atoms,
      If the atom has >3 distances to the determined atoms,
         Determine the atom uniquely.
         Check multiple structures with all these distances.
         Remove structures that violated the distance constraints.
      End
      If the atom has 3 distances to the determined atoms,
         Determine the atom rigidly.
         Record multiple structures generated from reflections.
      End
   End
4. If no atom can be determined in the loop, stop.
5. All atoms are determined.

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 whenever 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.

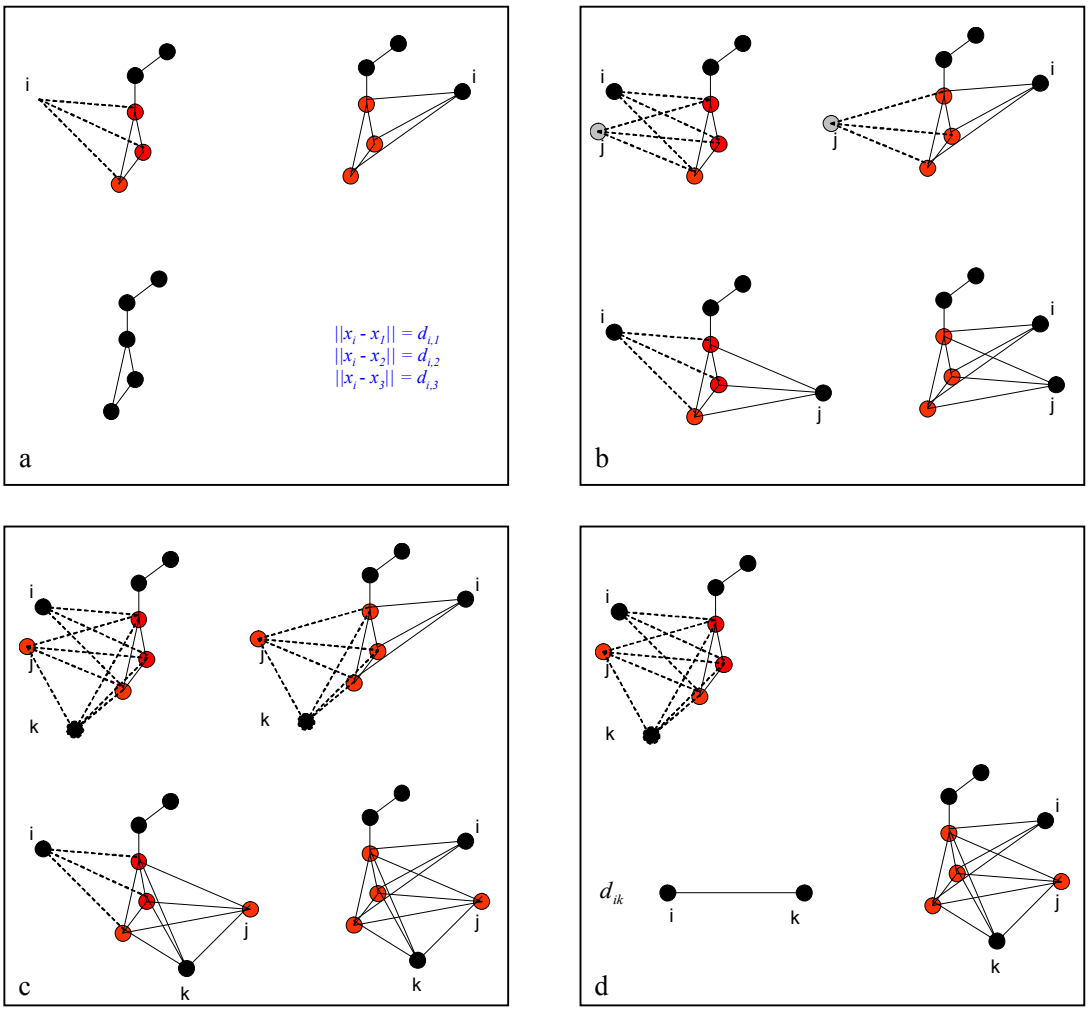


Figure 2 **Rigid structure determination** a. Atom *i* is determined. The number of structures is two. b. Atom *j* is determined. The number of structures is increased to four. c. Atom *k* is determined. d. Two structures are removed because they do not satisfy the distance constraint for atom *i* and *k*.

Figure 2 shows how a structure can be determined rigidly and how multiple structures can be generated and also reduced. Figure 2a 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 2b shows that atom *j* again is determined with three available distances, with two positions for each of the possible structures. In total, four possible structures are made. In Figure 2c, 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 2d.

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.

Definition 3.1 A set of points B in a space S is a reduced metric basis of S provided each point of S is rigidly determined by its distances from the points in B.

Definition 3.2 A set of k points in R^k is called independent if it is not a set of points in R^{k-2}.

Theorem 3.1 Any k independent points form a reduced metric basis for R^k.

Proof It can be proved by generalizing the modified geometric buildup step to the k-dimensional Euclidean space. □

Again, 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 modified 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 3.2 A necessary condition for the rigid determination of the coordinates of a set of atoms, $x_1, \ldots, x_n$, with a given set of distances among the atoms is that each atom must have at least three distances from three other atoms, assuming that this atom is not in the same line with any two of them.

Proof It follows from the fact that the position of an atom can be flexible if it has only two distances from two other atoms unless it is in the same line with the two atoms. □

Theorem 3.3 A sufficient condition for the rigid determination of the coordinates of a set of atoms, $x_1, \ldots, x_n$, with a given set of distances among the atoms is that in every step of the modified buildup algorithm, there is an undetermined atom with three distances from three determined atoms that are not in the same line.

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 atoms rigidly. □

4. Test Results For convenience, we call the general geometric buildup algorithm the unique geometric buildup algorithm, and the modified geometric buildup algorithm the rigid geometric buildup algorithm. We show the test results from applying the two algorithms to determining the structures for a group of proteins at the atomic or residue level, using a subset of inter-atomic or inter-residue distances computed from the known structures of the proteins.

Table 1 contains some results of using the rigid geometric buildup algorithm for the determination of the structures of a group of proteins. They are also compared with the results of using the unique geometric buildup algorithm. The first column contains the names of the proteins

in the PDB Data Bank [17]. The second column contains the numbers of atoms in the proteins. The remaining columns list the results of using the rigid and unique algorithms for the solution of the structures. Two sets of distance data were generated for each protein, one with all distances ≤ 4 Å and another ≤ 5 Å. With the rigid geometric buildup algorithm, some proteins were determined rigidly, but with multiple conformations. For example, two conformations were in fact determined for 1ABA with distances ≤ 5 Å, and one of them was very close to the original structure of the protein. However, there were more than hundreds of conformations for 1ABA and 1BKR with only distances ≤ 4 Å. For the rest of the test cases, the structures were determined uniquely. With the unique geometric buildup algorithm, because the data was too sparse, in most of the test cases, except for two, the structures were not even solvable. The algorithm stopped when it was not able to find the required distances for any of the undetermined atoms [18].

Table 1 **Rigid structure determination at atomic level**

PID	Atoms	Method	4 Å	5 Å
1ABA	699	Rigid	Multiple	Multiple
		Unique	/	/
1BKR	887	Rigid	Multiple	3.80e-07
		Unique	/	/
1EJG	637	Rigid	3.80e-09	9.90e-11
		Unique	/	8.80e-08
1HYP	656	Rigid	3.00e-07	1.80e-07
		Unique	/	2.90e-09

*The RMSD values of the structures compared with the reference structures. Table legend: PID – protein ID; Atoms – the number of atoms; Methods – rigid or unique buildup methods; 4 Å – distances < 4 Å; 5 Å – distances < 5 Å.

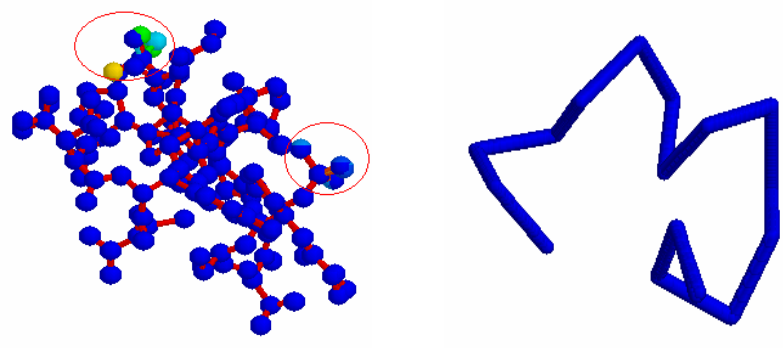


Figure 3 **Rigid structure determination of 1AKG** Shown is the structure of protein 1AKG, with 16 residues, 110 atoms. The distances < 3.5 Å were used. Total 8192 rigid structures were determined. They all were almost identical except for the circled small regions

Figure 3 further demonstrates the application of the rigid geometric buildup algorithm to a small protein, 1AKG, and the nature of the multiple structures it generated. The protein 1AKG is a

small polypeptide with 16 amino acids and 110 atoms. The unique geometric buildup algorithm was able to determine the structure for this protein completely, with distances ≤ 4.5 Å, and the RMSD value of the structure was 8.3e-07 Å against the original structure. Here, the number of distances used was 1638, which was about 14% of all the distances. However, with distances ≤ 3.5 Å, the unique geometric buildup algorithm failed, but the rigid geometric buildup algorithm was still able to find a reasonable number of rigid structures. Here, the number of distances used was 898, which was only 7.5% of all the distances. There were total 8192 multiple conformations found by the rigid algorithm. The one closest to the original structure had the RMSD value equal to 4.3e-07 Å. Note that $8192 = 2^{13}$, and therefore, the multiple structures were 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 happened with the side-chain atoms when they are in the surface of the protein, and the reflections only affected 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 were all uniquely determined.

Table 2 **Rigid structure determination at residual level**

PID	Residues	Method	7.5 Å	8.5 Å
1BKR	108	Rigid	2.20e-12	4.30e-12
		Unique	/	3.60e-11
1EJG	46	Rigid	4.70e-13	1.20e-09
		Unique	/	7.70e-10
1IO0	166	Rigid	3.7e-7 (2)	6.20e-12
		Unique	/	6.60e-12
1LIT	131	Rigid	8.7e-10 (2)	1.40e-11
		Unique	/	9.20e-11
1WRI	93	Rigid	/	5.6e-13 (4)
		Unique	/	/

*The RMSD values of the structures compared with the reference structures. Table legend: PID – protein ID; Residues – the number of residues; Methods – rigid or unique buildup methods; 7.5 Å – distances < 7.5 Å; 8.5 Å – distances < 8.5 Å.

We have also applied the rigid geometric buildup algorithm to determining a group of protein structures at the residual level, with a set of distances between the residue pairs (between the C_α atoms). Table 2 shows the results of using the rigid and unique geometric buildup algorithms for the determination of the structures. The first column contains the PDB names of the proteins. The second column contains the number of residues in each protein. The last two columns show the RMSD values of the structures obtained with rigid or unique buildup methods, against the original structures. Two sets of distance data were tested for each protein, one with distances ≤ 7.5 Å, and another with ≤ 8.5 Å. The table also shows the numbers of multiple conformations for each protein determined using the rigid algorithm. With distances ≤ 7.5 Å, the unique algorithm was not able to determine any of the structures, but the rigid algorithm determined the structures for 1BKR, 1EJG, 1IO0 and 1LIT, with two possible structures for 1IO0

and 1LIT. With distances ≤ 8.5 Å, the rigid algorithm determined the structures for all the listed proteins uniquely except for 1WRI with four possible conformations, but the unique algorithm was able to determine the structures only for 1BKR, 1EJG, 1IO0 and 1LIT [18].

Figure 4 further illustrates the detailed structures for 1IO0 determined by the rigid and unique geometric buildup algorithms. With distances ≤ 8.5 Å, the unique algorithm was able to find the structure, where total 1886 distances, about 7.5% of all distances, were used. On the other hand, with distances ≤ 7 Å, the structure was determined only by the rigid algorithm with 16 possible rigid conformations. The total number of distances used was 1386, about 5% of all distances. Almost all the residues were determined uniquely except for PRO 115, HIS 140 and THR 142 located on the surface of the protein. The PRO 115 had 2 possible positions, HIS 140 had 4 possible positions and THR 142 had 2 possible positions, which contributed to all the 16 conformations determined for the protein.

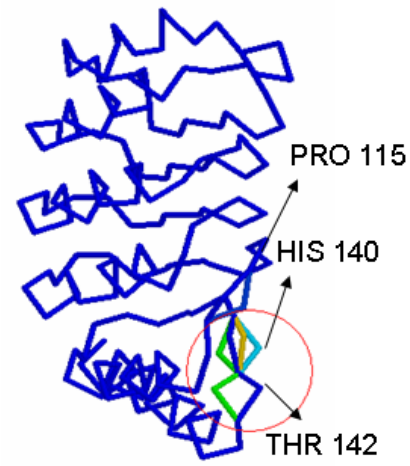


Figure 4 **Rigid structure determination of 1IO0** The structure for 1IO0 was determined rigidly, with 16 possible conformations, using residue distances ≤ 7 Å. Almost all the residues were determined uniquely expect for PRO 115 with 2 positions, HIS 140 with 4 positions, and THR 142 with 2 positions, making total 16 possible conformations.

5. Concluding Remarks In this paper, we have introduced a geometric buildup approach to the distance geometry problem in protein modeling, and discussed the necessary and sufficient conditions on the distances for rigid or unique determination of a protein structure. We have described a new buildup algorithm for determining protein structures rigidly instead of uniquely. The algorithm requires even fewer distance constraints than the general buildup algorithm. We have presented the test results from applying the algorithm to determining the protein structures with varying degrees of availability of the distances, and showed that the new algorithm was able to determine the structures for many of the tested proteins, while the general algorithm failed to do so, given the same limited numbers of distances.

The determination of rigid or unique protein structures with a given set of inter-atomic or inter-residual distances has potential applications in NMR protein modeling [19][20][21] or modeling with residue contact distances [22][23]. However, in these applications, the distances can only be estimated within certain ranges, while the algorithms we have described apply only to exact distances. The algorithms can in principle be extended to distance ranges, with the position

of the unknown atom in each buildup step determined within the given distance ranges by using some optimization method, but they have to be developed with special cares on the characterization of the ensemble of structures defined by the distance ranges and the potential instability of the algorithms due to error accumulation. Investigation along this line is underway and will be reported soon [24].

Note that although only a small subset of all distances is required for the rigid or unique determination of the structures, in practice, the distances may still be lacking in some regions. In that case, the physics-based potentials or other modeling efforts may be helpful for reducing some of the degeneracy of the structures in the uncertain regions, while the geometric buildup algorithm can help to build the initial structures efficiently using the available distances. For physics-based approaches or other modeling methods, the readers are referred to Schlick [25] and Bourne and Weissig [26].

Acknowledgements

The authors would like to thank the anonymous referees for their careful reading the manuscript and providing valuable suggestions. They would also like to acknowledge the support for Di Wu from the New Faculty Startup Fund provided by the Ogden College of Science and Engineering of Western Kentucky University, for Zhijun Wu from the NIH/NIGMS grant R01GM081680, and for Yaxiang Yuan from the NSF of China.

References

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

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

[3]. 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.

[4]. 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.

[5]. D. Wu, R. Jernigan, and Z. Wu, Refinement of NMR-determined protein structures with database derived mean-force potentials, *Proteins: Structure, Function, Bioinformatics*, 2007, (DOI: 10.1002/prot.21358).

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

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

[8]. T. F. Havel, Distance geometry: Theory, algorithms, and chemical applications, in *Encyclopedia of Computational Chemistry*, John Wiley & Sons, 1998, 1-20.

[9]. 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.

[10]. 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.

[11]. 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.

[12]. 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.

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

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

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

[16]. 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.

[17]. 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.

[18]. D. Wu, *Distance-Based Protein Structure Modeling*, Ph.D. Thesis, Program on Bioinformatics and Computational Biology and Department of Mathematics, Iowa State University, 2006.

[19]. J. L. Klepeis, C. A. Floudas, D. Morikis, and J. D. Lambris, Predicting peptide structures using NMR data and deterministic global optimization, *J. Comp. Chem*. **20**, 1999, 1354-1370.

[20]. J. L. Klepeis and C. A. Floudas., Prediction of beta-sheet topology and disulfide bridges in polypeptides, *J. Comp. Chem*. **24**, 2002, 191-208.

[21]. C. A. Floudas, H. K. Fung, S. R. McAllister, M. Mönnigmann, and R. Rajgaria, Advances in protein structure prediction and de novo protein design: a review, *Chem. Eng. Sci.* **61**, 2006, 966-988.

[22]. S. Vicatos, B. V. Reddy, and Y. Kaznessis, Prediction of distant residue contacts with the use of evolutionary information, *Proteins* **58**, 2005, 935-949.

[23]. J. Cheng and P. Baldi, Improved residue contact prediction using support vector machines and a large feature set, *BMC Bioinformatics* **8:113,** 2007.

[24]. A Sit, Z. Wu, and Y. Yuan, A geometric buildup algorithm for the solution of the distance geometry problems using least-squares approximation, in preparation, 2007.

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

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