# A Stochastic/Perturbation Global Optimization Algorithm for Distance Geometry Problems*

Zhihong Zou † Richard H. Byrd ‡ Robert B. Schnabel §

July, 1996

**Abstract**

We present a new global optimization approach for solving exactly or inexactly constrained distance geometry problems. Distance geometry problems are concerned with determining spatial structures from measurements of internal distances. They arise in the structural interpretation of nuclear magnetic resonance data and in the prediction of protein structure. These problems can be naturally formulated as global optimization problems which generally are large and difficult. The global optimization method that we present is related to our previous stochastic/perturbation global optimization methods for finding minimum energy configurations, but has several key differences that are important to its success. Our computational results show that the method readily solves a set of artificial problems introduced by Moré and Wu that have up to 343 atoms. On a set of considerably more difficult protein fragment problems introduced by Hendrickson, the method solves all the problems with up to 377 atoms exactly, and finds nearly exact solution for all the remaining problems which have up to 777 atoms. These preliminary results indicate that this approach has very good promise for helping to solve distance geometry problems.

*Research supported by AFOSR grant F49620-94-1-0101, ARO contract DAAH04-94-G-0228, and NSF grant CDA-9502956. Address for all authors: Department of Computer Science, University of Colorado, Boulder, Colorado 80309-0430.

†zzou@cs.colorado.edu

‡richard@cs.colorado.edu

§bobby@cs.colorado.edu

# 1 Introduction

Distance geometry problems, which are concerned with determining spatial structures from measurements of internal distances, arise in the structural interpretation of nuclear magnetic resonance data and in the prediction of protein structure. Methods of calculating the conformations of biological molecules from distance constraints have become an important tool in structural biochemistry. Solving the distance geometry problem in this context would determine the three-dimension shape of the protein, which is critical for understanding its chemical and biological properties. For general reviews of the distance geometry problem and its relationship to the structure of chemical molecules, see Crippen and Havel[6], Havel[8], Kuntz, Thomason and Oshiro[10],and Brünger and Nilges[1].

The distance geometry problem can be naturally formulated as a nonlinear global optimization, where the objective function is constructed such that the distance constraints are satisfied at the global solutions of the problem. A simple objective function can be defined to enforce the constraints. This optimization problem is believed to be computationally intractable in general because it has been shown to be strongly NP-complete in the one dimensional case [6, 13], and strongly NP-hard in the higher dimension case [14].

A large number of such methods for solving distance geometry problems have been proposed, such as Crippen and Havel[6], Havel[8], Hendrickson[9], Glunt, Hayden, Raydan[7], and Moré and Wu[12, 13]. The method we present in this paper is based on the stochastic/perturbation global optimization approach[4, 2, 3, 5] with several new features. The purpose of this paper is to describe this approach and to demonstrate its capabilities on some difficult distance geometry problems. The larger context of this research is to continue to develop and understand the capabilities of the stochastic/perturbation global optimization methodology, which we have found very successful for several large-scale global optimization problems arising from molecular chemistry.

Our stochastic/perturbation algorithm combines a first, stochastic phase that identifies an initial set of local minimizers, with a second, more deterministic phase that moves from low to even lower local minimizers. In the second phase, by incorporating the partially separable structure of the problem, we work on very small-dimensional global optimization subproblems and then for local minimizers in the full-dimensional space. Both the selection of small dimensional subproblems and some other important algorithmic features are specific to the distance geometry problem.

We experimented with our algorithm on Moré and Wu's artificial problems [12] and on the protein fragment problems from Hendrickson[9]. For the artificial problems, even our first phase can find the exact solutions with great success. For the protein fragment problems, which are considerably more difficult, we have found the exact solutions solutions for problems with up to 377 atoms (1131 parameters).

Another important issue in dealing with distance constraints is that there may not exist any molecular structure satisfying the given distance constraints, due to measurement or experimental errors. In practice, lower and upper bounds on distances are specified instead of exact distances. Therefore our algorithm is intended to deal with both exact and inexact distance geometry problems. For the larger protein problems of Hendrickson[9], with up to 777 atoms (2331 parameters), our algorithm has found approximate solutions with maximum relative error of distance at most 0.04. These computational successes on problems of considerable size appear to indicate that our algorithm is a powerful tool for solving distance geometry problems.

The remainder of this paper is structured as follows. Section 2 describes distance geometry problems and current approaches developed for these problems. In section 3, we describe the stochastic/perturbation algorithm used to deal with the distance geometry problems. The framework of our algorithm is outlined, and several new features are discussed in the section. These are followed by extensive experimental results for our method on distance geometry problems in Section 4. Section 5 contains some brief conclusions and comments about future research.

## 2 The Distance Geometry Problems

As mentioned earlier, solving the distance geometry problems is an important tool in determining the three-dimensional structure of a molecule. In the ideal case, distance geometry problems are concerned with finding positions $x_1, x_2, \cdots, x_m$ in $\mathcal{R}^3$ such that

$$\| x_i - x_j \| = \delta_{i,j}, (i,j) \in \mathcal{S}. \tag{2.1}$$

where $\mathcal{S}$ is a subset of the atom pairs, and $\delta_{i,j}, (i,j) \in \mathcal{S}$ is the given distance between atom i and j. Usually $\mathcal{S}$ has many fewer than $m^2/2$ elements, that is only a small subset of pairwise distances is known.

There may not exist any solution $x_1, x_2, \cdots, x_m$ to these distance constraints, due to the error in the theoretical or experimental data. For example, this is guaranteed to happen if data for atoms $i, j, k$ violate the triangle inequality.

In the more general distance geometry problem, lower and upper bounds on the distances are specified instead of exact values. In this case, the distance geometry problem is to find a set of positions $x_1, x_2, \cdots, x_m$ satisfying

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

where $l_{i,j}$ and $u_{i,j}$ are lower and upper bounds on the distance constraints, respectively.

The distance geometry problems with constraints (2.1) and (2.2) can be naturally phrased as a nonlinear global optimization problem. The objective function is constructed so that the constraints are satisfied at the global minimizers of the optimization problem. One simple approach, which we utilize in this paper, is to penalize all the unsatisfied constraints. We formulate the distance geometry problem in terms of finding the global minimizers of the function

$$f(x) = \sum_{(i,j)\in\mathcal{S}} h_{i,j}(x_i - x_j), \tag{2.3}$$

where

$$h_{i,j}(x) = min^2\{\frac{\| x_i - x_j \|^2 - l_{i,j}^2}{l_{i,j}^2}, 0\} + max^2\{\frac{\| x_i - x_j \|^2 - u_{i,j}^2}{u_{i,j}^2}, 0\}. \tag{2.4}$$

This is the approach taken by Moré and Wu[13]. In order to compare results with those in Moré and Wu[12], we also use the following function for the exact problem:

$$f(x) = \sum_{(i,j)\in\mathcal{S}} w_{i,j}(\| x_i - x_j \|^2 - \delta_{i,j}^2)^2 \tag{2.5}$$

where $w_{i,j}$ are positive weights. This function is same as (2.4) for the exact distance geometry problems if $l_{i,j} = u_{i,j} = \delta_{i,j}$ and $w_{i,j} = \delta_{i,j}^{-4}$. Clearly, $x = \{x_1, x_2, \cdots, x_m\}$ solves problem (2.2) if and only if $x$ is the global minimizer of $f(x)$ in (2.4) and $f(x) = 0$. Similarly for problem (2.1), a solution $x$ is equivalent to a global minimizer of the function (2.5) with function value 0.

The difficulty in solving the distance geometry problem arises from several sources. First, the problem in itself is strongly NP-complete in one dimension, and strongly NP-hard in higher dimension[6, 14, 13], therefore it is very unlikely that an efficient algorithm for solving all cases of the problem can be found. Also there are huge numbers of local minimizers for the functions (2.4) and (2.5), which makes it very challenging to locate the basin of attraction of the global minimizer. It can be shown that function (2.5) has an exponential number of local minimizers[9]. The another important aspect is that most distance geometry problems are large and contain hundreds or more atoms. Thus even the calculation of a local minimizer makes heavy demands on the computer time.

Most of the previous work by chemists on the distance geometry problems depends on the intrinsic properties of the chemical structure of solutions. Often, these methods use special heuristics based on chemical properties to generate initial configurations, and then perform local optimizations. The method of Hendrickson[9] is of particular interest to this paper because the more difficult test problems that we use are taken from that paper. Hendrickson's method also utilizes chemical structure information. It works well on his test problems, which are generated from bovine pancreatic ribonuclease, a protein containing 124 amino acids, but relies on the assumption that the distances are highly accurate.

In contrast, general global optimization approaches to the distance geometry problem do not utilize information about the solution structure. While utilizing problem-specific information may ultimately be desirable, it is important to understand the capabilities of general global optimization approaches for distance geometry problems. Work of this type includes the multistart method and Moré and Wu's continuation approach [12, 13]. In the former, a set of starting points are randomly generated and local minimization is performed from them. This method is simple and easy to implement. But it is highly unreliable in the context of distance geometry because the problems are usually large and have a huge number of local minimizers. In the continuation approach, the original function is transformed into a smoother function with fewer local minimizers. A series of local optimization algorithms is then applied to the transformed function, tracing its minimizers back to those of the original function. The experimental results of Moré and Wu show that this approach is markedly superior to the multistart method.

In the following section, we will describe our stochastic/perturbation algorithm for the distance geometry problem. Our algorithm is a general global optimization method, and applies to both exact and inexact problems. We do not utilize a smoothing approach as in the work of Moré and Wu, but our approach could readily be combined with smoothing techniques. Indeed, our experience with minimum energy conformation problems is showing that the combination of smoothing and our stochastic/perturbation global optimization approach is a promising framework (see e.g. [15]).

# 3 The Stochastic/Perturbation Global Optimization Method

Our new algorithm for the distance geometry problems is based on the stochastic/perturbation global optimization method, which was constructed to solve large-scale global optimization problems. This approach has been successfully applied to Lennard-Jones problems [4], water cluster problems [2], and protein conformation problems [3].

The basic framework of our global optimization method for distance geometry problems is outlined in Algorithm 3.1 below. The method combines an initial phase that locates some low local minimizers with a second phase for moving from low to even lower local minimizers. During the first phase, a full dimensional random sample is generated over the domain space by randomly and independently placing each atom. The worst configurations are discarded, and the better ones are improved by selecting and moving an atom or a pair of atoms, until the function value for the configuration falls below a specified threshold level. A subset of these improved configurations is used as start points for a full dimensional local optimization algorithm. Some of the local minimizers found in this phase are passed on to the second phase for improvement.

In the second phase, a local minimizer is successively selected for improvement as discussed below. A pair of atoms is chosen, and a small-scale stochastic global optimization algorithm is applied to the configuration with only these two atoms as variables and the remainder of the configuration fixed. This is followed by full dimensional local minimizations starting from the best configurations that resulted from the small-scale global optimization step. The lowest new configurations are then merged into the list of local minimizers, and this phase is iterated a fixed number of times.

A key feature of this approach is that both phases make use of strategies that work on a very small subset of the atoms at once. In the initial phase this approach is used to improve the sample points by sampling on only one atom or a pair of atoms at a time in step 1b while leaving the remaining atoms temporarily fixed. When a pair of atoms is chosen, it is a pair for

**Algorithm 3.1 – Framework of the Large-Scale Global Optimization Algorithm for Distance Geometry Problems**

1. **Initial Generation of Configurations Phase :**
   (a) **Sampling in Full Domain :** Randomly generate the coordinates of the sample points in the sampling domain, and evaluate $f(x)$ at each new sample point. Discard all sample points whose function value is below a global "cutoff level".
   (b) **One-atom/two-atom Sampling Improvement :** For each remaining sample point : While the energy of the sample point is above the threshold value, Repeat:
      - Select the atom that contributes most to the function value or a pair of atoms in $\mathcal{S}$ that violates the ideal distance most
      - Randomly sample on the location of the selected atoms
      - Replace the atoms in the sample point with the new sample coordinates that give the lowest energy value.
   (c) **Start Point Selection :** Select a subset of the improved sample points from step 1b to be start points for local minimizations.
   (d) **Full-Dimensional Local Minimizations:** Perform a local minimization from each start point selected in step 1c. Collect some number of the best of these minimizers for improvement in Phase 2.
2. **Improvement of Local Minimizers Phase:** For some number of iterations:
   (a) **Select a Configuration :** From the list of full-dimensional local minimizers, select the local minimizer and a pair of atoms in $\mathcal{S}$ to be optimized.
   (b) **Expansion :** Transform the configuration by multiplying the position of each atom relative to the center of mass of the configuration by a constant factor of between 1.25 and 2.0 or larger.
   (c) **Two-atom Small Global Optimization :** Apply a global optimization algorithm to the expanded configuration with only the two atoms chosen in step 2a as variables.
   (d) **Full-Dimensional Local Minimization :** Apply a local minimization procedure, using all the atoms as variables, to the lowest configurations that resulted from the two-atom global optimization.
   (e) **Merge the New Local Minimizers :** Merge the new lowest configurations into the existing list of local minimizers.

which a constraint distance is given in the problem formulation, that is an $(x_i, x_j)$ for which $(i, j) \in \mathcal{S}$. We will refer to this as a "constrained pair of atoms". When a constrained pair of atoms is used in step 1b, the sampling is done so that the distance between the atoms is always at the constraint value $\delta_{i,j}$. In the second phase, this small subproblem approach is used to move a constrained pair of atoms in an existing configuration to new positions via the small-scale global optimization in step 2c. This small global optimization locates the best possible position for the selected atoms in the current configuration with the remaining atoms temporarily fixed. Again, when step 2c samples on the constrained pair of atoms, the distance between them is kept at the constraint value. All these small subproblem steps are very efficient due to the partial separability of the objective functions for the distance geometry problem, which make the cost of evaluating the function when only one or two atoms are moved much less expensive than the cost of a full function evaluation.

The atom or pair of atoms that is chosen in each of these small-scale steps is one for which there seems to be potential for improving the overall configuration by moving this atom or pair. In phase 1, the atom which contributes most to the overall function value (and thus appears to have the greatest potential for reducing the function value) is selected. In phase 2 and alternatively in phase 1, the constrained pair of atoms that has the worst violation from the ideal distance is selected. In our early experiments, we also employed a one-atom small-scale global optimization heuristic in the second phase. The experimental results suggested that for some difficult problems the two-atom approach is more effective and that it locates lower configurations than the one-atom strategy, while for many other problems the two strategies are about equally effective. For this reason we have used the two-atom approach in phase 2. It may be that in some cases, moving only one atom is not enough to force the configuration out of the region of attraction of the current local minimizer.

The choice of which local minimizer to improve in phase 2 is an important heuristic in this method. Through our experimentation on previous problems, we have concluded that a strategy that balances selecting a breadth of configurations with working on the best current configurations is most effective. This is the approach we have used for the distance geometry problem.

Another important feature in Algorithm 3.1 is the expansion step, step b of phase 2, which expands the configuration around its center of mass prior to the small-scale global optimization. The expansion step was used first with minimum energy water cluster problems, where it significantly improved the ability of the phase 2 iteration to find improved configurations. This improvement apparently is due to expansion creating more room to move the atom or atoms that are the variables in the small-scale global optimization. The local optimization step then contracts the configuration again, into a new and hopefully improved local minimizer. The physical analogy of expansion is to heating in annealing. The drawback of expansion is that the local minimizations in phase 2 become more expensive.

Our initial intuition was that the expansion would not be appropriate for distance geometry problems, because it leads to large violations in all the distance constraints. But we found that the algorithm without the expansion step didn't perform well in phase 2; after step 2d, most of the configurations returned to the local minimizer that the step had started with in step 2a. Possibly this is due to the relatively small number of constraints in distance geometry problems, which means that each atom is only involved in a small number of distance constraints. Therefore moving only the pair of atoms doesn't change the configuration sufficiently from the current local minimum. We found that when the expansion step is added, the small-scale global optimization step is far more successful at producing significantly different, and improved, configurations. We have experimented with different expansion factors in the range 1.25 to 2.5. For most of the problems, an expansion factor between 1.25 and 2.0 is most effective, but for some problems, an expansion factor between 2.0 and 2.5 is more helpful. In order to be consistent with the stochastic aspect of our algorithm, for most of the test problems in next section, the expansion factor is chosen randomly between 1.25 and 2.0.

A final useful feature in our implementation, which is not captured in Algorithm 3.1, is a progression of problems approach. By this we mean that for some more difficult problems, we solve a sequence of problems of the form (2.4), with increasingly tight constraints (i.e. lower and upper bounds on the distances). Phase 1 is applied only to the first problem (with the loosest constraints), and phase 2 is applied to each problem, with the best configurations from the previous, looser-constrained problem forming the starting set of configurations for phase 2 for the next, tighter-constrained problem. The motivation for this approach came from experimentation. For the easier problems, applying Algorithm 3.1 to the problem directly was very successful. For the more difficult problems, we noticed that finding the global minimizer of function (2.4) directly was much more difficult than starting to work on the distance geometry problem with looser distance constraints, and then gradually tighting the constraints and applied Phase 2 until the expected accuracy. This approach is in some sense similar to the continuation methods. It has the additional advantage that for real problems, we may not know exactly how accurate the given distances are, and with this approach we can find the solutions to the best accuracy that is possible. In fact the chemists care more about finding reasonable solutions than exact solutions, and this approach coincides with this idea and yields an algorithm that can be tailored to find reasonable configurations.

# 4 Experimental Results

We have run our algorithm on a set of artificial distance geometry problems from Moré and Wu [12, 13], and on a set of protein fragment test problems generated by Hendrickson[9]. In all of these problems, we are given the exact distances. In order to have both exactly and inexactly constrained problems for function (2.4), we used

$$l_{i,j}^2 = \delta_{i,j}^2(1-\epsilon), \qquad u_{i,j}^2 = \delta_{i,j}^2(1+\epsilon)$$

with values of *epsilon* that are given later. Most of our test problems are large, with hundreds or thousands of atoms. Therefore in our implementation, we used the limited memory BFGS algorithm[11] to perform the full dimensional local minimizations, whereas the small-scale local minimizations within the small-scale global optimization step used the standard BFGS method. All the experiments were conducted on an Intel Paragon multiprocessor.

In the artificial distance geometry problems from [12, 13], the molecule has $m = s^3$ atoms located in the three-dimensional lattice

$$\{(i_1, i_2, i_3) : 0 \leq i_1 < s, 0 \leq i_2 < s, 0 \leq i_3 < s\}$$

for some integer $s \geq 1$. The problem is to determine the structure of this molecule if we are given $\delta_{i,j} \in \mathcal{S}$, where

$$\mathcal{S} = \{(i,j) : |i-j| \leq r\}$$

and r is an integer between 1 and m.

Our first experiment for the artificial problem is to find the global minimizers using the exact function (2.5), using $w_{i,j} = 1$ for all $(i,j) \in \mathcal{S}$. First

| *Problem* | | *Domain $\mathcal{B}$* | | | *Domain $2\mathcal{B}$* | | |
|---|---|---|---|---|---|---|---|
| m | r | #fval | #gval | #global | #fval | #gval | #global |
| 27 | 9 | 2305 | 2005 | 5/15 | 2098 | 2089 | 7/15 |
| 64 | 16 | 3829 | 3445 | 1/15 | 3814 | 3441 | 1/15 |
| 125 | 25 | 5925 | 5520 | 1/15 | 5791 | 5404 | 3/15 |
| 216 | 36 | 7795 | 7336 | 0/15 | 8003 | 7483 | 0/15 |
| 343 | 49 | 10726 | 10160 | 0/15 | 10660 | 10175 | 0/15 |

Table 4.1: Phase 1 results for Artificial Problems and Function (2.5)

we ran Phase 1 for the molecules with $(m, r) = (s^3, s^2)$ where 3$\leq$ s$\leq$ 7, starting from the domain space

$$\mathcal{B} = \{x \in \mathcal{R}^{3m} : 0 \leq (x_i)_k \leq s - 1,\ i = 1, \cdots, m,\ k = 1, 2, 3\}$$

and the domain $2\mathcal{B}$. (These are the same problems reported in [12, 13].) One-atom moves were used in step 1b. The numerical results are presented in Table 4.1, where # fval, #gval and #global are the number of function and gradient evaluations, the global solutions out of the total minimizers, respectively. Although Phase 1 can't find the global solution in either the smaller domain space or the larger one for the last two problems, recall that Phase 1 is simply designed to provide initial configurations for the second phase, which accounts for most of the work in the algorithm. In our experience, problems that can already be solved in Phase 1 are not particularly difficult global optimization problems. Note also that Phase 1 is in some sense similar to the multistart algorithm except for the one-atom or two-atom steps. Compared to the multistart results in [12], the phase 1 works quite well in this context.

We also ran the same artificial problems using the exact version of function (2.4), i.e. with $\epsilon = 0$. The Phase 1 results for these runs are given in Table 4.2. Again, one-atom moves were used in step 1b. Table 4.2 shows that using function (2.4), Phase 1 already finds the global solutions for each problem in both cases. This indicates that these are not particularly difficult global optimization problems. The comparison of these to those in Table 4.1 seems to indicate that using the exact version of function (2.4) creates an easier problem than using function (2.5). We believe that this is because function (2.4) is smoother. Finally, there is a marked difference between

| *Problem* | | *Domain $\mathcal{B}$* | | | *Domain $2\mathcal{B}$* | | |
|---|---|---|---|---|---|---|---|
| m | r | #fval | #gval | #global | #fval | #gval | #global |
| 27 | 9 | 1822 | 1443 | 3/15 | 1699 | 1398 | 5/15 |
| 64 | 16 | 2622 | 2253 | 7/15 | 2493 | 2091 | 4/15 |
| 125 | 25 | 3268 | 2829 | 4/15 | 3231 | 2764 | 3/15 |
| 216 | 36 | 4323 | 3840 | 5/15 | 3779 | 3340 | 4/15 |
| 343 | 49 | 5206 | 4646 | 3/15 | 4856 | 4338 | 4/15 |

Table 4.2: Phase 1 results for Artificial Problems and Function (2.4)

the results in Table 4.2 and the multistart results given for these problems and function (2.4) in [13], in which the global minimizers were never found. However these results are not directly comparable because the results in [13] are for (2.4) with $\epsilon = 0.1$.

Tables 4.3 and Table 4.4 give the results from applying Phase 2 to the same set of artificial problems, again using functions (2.5), and (2.4) with $\epsilon = 0$, respectively. In these tables, FLS, SLS are the total number of full dimensional and small dimensional local optimizations, respectively, and Ffval and Pfval are the number of full function evaluations and partial (only two atoms change) function evaluations, respectively. The results show that for function (2.5), applying Phase 2 allows the method to readily solve the two problems not solved in Phase 1 and to relocate the global minimizers for all the problems many times. For function (2.4), more than half the local minimizations find the global minimizer. These results show that Algorithm 3.1 is very successful on these artificial problems. A comparison with the efficiency of the method of Moré and Wu [12, 13] is difficult in part because they do not give costs in the second of these papers, but mainly because smoothing is a very useful technique that makes problems easier to solve, but has not been used in our algorithm.

We then proceeded to a set of significantly more difficult test problems. These problems are generated by Hendrickson [9] from the bovine pancreatic ribonuclease, a typical, rather small protein containing 124 amino acids. Hendrickson derived the set of twelve test problems by using fragments consisting of the first 20, 40, 60, 80, and 100 amino acids as well as the full protein, with two sets of distance constraints for each size. The distance information given for these problems is exact, meaning that there is a global

| Problem | FLS | #Ffval | SLS | #Pfval | #global |
|---|---|---|---|---|---|
| 27 | 50 | 2905 | 6779 | 239165 | 23/42 |
| 64 | 50 | 7786 | 11616 | 1077940 | 18/39 |
| 125 | 50 | 11570 | 17305 | 1627957 | 6/32 |
| 216 | 50 | 14267 | 24810 | 1834247 | 13/44 |
| 343 | 50 | 33624 | 15168 | 1827104 | 9/40 |

Table 4.3: Phase 2 results for Artificial Problems and Function (2.5)

| Problem | FLS | #Ffval | SLS | #Pfval | #global |
|---|---|---|---|---|---|
| 27 | 89 | 7292 | 4065 | 248811 | 37/70 |
| 64 | 100 | 10278 | 9333 | 606992 | 58/82 |
| 125 | 100 | 13371 | 18817 | 1338752 | 49/78 |
| 216 | 100 | 15977 | 29439 | 2725847 | 54/91 |
| 343 | 100 | 17853 | 39068 | 4097761 | 47/81 |

Table 4.4: Phase 2 results for Artificial Problems and Function (2.4)

minimizer where all constraints are satisfied exactly. The problems have from 63 to 777 atoms (189 to 2331 parameters). Moré and Wu [13] tested their method on the smallest of these problems, using function (2.4), and reported success for the values $\epsilon = 0.10, 0.06, 0.04, 0.02$, but not below.

For each of the twelve problems of Hendrickson [9], we applied the Phase 2 repeatedly to the function (2.4) with decreasing $\epsilon$ until $\epsilon = 0$. In our runs, we usually chose the values $\epsilon = 0.10, 0.06, 0.04, 0.02, 0.01, 0.00$ used by Moré and Wu [13]. For the two smallest problems, however, we started from 0.02, whereas for the largest problems, we may find the global solutions for the $\epsilon = 0.15$ subproblem and then proceed to the $\epsilon = 0.10$ subproblems. For each value of $\epsilon$, we generally performed 50 iterations in Phase 2. We applied Phase 1 on the first subproblem only, using the constrained pair of atom moves in step 1b.

Our numerical results on the Hendrickson problems are summarized in the Table 4.5. The first two columns record the number of atoms and the given distance constraints, the third column is last $\epsilon$-problem which we can solve successfully in the progression of problems approach, and the final column is the maximum relative error of the distances with respect to the ideal distances in our best solution to this subproblem, where the maximum relative error is defined as

$$max\{|(\| x_i - x_j \|^2 / \delta_{i,j}^2) - 1.0|\} \qquad (i,j) \in \mathcal{S}.$$

Table 4.5 shows that for the first seven problems, which have 63 to 377 atoms (189 to 1131 parameters) we can find the global solutions to the problems. For the remaining five problems, which have 472 to 777 atoms (1416 to 2331 parameters), we solve the problems down to relative accuracy levels of between 0.01 and 0.04. These results indicate that our approach has the potential to locate exact or nearly exact solutions of quite large distance geometry problems without using any particular structural information.

Finally, we briefly discuss the costs of Algorithm 3.1 on the protein fragment problems. We have analyzed the costs of two typical runs for the 63-atom and 102-atom problems. For the 63-atom problem, the entire solution process required 692,488 full function evaluations and 2,416,214 partial function evaluations. For the 102-atom problem, the entire solution process required 498,500 full function evaluations and 952,836 partial function evaluations. Detailed timings show that the full dimensional local optimization steps in Phase 1 and 2 dominate the cost of the algorithm for these problems, accounting for 80% of the total time. The other steps are relatively inexpensive. Therefore a cheaper local optimization algorithm will

| Number of Atoms | Distance Constraints | Final Subproblem | Final Relative Error |
|---|---|---|---|
| 63 | 236 | 0.00 | 1.0e-6 |
| 102 | 336 | 0.00 | 1.0e-6 |
| 174 | 786 | 0.00 | 8.1e-4 |
| 236 | 957 | 0.00 | 8.1e-4 |
| 287 | 1319 | 0.00 | 8.1e-4 |
| 362 | 1526 | 0.00 | 8.1e-4 |
| 377 | 1719 | 0.00 | 3.7e-3 |
| 472 | 2006 | 0.02 | 0.02 |
| 480 | 2169 | 0.01 | 0.01 |
| 599 | 2532 | 0.01 | 0.01 |
| 695 | 3283 | 0.04 | 0.04 |
| 777 | 3504 | 0.04 | 0.04 |

Table 4.5: Results on Protein Fragment Problems

reduce the total cost of our algorithm. An alternative way to reduce the cost would be to work in internal coordinates. The advantage of this approach is that the number of variables is greatly reduced, generally by a factor of ten or more. Although the problem is no longer partially separable in this parameterization, the overhead in re-evaluation of the partial function and gradient values is relatively minor compared to the reduction in the local optimization cost. A disadvantage of this approach is that it already uses some structural information of the problem, namely the primary structure of the protein, which may be considered contrary to the goal of assessing the potential of general-purpose global optimization methods for distance geometry problems.

# 5 Summary and Future Research

We have presented a new global optimization algorithm intended to solve exactly or inexactly constrained distance geometry problems. The algorithm utilizes small-scale global optimization calculations on selected subsets of the parameters, performed by a stochastic global optimization method, as a key part of its approach. Its structure is related to our previous stochastic/perturbation global optimization methods for molecular clusters and protein folding, but there are several important differences. In particular, three algorithmic choices helped in the success of the method and none were anticipated by us initially. The first is to expand the configuration before each small-scale global optimization. The second is to select a linked pair of atoms rather than a single atom as the parameters for the small-scale global optimization. The third is to solve a progression of distance geometry problems where the constraint bounds become tighter and tighter.

Our computational tests on artificial problems with up to 343 atoms (1029 variables) and on more difficult protein fragment problems with up to 777 atoms (2331 variables) indicate that the method is very successful in located the configurations satisfying the given distance constraints. The results on the protein problems especially indicate that the method is quite successful in solving large and apparently difficult distance geometry problems without using any information about the solution structure.

There are many directions for continuing this research that we are considering. One is to use more than two atoms in the small-scale global optimization step of phase 2. While this increases the cost of the small-scale global optimization, this is not a dominant cost, and with such a strategy we were able to solve the 480 atom problem at the subproblem 0.00 level. A second possible direction is to use a buildup approach, where one finds (approximate) solutions to subsets of the problem on the way to solving the full problem. The third, and perhaps most intriguing direction, is to combine a smoothing approach like that of Moré and Wu [12, 13] with our stochastic/perturbation global optimization technique. Our recent research in protein folding has shown that this combination has great potential. Finally, if this stochastic/perturbation approach were to be part of a production distance geometry code, one would want to find ways to combine it with the chemical structural knowledge that chemists use in the solution of these problems.

# Acknowledgement

We thank Zhijun Wu and Bruce Hendrickson for providing the test problems, and also thank Zhijun Wu for many helpful communications about this research.

## References


[1] A.T.Brünger and M.Nilgers, *Computational Challenges for macromolecular structure determination by X-ray crystallography and solution NMR-spectroscopy*, Q. Rev. Biophys., 26(1993), pp.49-125.

[2] R. H. Byrd, T. Derby, E. Eskow, K. Oldenkamp and R. B. Schnabel, *A New Stochastic/Perturbation Method for Large-Scale Global Optimization and its Application to Water Cluster Problems*, in Large-Scale Optimization: State of the Art, W. Hager, D. Hearn, and P. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994, pp. 71–84.

[3] R. H. Byrd, E. Eskow, A. van der Hoek, R. B. Schnabel, K. P. B. Oldenkamp, *A Parallel Global Optimization Method for Solving Molecular Cluster and Polymer Conformation Problems*, Proceedings of the Seventh Siam Conference on Parallel Processing for Scientific Computing, D. H. Bailey, P. E. Bjørstad, J. R. Gilbert, M. V. Mascagni, R. S. Shreiber, H. D. Simon, V. J. Torczon, L. T. Watson, eds., SIAM, Philadelphia, 1995, pp. 72–77.

[4] R. H. Byrd, E. Eskow and R. B. Schnabel, *A New Large-Scale Global Optimization Method and its Application to Lennard-Jones Problems*, University of Colorado Technical Report CU-CS-630-92.

[5] R. H. Byrd, E. Eskow, A. van der Hoek, R. B. Schnabel, C.S. Shao, Z. Zou, *Global Optimization Methods for Protein Folding Problems*, DIMACS Vol. 23, Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding, DIMACS Workshop, Mar. 20-21, 1995, edited by P. Pardalos, D. Shalloway, and G. Xue.

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

[7] W. Glunt, T.L.Hayden and M. Raydan, *Molecular Conformation from Distance Matrices*, J.Comp. Chem., 14(1993), pp. 114-120.

[8] T.F.Havel, *An Evaluation of Computational Strategies for Use in the Determination of Protein structure from Distance Geometry Constraints Obtained by Nuclear Magnetic Resonance*, Prog. Boiphys. Mol. Biol., 56(1991), pp. 43-78.

[9] B.A.Hendrickson, *The Molecule Problem: Exploiting Structure in Global Optimization*, SIAM J. Optimization, Vol. 5(1995), No.4, pp. 835-857.

[10] I.D.Kuntz, J.F.Thomason and C.M.Oshiro, *Distance Geometry*, in *Methods in Enzymology*, N.J.Oppenheimer and T.L. James, eds., Vol. 177, Academic Press, 1993, pp. 159-204.

[11] D.C. Liu and J.Nocedal, *On the Limited Memory BFGS Method for Large Scale Optimization*, Math. Prog., 45(1989), pp. 503-528.

[12] Jorge J. Moré and Zhijun Wu, *Global Continuation for Distance Geometry Problems*,Preprint MCS-P505-0395, Argonne National Laboratory, Argonne, Illinois, 1995.

[13] Jorge J. Moré and Zhijun Wu, *$\varepsilon$-Optimial solutions to Distance Geometry Problems via Global Continuation*, DIMACS Vol. 23, Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding, DIMACS Workshop, Mar. 20-21, 1995, edited by P. Pardalos, D. Shalloway, and G. Xue.

[14] J.B.Saxe, *Embeddability of Weighted Graphs in k-space is Strongly NP-hard*, Proc. 17th Allerton Conference in Communications, Control and Computing, 1979, pp. 480-489.

[15] C.-S.Shao, R.H. Byrd, E.Eskow and R.B.Schnabel, *Global Optimization for Molecular Clusters Using a New smoothing approach*, (1995).