

# Combining Constraint Programming and Multidimensional Scaling to solve Distance Geometry Problems.

Ludwig Krippahl<sup>1</sup>, Michael Trosset<sup>2</sup> and Pedro Barahona<sup>1</sup>

[ludi@dq.fct.unl.pt](mailto:ludi@dq.fct.unl.pt)

[trosset@math.wm.edu](mailto:trosset@math.wm.edu)

[pb@di.fct.unl.pt](mailto:pb@di.fct.unl.pt)

1-Dep. de Informática, Universidade Nova de Lisboa, 2825 Monte de Caparica, Portugal

2-Dept. Mathematics, College of William & Mary, P.O.Box 8795, Williamsburg, Virginia 23187-8795 USA

## Abstract

In this paper we address the merits of combining constraint propagation, typical of Constraint Programming (CP), with local search to solve problems where each method alone can perform poorly. This combined approach is used in solving distance geometry problems, namely those arisen in the determination of spatial configuration of molecular structures from Nuclear Magnetic Resonance (NMR) data. In this problem, we apply a special purpose constraint propagation with local search optimisation based on a Multidimensional Scaling (MDS) method.

In this case the methods are combined by using the CP algorithm to provide an approximate solution, which the MDS algorithm then refines until the desired solution is obtained. Thus we overcome two difficulties in this type of problem: on the one hand, great precision is difficult to achieve with the CP approach due to the combinatorial explosion of enumerating the variable domains; on the other hand, a good initial guess is very important for the efficiency of an MDS algorithm.

## 1 Introduction

There are two major types of methods to solve constraint satisfaction or optimisation problems, whose goal is to assign to each variable of the problem a value within its domain such that all the constraints of the problem are satisfied and (in optimisation problems) these values optimise some objective function. Constructive methods complete a partial solution (i.e. an assignment to only some of the variables) by progressively assigning values to the remaining variables. Repairing methods take complete solutions, that do not meet the requirements (by either violating some constraints or by not optimising the objective function) and repair them, changing the values assigned to some of the variables. Due to the combinatorial nature of the problems, both methods face some difficulties.

In constructive methods, it is often the case that partial solutions, although not violating any constraint between the already assigned variables, cannot be extended to a complete solution, but this inconsistency is difficult to check without testing all possible combination of values to the unassigned variables (a combinatorial procedure). A number of techniques can be used that circumvent this difficulty in many problem instances, namely a) the use of heuristics to guess the most promising values to assign to the variables; and b) constraint propagation. In this latter case, as the partial solution is constructed, values that cannot complete the partial solution are removed from the domains of the remaining variables.

A number of techniques can be used that typically prune a significant number of values from the domains at acceptable computing costs. This is the case with general purpose techniques such as forward-checking arc-consistency techniques [28], as well as with special purpose techniques for some kinds of problems,

namely global constraints [29] that achieve good pruning efficiency with algorithms of polynomial complexity.

Notwithstanding the pruning achieved by these constraint propagation techniques, they cannot be used alone to solve many combinatorial problems, if their size is too large. However, constraint propagation is an important technique, specially if it can be integrated with other methods (in addition to the typical use of heuristics).

Repairing methods depend on the ability of searching efficiently and effectively around the current solution(s). Notice that in this case it is quite common to have the current solutions to violate some constraints; if the violation of the constraints is taken into account in the objective function, the optimality of this function guarantees the feasibility of the solution. With the notable exception of genetic algorithms [30], repairing methods often rely on local search, whereby a solution is changed into another in its neighbourhood, by changing slightly the value of its variables.

A major pitfall of these methods is the possibility of the current solutions to be trapped into local optima, from where no neighbour solution has a better objective function. Although there are many possibilities of escaping from these local maxima (using a number of metaheuristic techniques, such as simulated annealing or tabu search or random restarts [30]) repairing methods are strongly dependent on the possibility of modelling the problem in such a way that the objective function does not exhibit too many local optima and/or that the starting solution is relatively close to the optimum.

In continuous domains, or continuous approximations to finite domains, changes in the current solution are directed towards the optimality of the objective function. In this case the efficiency of the local search rely heavily on the possibility of determining the adequate changes on the variables, and changes are often directed by the gradient of the objective function near the current solution.

Both problems (local optima traps and determination of the direction of change) may depend significantly on the modelling of the problem that is chosen. Regardless of the model, however, starting the search in the vicinity of an optimal solution is a key factor for the success of the local search.

In this paper we describe the combination of two methods, based on constraint propagation and local search respectively, and show how this combination achieves results that either of them, alone, would hardly achieve. The problem to be solved is one of distance geometry, whose goal is to find, given the bounds of the distances between pairs of points in a  $p$  dimensional space, the localisation in space of all the points. A special instance of this problem arises in trying to determine the structure of proteins from nuclear magnetic resonance (NMR) data, in which case the points are the atoms that compose the proteins and the bounds of the distances between two atoms are obtained by analysis of the NMR data and the knowledge of the amino acids that compose the protein.

The paper is organised as follows. Section 2 describes two techniques that were previously used separately to determine the structure of proteins from bound distances. Section 2.1 describes a constraint propagation technique that was previously used to handle NMR data, and discusses its main features. Section 2.2 describes a local search optimisation technique, whose model based on Multidimensional Scaling (MDS) seems particularly suited to these kinds of problems, and discusses the limitations of the method. Section 3 describes some experiments that were performed by combining the two methods and discusses the results obtained. Finally, section 4 summarises the main conclusions and discusses further work.

## 2 The Constraint Propagation and Multidimensional Scaling Techniques

The objective of both techniques is to determine the coordinates (in an arbitrary  $p$  dimensional space) of a set of variables such that the distances between pairs of variables respect or approach as much as possible predefined boundaries.

### 2.1 Constraint Programming

Without loss of generality, the Constraint Programming (CP) algorithm is described for the particular case of molecular structure determinations, where each variable corresponds to the position of one atom and the problem is restricted to three dimensions. In this initial stage each atom domain is represented as sets of cuboid regions [15]. This is a simplified representation, with one cuboid defining the allowed positions for the atom and a set of zero or more cuboids defining the regions within the former from which the atom must be excluded. The propagation of distance constraints is also simplified to preserve this representation of the domains, by considering that a distance constraint to a point defines a cubic and not a spherical volume. This compromises the precision of the results but that is not a significant problem since precision is not an important factor at this stage.

The algorithm is represented schematically in pseudo-code below.

```
Place all atoms in AtomList
Place all atoms in CurrentEnumeration
Initialise domains for three atoms in AtomList
repeat
  if ArcConsistency then
    for all X in AtomList do
      if DomainSize(X)<MinimumSize then
        begin
          Remove(X, AtomList)
          Remove(X, CurrentEnumeration)
        end
      end for
    else
      Backtrack
      RemoveSmallestDomain(X, CurrentEnumeration)
    SplitDomain(X)
    if IsEmpty(CurrentEnumeration) then
      Copy AtomList to CurrentEnumeration
  until IsEmpty(AtomList) or Fail or BacktrackCount>MaximumBacktrack
```

The initialisation of the three atom domains fixes the whole structure in rotation and translation space, which is irrelevant for the structure determination. Each iteration consists on imposing arc-consistency on the whole system and then splitting the domain of an atom in two halves and selecting one (SplitDomain). By imposing arc-consistency [3, 11] we insure that all atom pairs have mutually consistent domains. If an atom domain is sufficiently reduced this atom is removed from AtomList, which means it will no longer be split. The list CurrentEnumeration is used to perform the domain splitting in a round robin sequence, ensuring that all atoms in AtomList are split at least once before any is split a second time.

Whenever ArcConsistency fails (meaning that one atom domain has become empty) the program backtracks to the last domain split and selects an alternative. It is easy to see that demanding a greater

precision, i.e. a smaller domain size needed for an atom to be removed from the active AtomList, will increase the number of domain splitting operations in which an enumeration choice must be made. This results in a combinatorial explosion in the size of the search tree, making it impractical to obtain a precise result with this method alone.

Thus the algorithm proceeds until either it succeeds by enumerating all atoms to the desired precision, or fails. Failure can be the result of exploring completely the enumeration tree in an impossible constraint system. However, since the problems tend to be large (thus generating huge enumeration trees) and the aim is to rapidly produce an approximate solution, typically the CP algorithm will terminate in failure after a predefined number of Backtrack operations. Regardless of the success of the CP algorithm, the present domains, even if partially incorrect, are saved for use in the optimisation stage.

This allows the program to provide some solution even if it is impossible to respect all constraints, a valuable feature since the interpretation of experimental data is often a hard and error prone process.

The algorithm was implemented in Borland's Object Pascal language using Delphi, and a more detailed description can be found in reference [15].

## 2.2 Multidimensional Scaling

The problem of determining a structure for which (some of) the interatomic distances satisfy lower and upper bounds can be formulated in many ways. This section relies on a formulation derived from classical multidimensional scaling (MDS).

### 2.2.1 Classical Multidimensional Scaling

We begin with some formal definitions. First, a symmetric  $n \times n$  matrix  $\Delta = (d_{ij})$  is a *dissimilarity matrix* if  $d_{ij} \geq 0$  and  $d_{ii} = 0$ . Next, an  $n \times n$  matrix  $\Delta = (d_{ij})$  is a  $p$ -dimensional Euclidean *distance matrix* if there exists a configuration of points  $x_1, \dots, x_n \in \mathbb{R}^p$  such that  $d_{ij}$  is the Euclidean distance between  $x_i$  and  $x_j$ . We denote the closed cone of all such matrices by  $\Phi_n(p)$ . Finally, we store the coordinates of  $x_i$  in row  $i$  of the  $n \times p$  *configuration matrix*  $X$ .

In traditional psychometric or statistical applications of MDS, one observes a set of  $n$  objects, e.g. colour stimuli. The data, stored in a dissimilarity matrix  $\Delta$ , represent information about the dissimilarity of each pair of objects. Metric MDS constructs a configuration matrix  $X$  from  $\Delta$ .

The  $x_i$  correspond to the objects and the  $d_{ij}$  approximate the  $d_{ij}$ . Thus,  $X$  is a graphical representation of  $\Delta$ . In the present case of molecular conformation, the objects are atoms and the goal is to recover the 3-dimensional structure of a molecule.

Anticipating Section 2 below, we consider a molecule with  $n$  atoms and ask:

**Question 1** Suppose that we knew all  $m = n(n-1)/2$  interatomic distances, stored in the  $n \times n$  distance matrix  $\Delta^0$ . Can we determine a configuration matrix  $X$  such that  $D(X) = \Delta^0$ ?

A constructive answer to Question 1 is contained in a famous embedding theorem from classical distance geometry. Let  $\Omega_n(p)$  denote the positive semidefinite matrices of rank  $\leq p$ . For any  $n \times n$  matrix  $A$ , let

$$t(A) = \frac{1}{2} \left( I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \right) A \left( I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \right) \quad (1)$$

where  $I_n$  denotes the  $n \times n$  identity matrix and  $\mathbf{1}_n$  denotes the  $n$ -vector of ones. Let  $*$  denote the Hadamard product of matrices. Then

**Theorem 1**  $D^0 \hat{\mathbf{I}} \mathbf{F}_n(p)$  if and only if  $B^0 = \mathbf{t}(D^0 * D^0) \hat{\mathbf{I}} \mathbf{W}_n(p)$ . Furthermore, if  $B^0 \hat{\mathbf{I}} \mathbf{W}_n(p)$ , then the factorisation  $B^0 = XX'$  produces an  $n \times p$  configuration matrix  $X$  for which  $D(X) = D^0$ .

Theorem 1 is a slight modification, due to Torgerson [24], of a result that was independently discovered and proved by Schoenberg [22] and by Young and Householder [30]. It provides an affirmative answer to Question 1: if we compute the spectral decomposition  $\tau(\Delta^0 * \Delta^0) = Q \Lambda Q'$  and let  $X$  denote the first  $p=3$  columns of  $Q \Lambda^{1/2}$ , then  $D(X) = \Delta^0$ .

Now we expand the scope of our inquiries:

**Question 2** Suppose that we obtained measurements (subject to error) of all  $m=n(n-1)/2$  interatomic distances, stored in the  $n \times n$  dissimilarity matrix  $\Delta^0$ . Can we determine a configuration matrix  $X$  such that  $D(X) \doteq \Delta^0$ ?

Question 2 virtually defines metric MDS, for which a variety of approaches are available. The classical approach of Torgerson [24] and Gower [9] finds an approximate solution to the embedding problem with "fallible" data. This approach can be formulated as follows:

$$\text{minimize } \|B - \mathbf{t}(\Delta * \Delta)\|_F^2, \text{ subject to } B \in \Omega_n(p) \quad (2)$$

The objective function is sometimes called the strain criterion. The elegance of the classical approach derives from the following well-known result:

**Theorem 2** Let  $I_1 \geq I_2 \geq \dots \geq I_n$  denote the eigenvalues of  $B^0 = \mathbf{t}(D^0 * D^0)$ ,  $L = (I_1, \dots, I_n)$ , and let  $Q L Q'$  be the spectral decomposition of  $B^0$ . Let  $\bar{I}_i = \max(I_i, 0)$  for  $i=1, \dots, p$  and  $\bar{I}_i = 0$  for  $i = p+1, \dots, n$ . Let  $\bar{\Lambda} = \text{diag}(\bar{I}_1, \dots, \bar{I}_p)$ . Then  $B^* = Q \bar{\Lambda} Q'$  is a global minimiser of Problem 2 and the global minimum is

$$F_p \circ \mathbf{t}(\Delta_0 * \Delta_0) = \sum_{i=1}^p [I_i - \max(I_i, 0)]^2 + \sum_{i=p+1}^n I_i^2 \quad (3)$$

This result provides an affirmative answer to Question 2: if we let  $X$  denote the first  $p=3$  columns of  $Q \bar{\Lambda}^{1/2}$ , then  $D(X) \doteq \Delta^0$ .

### 2.2.2 Problem Formulation

Let  $[L, U]$  denote a rectangle of bound constraints on the dissimilarities. Continuing in the spirit of Section 1.2.1, we now ask:

**Question 3** Suppose that  $L$  and  $U$  are dissimilarity matrices of lower and upper bounds on the interatomic distances. Can we determine a configuration matrix  $X$  such that  $D(X) \doteq D \hat{\mathbf{I}} [L, U]$ ?

The extension of classical MDS to the case of bound constraints for the purpose of solving distance matrix completion problems was proposed by Trosset [29], who briefly noted its relevance to molecular conformation problems, on which he focussed in [26]. The same author proposed an analogous extension to order constraints in [27] and unified the mathematical theory common to these extensions in [28].

To address Question 3, we allow the  $\delta_{ij}$  to vary subject to  $l_{ij} \leq \delta_{ij} \leq u_{ij}$  and search for the dissimilarity matrix for which (3) is minimal. This results in the optimisation problem

$$\text{minimize } F_3 \circ \mathbf{t}(\Delta * \Delta), \text{ subject to } \Delta \in [L, U]$$

We reparameterise this problem in terms of the squared dissimilarities, obtaining

$$\text{minimize } F_3 \circ \mathbf{t}(\Delta), \text{ subject to } \Delta \in [L * L, U * U] \quad (4)$$

We provide an affirmative answer to Question 3 by finding a matrix  $\Delta^*$  of squared dissimilarities that solves Problem (4), then applying the methods available for Questions 1 and 2 to recover  $X$ .

Problem (4) is a nonlinear optimisation problem with simple bound constraints. The objective function  $F_3 \circ \mathbf{t}$  was studied by Trosset [28], who derived an explicit formula for its gradient. However, analytic second derivatives of  $F_3 \circ \mathbf{t}$  are not available.

Problem (4) requires  $m=n(n-1)/2$  variables, one for each interatomic distance. For large molecules, this is a great many variables. In contrast, the more conventional parameterisation by Cartesian coordinates,  $D=D(X)$ , requires only  $3n - 6$  variables (one for each atomic coordinate, less six to remove translational and rotational indeterminacy). However, certain advantages appear to accrue from allowing each (squared) interatomic dissimilarity to vary independently of the others - additional memory is the price that we pay for these advantages.

### 2.2.3 Numerical Algorithm

For  $n$  small, as in [29], one can solve Problem (4) with a quasi-Newton method for bound-constrained optimisation. For molecules with hundreds or thousands of atoms ( $n$  large), as is typical in structural molecular biology, the memory requirements of such methods are prohibitive.

The Hessian matrix is completely dense; hence, to store a symmetric approximation of the Hessian matrix requires  $(n^4 - 2n^3 + 3n^2 - 2n)/8$  variables, which is unrealistic for large molecules.

One possibility is to try to solve Problem (4) with a first-order method, *e.g.* the gradient projection method discussed in [28]. Unfortunately, the objective function does not have much curvature near solutions. This can be inferred from the well-known fact that classical MDS is extremely stable under perturbations of the dissimilarity data, a phenomenon that has been studied by Mardia [17] and by Sibson [23]. Hence, we would expect first-order methods to be inefficient.

In light of the preceding, our computational dilemma is to balance the need for second-order information with the expense of managing it. A natural compromise is a limited-memory method. Such methods use a small number of recent updates to approximate the current Hessian matrix of the objective function. The numerical experiments reported in the following section used a limited memory algorithm that was developed by Byrd, Lu, Nocedal, and Zhu [5] for solving large nonlinear optimisation problems with simple bound constraints. The code (L-BFGS-B, version 2.1) was described by Zhu, Byrd, Lu, and Nocedal [31] and is known to run very efficiently.

## 2.3 Combining CP and MDS

We believe that the CP and the MDS approaches described in the previous sections compensate for each other's weaknesses. The strength of the constraint programming approach lies in its ability to find an approximate solution (a solution in which some constraints are violated) with remarkable ease. However, as one demands solutions that violate fewer and fewer constraints, the computational expense of finding a solution grows exponentially. In contrast, the most difficult aspect of the MDS approach is finding good initial dissimilarities from which to begin optimising. Although this is not a substantial difficulty in simple problems with a small number of atoms, experiments with larger molecules and more realistic problems suggest that it will become so as the number of atoms increases. Furthermore, in real life problems only a relatively small set of interatomic distances are known in advance to be constrained, which makes it necessary to propagate these constraints to obtain enough information to generate a good initial dissimilarity matrix.

Our experiments suggest that, given good initial dissimilarities, MDS algorithms are more efficient than constraint programming at producing very accurate solutions. Accordingly, we propose a hybrid approach in which constraint programming is used to produce the inputs for an MDS algorithm.

A tighter coupling between the constructive CP method and the reparative optimisation would be desirable. One possibility would be to use an optimisation algorithm to help with the variable and domain choices during enumeration. However, the MDS algorithm works in a different domain, with the variables being the interatomic distances in a dissimilarity matrix and not the atom positions. Furthermore, the MDS optimisation can take approximately one hundred times longer than the CP stage for very accurate solutions due to the large number of variables (on the order of the square of the number of atoms). These two factors make it hard to efficiently integrate MDS into the CP stage.

Currently under study is the integration of another optimisation method with CP, consisting of a conjugate gradient minimisation of the molecule torsion angle [16]. This has the advantages of high computation speed (due to the small number of variables, which are approximately 5 times less than the number of atoms) and of operating in a three-dimensional structure, which may allow a tighter integration of the two methods. However, there are two shortcomings to this optimisation method; considerably lower accuracy than MDS, due to the lower number of variables increasing the chances of stopping at local minima, and being problem dependent, as the torsion angle model can only be applied to molecular structure determination.

At this stage it is conceivable that the best solution for the protein structure determination problem will involve all three approaches, but in this paper we present only the CP-MDS combination.

### 3 Results

In this section we show the results obtained using either method (MDS and CP) alone and using both methods combined. In all examples the Hydrogen atoms were ignored, so whenever the atom numbers are mentioned these refer only to non-Hydrogen atoms.

#### 3.1 Preliminary MDS Results

We report results for a molecule with  $n=38$  atoms and a molecular fragment with  $n=63$  atoms using only the MDS method.

##### Example 1

Crisma et al [7] described a small molecule, an analogue of the antibiotic trichogin A IV, that contains 7 oxygen atoms, 4 nitrogen atoms, and 27 carbon atoms (we did not consider hydrogen atoms). Trosset and Phillips [25] derived plausible lower (L) and upper (U) bounds on the  $703 = 38 \times 37 / 2$  interatomic distances associated with these  $n=38$  atoms. We attempted to solve Problem 4 with these bounds, i.e. to find squared dissimilarities  $\Delta \in [L * L, U * U]$  for which  $F_3 \circ t(\Delta) \stackrel{\Delta}{=} 0$ , using version 2.1 of L-BFGS-B, written in Fortran 77 and run on a Sun Ultra 2 (Model 1200). A typical run required just 25 seconds to obtain  $F_3 \circ t(\Delta^{897}) = 1.888 \times 10^{-7}$ . When classical MDS was used to obtain an actual 3-dimensional Euclidean distance matrix D from  $\Delta^{897}$ , only six constraints were violated and by a maximum of  $2.7 \times 10^{-4}$  angstroms — much greater accuracy than is required in practice.

## Example 2

Hendrickson [10] provided a set of 236 (exact) interatomic distances for a 63-atom molecular fragment. Given a partial distance matrix with these 236 distances specified, the problem is to complete the partial matrix to a 3-dimensional distance matrix. Moré and Wu [19] attempted to solve a closely related (and presumably easier) problem in which each specified distance  $d_{ij}$  was replaced with lower and upper bounds  $[l_{ij}, u_{ij}] = (1 \pm \epsilon) d_{ij}$ , for various choices of  $\epsilon$ . These problems are fairly realistic in the sense that comparable bounds might be derived if the molecule's 3-dimensional structure was unknown. Furthermore, Hendrickson [10] determined that a structure with the 236 specified distances is nearly rigid and quite difficult to determine. Indeed, for  $\epsilon \leq 0.02$ , Moré and Wu were rarely able to find global minimizers using a global continuation method with 100 random starts on Argonne National Laboratory's IBM SP parallel system:

Relative error ( $\epsilon$ )	.10	.08	.06	.04	.02	.01
Number of successes	96	76	69	30	4	0

Notice that the problem became more difficult as  $\epsilon$  decreased.

In contrast, we formulated the embedding problem as Problem (4) and set the appropriate 236 squared dissimilarity variables equal to the 236 specified squared distances. Lower and upper bounds on the remaining variables were derived by (i) imposing lower bounds that correspond to the repulsive forces that keep unbonded atoms apart, and (ii) applying the triangle inequality. After obtaining a solution, classical MDS was used to obtain a 3-dimensional Euclidean distance matrix and the maximal relative error of the calculated distances corresponding to the 236 known distances was computed. (Because the 236 distances are fixed while optimising, error is only introduced during the final application of classical MDS. Assuming that one can find a global minimiser, the maximal relative error can be controlled by the tolerance specified in the convergence criteria.) No global optimisation strategy was employed. Ten random starts were obtained by simulating  $Z_{ij} \sim \text{Uniform}(-.2, +.2)$  and setting

$$\delta_{ij} = (.75 + Z_{ij}) u_{ij}^2 + (.25 - Z_{ij}) l_{ij}^2$$

The convex problem in  $p=63$  dimensions was solved with a tolerance on the norm of the projected gradient of  $10^{-3}$  (this required 20—29 seconds) and the problem in  $p=3$  dimensions was then solved with a very stringent tolerance of  $10^{-8}$  (this required 732—1088 seconds). A global solution was found on each run and the *largest* relative error obtained was  $1.773 \times 10^{-6}$ .

These results show that the MDS algorithm is capable of finding very precise solutions to structural problems. However, when the number of atoms is significantly larger (as is typically the case in applications of NMR spectroscopy), it may be difficult to find an imprecise solution for MDS to refine. This occurs when the optimisation problem is plagued by nonglobal minimizers. A natural way to address this difficulty is to provide the optimisation algorithm with a good initial guess. This is especially helpful for molecular conformation problems because one usually can bound only the distances between nearby atoms, whereas the MDS algorithm requires an initial guess for each interatomic distance. Thus, a complementary approach that can quickly calculate plausible initial guesses is precisely what is needed to improve the MDS algorithm.

### 3.2 Preliminary CP Results

The CP algorithm described in section 1 was implemented in Borland's Object Pascal using the Delphi 5 compiler, and was run on a Pentium II PC at 300MHz. In the next section we will show the approximate solutions obtained in the cases where a large amount of information about the structure is available and the distance constraints can be known with great precision. However, in practical applications of structural NMR the number of constraints available can be one order of magnitude smaller and the Lower to Upper bounds range up to two orders of magnitude larger. Fortunately the CP algorithm is quite robust to the degradation of information available. As shown in [13], the CP algorithm can generate approximate solutions with experimental data supplied by [8], and in simulations of low quality realistic distance constraint information [14] the algorithm performed well and generated solutions where the root mean square deviation for the atomic positions differed from the target structure only by about 10%-15% of the structure size.

The great disadvantage of the CP algorithm by itself is the combinatorial explosion in the enumeration when more precise solutions are required. Experimental results [14, 13, 15] show that when the enumeration is made to proceed to domain sizes smaller than  $2\text{\AA}$ , performance degrades rapidly and the application of this algorithm alone is not practical at this level.

### 3.2 Combined CP-MDS Results

The method described in Section 1 was tested on five proteins of known structure. The structures were obtained from the Protein Data Bank, maintained by the Research Collaboratory for Structural Bioinformatics. Two structures (Mutant P53 Anti-Oncogene and Phosphotransferase, PDB codes 1AU1 and 1PHO respectively) were obtained by Nuclear Magnetic Resonance (NMR) spectroscopy. The remaining three (Desulfiredoxin, Bovine Trypsin inhibitor and Barstar, PDB codes 1DXG, 1BPI and 1BRS respectively) are X-Ray structures.

The sizes of the structures chosen are representative of the practical range of current NMR technology, and to simulate the Nuclear Overhauser Effect (NOESY) technique, widely used in NMR structural studies, each structure was used to generate a set of distance constraints between atom pairs.

All atom pairs closer than  $7\text{\AA}$  were used to generate a constraint pair, where the distance between two atoms is constrained by both a lower and upper boundary. The values for the boundaries were taken to be the distance value in the known structure plus or minus  $0.1\text{\AA}$ .

This method provides a large number of very tight constraints. The number of constraints expected from an actual NMR experiment would be an order of magnitude smaller, and the distance boundaries at least an order of magnitude larger than the ones used for these tests. Although this way we are not accurately simulating a real NMR structural determination, these 'tighter' problems allow for a more stringent test on the performance of the algorithm.

The CP stage was run on a PII processor at 300MHz, and the calculations took between 20 seconds (Desulfiredoxin) and 2 minutes (Barstar). The MDS stage was run on a PIII processor at 6000MHz, with the calculations taking between 15 minutes (Desulfiredoxin) and 6 hours (Barstar) hours, although computation time can be reduced by allowing for less accurate solutions.

For each test case below we show the name of the protein, the number of atoms and of the distance constraint pairs (upper and lower bound) taken from the known structure. These were all interatomic distances below  $7\text{\AA}$  and the boundaries were considered to be equal to the interatomic distance  $\pm 0.1\text{\AA}$ .

For all other distance constraints, for a total of  $n \times (n-1)$  pairs, the boundaries were considered to be 7.0Å and infinity (100.0Å).

For each case we also show the quality of the CP and MDS structures by measuring the root mean square deviation (Rmsd) from the target structure and the average constraint violations for the constraints taken from the structure. The violations that can still be observed are caused by the premature termination of the optimisation. Since the dissimilarity matrix when the process is terminated is still significantly different from a configuration matrix, when the nearest configuration matrix is found constraints are still violated. Given enough time the program would converge to a solution with no constraint violations, but the trade-off between accuracy and calculation economy is still being under study, so these results show a tentative exploration of the possibilities.

As these results show, the combination of these two complementary approaches makes it possible to solve problems of a size and with a precision beyond the practical capabilities of each method alone, for neither can the CP algorithm reach an accurate solution in an acceptable time, nor can the MDS algorithm without an approximate initial guess.

#### **Desulforedoxin (monomer) [2]: 260 Atoms, 5951 Constraint pairs**

CP Structure



Rmsd: 2.1 Å  
Average Violation: 0.82 Å

MDS Structure



Rmsd: 0.2 Å  
Average C. Violation: 0.04 Å

Target Structure



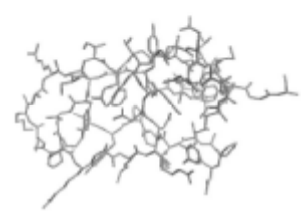
#### **Trypsin Inhibitor [21]: 448 Atoms, 11613 Constraint pairs**

CP Structure



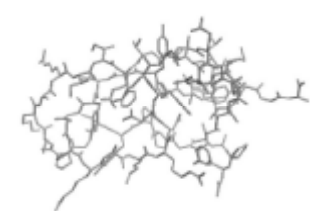
Rmsd: 2.8 Å  
Average Violation: 0.73 Å

MDS Structure



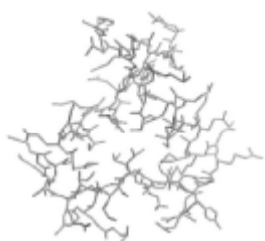
Rmsd: 0.02 Å  
Average C. Violation: 0.05 Å

Target Structure



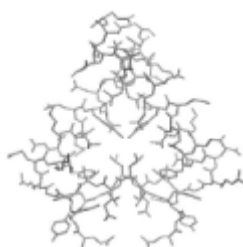
**Mutant P53 Anti-Oncogene [18]:** 514 Atoms, 12938 Constraint pairs

CP Structure



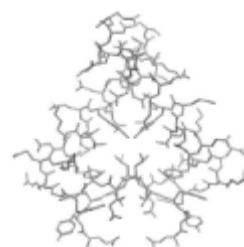
Rmsd: 2.5 Å  
Average Violation: 0.68 Å

MDS Structure



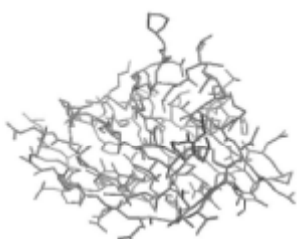
Rmsd: 0.02 Å  
Average C. Violation: 0.05 Å

Target Structure



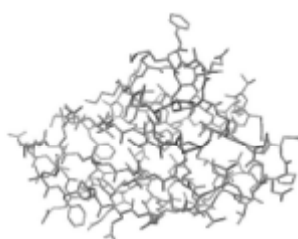
**Phosphotransferase [12]:** 639 Atoms, 17206 Constraint pairs

CP Structure



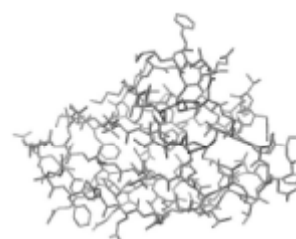
Rmsd: 2.8 Å  
Average Violation: 0.67 Å

MDS Structure



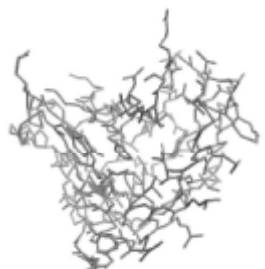
Rmsd: 0.02 Å  
Average C. Violation: 0.01 Å

Target Structure



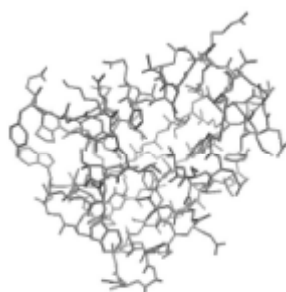
**Barstar Mutant [4]:** 693 Atoms, 18996 Constraint pairs

CP Structure



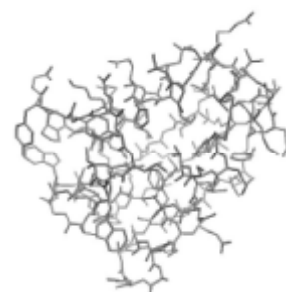
Rmsd: 4.1 Å  
Average Violation: 0.76 Å

MDS Structure



Rmsd: 0.02 Å  
Average C. Violation: 0.01 Å

Target Structure



## 4 Conclusion and Further work

In this paper we have shown the potential of combining constraint propagation with local search in order to solve problems that cannot be solved, with reasonable efficiency, by neither of the two techniques alone. Specifically, we have explored a two stages method (first we used a purely constraint propagation to produce an approximate solution that is then refined by a local search algorithm), but other combinations could be used, namely using some degree of local search to produce heuristics to split the domains of the variables between two constraint propagation steps. We intend to try this method in the future, but we have not exploited it yet.

The results that we have obtained clearly show the merit of applying this combination of techniques for the determination of protein structure from nuclear magnetic resonance data. Nevertheless there are some general issues that should be pointed out.

The examples shown in the previous section are not a realistic representation of the experimental data available in real life structural NMR problems. Although we know that the CP algorithm can perform satisfactorily in real life situations [13], the MDS algorithm was still not properly tested with experimental data. Nevertheless, the MDS optimisation algorithm gave extremely accurate solutions to the problems tested so far.

Despite the improvements of their combination over their independent use, finding the adequate models for both constraint propagation and local search is still an important issue. For the particular problem that we described, it was critical to develop a special purpose model of the atoms domains together and a special purpose kind of arc consistency adequate for distance constraints applied to these domains. On the other hand, the local search method, based on the multidimensional scaling turned out to be quite appropriate to address this kind of problems. Some rough experiments that were done with a much more naïve (but also efficient) local search optimisation (Powell multi dimensional minimisation) could not achieve nearly as good results as those obtained with the MDS optimisation. Torsion angle optimisation using more sophisticated methods such as conjugated gradient has a good potential for the particular problem of protein structure determination, since the number of variables is reduced from the order of  $n^2$  of the MDS method to the order of  $n$ . However, this optimisation technique takes advantage of specific features of this problem and so has the drawback of not being generally applicable to other distance geometry problems.

Regarding the efficiency of the method in the particular problem of protein structure determination from NMR data, at this moment we do not know what is the best trade-off between the time to run the MDS stage (the CP stage is very fast) and the quality of the solution produced. The calculation times for the examples in the previous section were between 20 minutes and 3 hours, but the accuracy obtained in most cases is still significantly higher than what is meaningful for real life problems and so the method presented here can potentially out perform other CP-Optimisation combinations such as described in [15] and [13], with the added significant advantage of being a generic distance geometry method.

## 5 References

1. A.K. Mackworth and E.C. Freuder, The complexity of some polynomial consistency algorithms for constraint satisfaction Problems, *Artificial Intelligence*, vol. 25, Elsevier, pp. 65-74, 1985

2. Archer M., Huber R., Tavares P., Moura I., Moura J.J., Carrondo M.A, Sieker L.C., LeGall J., Romão M.J., *Crystal Structure of Desulforedoxin from Desulfovibrio gigas Determined at 1.8 Å Resolution: A Novel non-Heme Iron Protein Structure* J.Mol.Biol. V. 251 690 1995
3. Bessière, C., *Arc-consistency and arc consistency again*. Artificial Intelligence, 65 (1994) 179-190
4. Buckle A.M., G.Schreiber, A.R.Fersht, *Protein-Protein Recognition: Crystal Structural Analysis of a Barnase-Barstar Complex at 2.0-Å Resolution* Biochemistry V. 33 8878 1994
5. Byrd R.H., P. Lu, J. Nocedal, and C. Zhu. *A limited memory algorithm for bound constrained optimization*. Journal on Scientific Computing, 16:1190--1208, 1995.
6. C.R. Reeves, *Modern Heuristic Techniques for Combinatorial Problems*", Mc Graw-Hill, 1995.
7. Crisma M., G. Valle, V. Monaco, F. Formaggio, and C. Toniolo. *N alpha-benzyloxycarbonyl-alpha-aminoisobutyryl-glycyl-L-isoleucyl-L-leucine methyl ester monohydrate*. Acta Crystallographica, 50:563--565, 1994.
8. Goodfellow B. J, Rusnak F., Moura I., Domke T., Moura J.J.G., NMR determination of the global structure of the <sup>113</sup>Cd derivative of Desulforedoxin: Investigation of the Hydrogen bonding pattern at the metal center, Protein Sc.7, 928-937 (1998)
9. Gower J.C.. *Some distance properties of latent root and vector methods in multivariate analysis*. Biometrika, 53:315--328, 1966.
10. Hendrickson B.A.. *The Molecule Problem: Determining Conformation from Pairwise Distances*. PhD thesis, Cornell University, 1991.
11. Hentenryck, P., Deville, Y., Teng, C. *A generic arc-consistency algorithm and its specializations*, Artificial Intelligence 57 (1992) 291-321
12. Jia Z., J.W.Quail,E.B.Waygood,L.T.J.Delbaere *The 2.0 Ångstroms Resolution Structure of Escherichia coli Histidine-Containing Phosphocarrier Protein HPR: a Redetermination* (to be published).
13. Krippahl L, Barahona P., *PSICO: Combining Constraint Programming and Optimisation to Solve Macromolecular Structures*, presented at ERCIM-2000, Padova, Italy
14. Krippahl L. *Determining Protein Structure through Constraint Programming*, 1999, MSc Thesis (in Portuguese) F.C.T./U.N.L.
15. Krippahl, L., Barahona, P., *Applying Constraint Programming to Protein Structure Determination*, Principles and Practice of Constraint Programming, Springer Verlag, 1999 289-302
16. Krippahl, L., Barahona, P., *PSICO: Solving protein structures with constraint programming and optimisation*. Submitted to Constraints, 2001
17. Mardia K.V.. *Some properties of classical multi-dimensional scaling*. Communications in Statistics---Theory and Methods, A7:1233--1241, 1978.
18. McCoy M., E.S.Stavridi, J.L.Waterman, A.M.Wieczorek, S.J.Opella, T.D.Halazonetis *Hydrophobic Side-Chain Size is a Determinant of the Three-Dimensional Structure of the P53 Oligomerization Domain* Embo J. V. 16 6230 1997
19. Moré J.J. and Z. Wu. *e-Optimal Solutions to Distance Geometry Problems via Global Continuation*. Preprint MCS-P520-0595, Mathematics & Computer Science Division, Argonne National Laboratory, Argonne, IL 60439-4844, May 1995.
20. N. Beldiceanu and E. Contejean, *Introducing Global Constraints in CHIP*, Journal of Mathematical and Computer Modelling, vol. 20, no. 12, Elsevier, pp. 97-123, 1994

21. Parkin S., B.Rupp, H.Hope, *The Structure of Bovine Pancreatic Trypsin Inhibitor at 125K: Definition of Carboxyl-Terminal Residues* (To be published)
22. Schoenberg I.J.. Remarks to Maurice Fréchet's article "*Sur la définition axiomatique d'une classe d'espaces distanciés vectoriellement applicable sur l'espace de Hilbert*". *Annals of Mathematics*, 38:724--732, 1935.
23. Sibson R.. *Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling*. *Journal of the Royal Statistical Society, Series B*, 41:217--229, 1979.
24. Torgerson W.S.. *Multidimensional scaling: I. Theory and method*. *Psychometrika*, 17:401--419, 1952.
25. Trosset M.W. and G.N. Phillips. *Deriving interatomic distance bounds from chemical structure*. In F. Seillier-Moiseiwitsch, editor, *Statistics in Genetics and Molecular Biology*, pages 276--287, Institute of Mathematical Statistics, Hayward, CA, 1999.
26. Trosset M.W. *Applications of multidimensional scaling to molecular conformation*. *Computing Science and Statistics*, 29:148--152, 1997.
27. Trosset M.W.. *A new formulation of the nonmetric STRAIN problem in multidimensional scaling*. *Journal of Classification*, 15:15--35, 1998.
28. Trosset M.W.. *Computing Distances Between Convex Sets and Subsets of the Positive Semidefinite Matrices*. Technical Report 97-3, Department of Computational & Applied Mathematics---MS 134, Rice University, Houston, TX 77005-1892, 1997.
29. Trosset M.W.. *Distance matrix completion by numerical optimisation*. *Computational Optimization and Applications*, 17:11--22, 2000.
30. Young G. and A.S. Householder. *Discussion of a set of points in terms of their mutual distances*. *Psychometrika*, 3:19--22, 1938.
31. Zhu C., R.H. Byrd, P. Lu, and J. Nocedal. *L-BFGS-B: Fortran Subroutines for Large-scale Bound Constrained Optimization*. Technical Report NAM-11, Department of Electrical Engineering & Computer Science, Northwestern University, Evanston, IL 60208, December 1994. Revised October 8, 1996