Molecular Conformation Search by Distance Matrix Perturbations

Ioannis Z. Emiris
Dept. Informatics & Telecoms, National Univ. of Athens, Greece
emiris@di.uoa.gr, http://www.di.uoa.gr/~emiris

Theodore G. Nikitopoulos
Univ. of Crete, Computer Science Dept., Heraklio 71409 Greece
nikitop@csd.uoc.gr, http://www.csd.uch.gr/~nikitop

July 10, 2003

Abstract: Three-dimensional molecular structure is fundamental in drug design and discovery, docking, and chemical function identification. The input to our algorithm consists of a set of approximate distances of varying precision; some are specified by the covalent structure and others by NMR experiments and application of the triangle or tetrangle inequality. The output is a valid conformation in a specified neighborhood of the input. We aspire that our approach helps in detecting outliers of the NMR experiments, and that it manages to handle inputs with partial information. Numerical linear algebra methods are employed for reasons of speed and accuracy. The main tools include, besides iterative local optimization, distance geometry and matrix perturbations for minimizing singular values of real symmetric matrices. Our algorithm is able to bound the number of degrees of freedom on the conformation manifold. A public domain MATLAB (or SCILAB) implementation is described; it can determine a conformation of a molecule with 20 backbone atoms in 3.79sec on a 500MHz PENTIUM-III.

Keywords: local conformation search, distance geometry, NMR, symmetric matrix perturbation, singular value minimization, MATLAB software

1 Introduction

Structural proteomics is today a major challenge in computational molecular biology and chemistry. More generally, drug design and discovery relies increasingly on structure-based methods in order to improve efficiency and guarantee completeness. Three-dimensional geometric (i.e. tertiary) structure is essential in function identification, docking of relatively small flexible ligands to macromolecules, as well as pharmacophoric pattern matching and site mapping. We use local search to identify molecular conformations where the pairwise Cartesian distances of backbone atoms (which correspond to a protein's residues), are known approximately. This is a crucial step in finding all, or the energetically favorable geometries, and in identifying degenerate conformations.

The problem of identifying the conformation of proteins of known amino-acid sequence, by using a model of residue-residue energy-like potential, was the underlying motivation in exploring the theory of *distance geometry* [Blu70, CH88, Hav98]. Given all pairwise distances between a set of points, their 3-D coordinates can be immediately obtained. Thus, the coordinates of the backbone atoms (e.g. on a protein's skeleton) are computed and the entire tertiary structure can be

eventually determined. Distance geometry applications have been quite successful (e.g. [BCDD90, Cou95, DH88, HKC83, Hav98]), achieving the conformational analysis of molecules with about 200 residues [MD02] as well as in the applications mentioned above, where the input data is obtained by different procedures, most notably by NMR (Nuclear Magnetic Resonance) experiments relying on the NOE (Nuclear Overhauser Effect). Distance information has been succesfully used by optimization approaches to compute molecular conformations. Today, distance geometry is revisited (e.g. [Cou95]) since it offers the possibility to assign a level of confidence to distance information, which may be quite imprecise, by treating some input as accurate while allowing us to perturb input that seems false or is simply not yet available due to the length of the phase of spectrum assignment in NMR. Certain methods consider entire families of proteins with homological similarities, and try to treat them with the minimal possible distance information. The latter may come from an on-line NMR process, where some of the intervals are inaccurate or simply not known. The goal is increased throughput, cf. e.g. [BKWK$^+$00, MBD01].

Our method, given a known valid conformation, explores nearby conformations lying on the same manifold, hence also topologically close to each other. This answers the need of biased sampling in order to avoid previously sampled configurations. Our algorithm offers the freedom to choose the direction of exploration and to search valid conformation in a neighborhood of the input. By systematically sampling the conformation space, we may compute several possible geometries when the input are interval constraints. For molecules or molecular substructures with few degrees of freedom (about 10), our methods are able to fully enumerate all realizable conformations. Note that a regular sampling may be suboptimal, because some solutions have bigger attractive regions in the space of starting parameters, and there can even be fractal boundaries between attractive regions [XOW$^+$00]. In addition, the algorithm is able to bound the dimension of the manifold of all allowable conformations.

In this work we propose numerical linear algebra methods for computing conformations of geometrically constraint molecules. We formulate molecular embedding as a *structured singular value (or eigenvalue) minimization* problem: Given distance approximations (or interval constrains, respectively), the aim is to find values near the given approximations (or in these intervals) so that the structure be embedded in euclidean space E_3. Although the algorithm is numerical for reasons of speed, it guarantees its output under certain assumptions. Our algorithm is based on an iterative local optimization method. It outputs conformations such as the one shown in Figure 1, for the example of a cyclic molecule.

The paper is organized as follows. The next section overviews the theory of distance matrices and how it has been used in conformation search. Section 3 contains the background on linear algebra perturbations and existing work in the area. Section 4 elaborates on our algorithm and presents our MATLAB and SCILAB implementations. Section 5 applies our implementation to cycloalkanes, whereas the following section reports on experimental results for more general molecules. Section 7 sketches current and further work.

2 Distance matrices

We review techniques related to distance matrices and then formalize their algebraic properties.

In our setting, the primary structure is considered as known, which enables us to deduce certain distances, such as those between covalently bonded pairs of backbone atoms. Holonomic constraints may also specify the distance of a pair of such atoms. Similarly, the bond angles can usually be determined from the covalent structure, while for fixed bond lengths there is a one-to-one relation between the bond angle and the geminal distance so that these distances can also be determined.

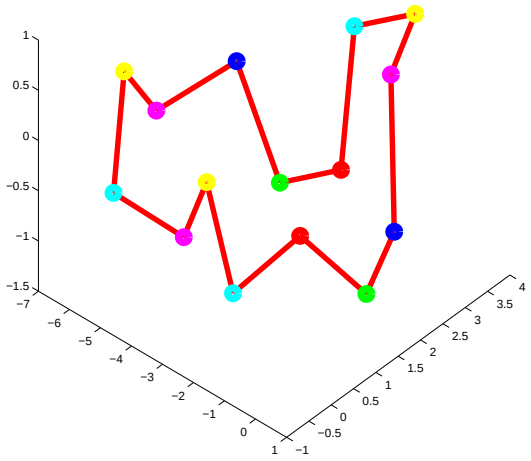

Figure 1: One backbone conformation computed by our program, for a cyclic molecule with 15 backbone atoms.

The distances across rotatable bonds usually vary within their *cis/trans* limits, to enable free rotation, and all the distances within any known rigid group of atoms (e.g., amino-acid residues or phenyl rings) are constrained to their known values. Distances that are unknown and not given by NMR are constrained to intervals obtained by the triangle (or tetrangle) inequality and must satisfy certain obvious bounds, such as the one corresponding to van der Waals forces. The problem of generating all triangle inequality bounds is equivalent to a shortest-path problem, which is well studied, of polynomial complexity, and for which efficient software exists [DH88, KNB92].

One variant of the problem can be stated as follows: Given an incomplete, undirected, weighted graph G, the *molecular Euclidean embedding* problem is that of mapping the nodes (or vertices) of G to points in the 3-dimensional Euclidean space E_3 so that any two nodes with an edge between them are mapped to points whose Euclidean distance equals the weight of that edge. This problem is NP-hard in E_k for any $k \geq 2$, even if all given distances lie in $\{1, 2\}$ [Sax79]. It is a global optimization problem, in the space of all molecular conformations, and it has been treated as such, c.f., e.g. [Hav98, KTO93, PL97].

Distance geometry had been implemented in certain packages, which seem today either not maintained or not publicly available. One package that is currently maintained and freely available is DGSOL [MW99], which relies on continuation methods for global optimization in order to trace the minimizing configurations; it uses a different approach from ours. A package that used to be widely used is EMBED, which explores the conformation space by random sampling. The key idea lies in minimizing an *error function*, which measures the total violation of the distance constraints after a certain best-fit embedding of the structure in E_3. Since there is a lot of freedom in choosing this function, it is possible to make it smooth and well-behaved for optimization. A number of different conformations have been obtained for molecules with about 100 atoms or more [Hav98]. To ensure completeness, *linearized embedding* uses the *metric matrix*, which contains the inner products between vectors defining local coordinate systems within the molecule [Cri92]. Other packages for molecular conformations using distance geometry include DGSOL [MW99], HELIX, DGEOM, DPSACE, VEMBED, and DYANA. The latter relies on local information, hence it handles well nearby atoms but has problems with those lying far apart on the chain. DYANA is an example of using local (spherical) coordinates, which offers an interesting general approach

(e.g. [Cri92, CGS89]). This and other software is found at [Exc, SS].

The speed of modern hardware has revived an interest into algebraic methods, which may handle efficiently substructures of small size as part of larger problems in structural biology. Hence, algebraic techniques have been applied to conformational search, since they offer completeness, raise no issues of convergence, and can certify their results. [Cri92, EM99, Hav97, HN95, MZW95, WS99]. Modeling the molecular problem in algebraic terms is achieved, in a general manner, by distance geometry. For instance, the one-dimensional manifold of boat conformations of cyclohexane is the solution of a system of quadratic equations. However, all of these methods have complexity exponential in the number of degrees of freedom, so they are limited to small molecules, say with at most a dozen of degrees of freedom. The goal of this paper is to exploit the power of distance matrices while studying molecules of larger sizes, by employing numerical linear algebra techniques.

It is time now for a formal presentation of distance geometry; for further details see [CH88, Hav98]. Suppose that there are n points; these shall correspond to the backbone atoms. Let $d_{ij}, i, j \in \{1, \ldots, n\}$, denote the euclidean distance between the corresponding nodes, with $d_{11} = \cdots = d_{nn} = 0$. We may consider the corresponding symmentric *distance matrix (or Cayley-Menger matrix)*

$$D(1, \ldots, n) := \begin{bmatrix} 0 & \frac{1}{2}d_{12}^2 & \ldots & \frac{1}{2}d_{1n}^2 & 1 \\ \frac{1}{2}d_{12}^2 & 0 & \ldots & \frac{1}{2}d_{2n}^2 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{2}d_{1n}^2 & \frac{1}{2}d_{2n}^2 & \ldots & 0 & 1 \\ 1 & 1 & \ldots & 1 & 0 \end{bmatrix},$$

which contains, besides the adjacency matrix as a (principal) submatrix, an additional row and column of units, with the diagonal being zero. In what follows, we use the sign function $\operatorname{sgn}(\cdot) \in \{-1, 0, 1\}$.

Theorem 2.1 [Blu70] *Let d_{ij} denote the distance between nodes i, j, $1 \leq i, j \leq n$. A necessary and sufficient condition for the distance matrix $D(1, \ldots, n)$ to express a point set embeddable in m-dimensional euclidean space E_m, $m \leq 3$ is, for some ordering of the nodes, to have (i)* $\operatorname{sgn} \det D(1, \ldots, k+1) = (-1)^{k+1}$ *for $k = 1, \ldots, m$, and (ii) for any nodes u, v, $m+1 \leq u, v \leq n$, $\det D(1, \ldots, m+1, u) = \det D(1, \ldots, m+1, v) = \det D(1, \ldots, m+1, u, v) = 0$.*

Part (i) for $k = 1$, $m \geq 1$ becomes $\operatorname{sgn} d_{i_1 i_2}^2 = 1$, which ensures that points $i_1, i_2 \in \{1, \ldots, n\}$ are distinct. For $k = 2$ and $m \geq 2$,

$$4 \det D = (d_{i_2 i_3}^2 - d_{i_1 i_2}^2 - d_{i_1 i_3}^2)^2 - 4 d_{i_1 i_2}^2 d_{i_1 i_3}^2,$$

so part (i) is equivalent to saying that, any three nodes $i_1, i_2, i_3 \in \{1, \ldots, n\}$ are not collinear and the respective distances satisfy the triangle inequality, i.e. $d_{i_1 i_2}^2 + d_{i_2 i_3}^2 > d_{i_1 i_3}^2$. Part (ii) is satisfied if matrix $D(1, \ldots, n)$ has rank $m + 2$.

Corollary 2.2 *With the notation of the previous theorem, $n \geq 4$ distinct points (not all coplanar) are embeddable in E_3 if and only if* $\operatorname{rank}(\mathrm{D}(1, \ldots, \mathrm{n})) = 5$.

The problem of mapping the input nodes to E_3, is equivalent to perturbing (or completing) matrix $D(1, \ldots, n)$ so that its rank becomes 5. In fact, we may use the matrix containing simply the squared distances d_{ij}^2, zeros on the diagonal and units on the last column and last row.

3 Matrix perturbations

This section presents relevant numerical linear algebra approaches.

Reducing a specific subset of eigenvalues and bringing them close to zero has been addressed in numerical analysis, e.g. [BJ95, Dem92, GHHW90, LEE96, WD95] and the references thereof. We shall focus on the latter approach, which studies the minimization of the last singular value. It proposes a modified Newton iteration in order to avoid instabilities near the minimum, where the derivative vanishes. Moreover, this modification ensures global convergence at a nearly quadratic rate, including in the case of arbitrary complex rectangular matrices. We shall use these results for real symmetric square matrices, because it is possible to devise *structured rank-reducing perturbations* which preserve (at least) symmetry, reality and zero diagonal by modifying only certain entries. We shall extend these methods in order to reduce more than one singular values, while mainting the structure nature of the perturbations. The latter means that it is possible to specify the set of perturbable entries, hence defining the direction of the search. If, moreover, we limit the magnitude of the perturbation per entry, we are able to search in a neighborhood of our choice.

Fundamental work exists concerning general distance matrices. A point set is said to be embeddable in a k-dimensional euclidean space E_k if and only if its distance matrix expresses the euclidean distances between the points in E_k. Let $\| \cdot \|_2$ stand for the 2-norm of vectors or of matrices. A relevant property of distance matrices is the following.

Theorem 3.1 [Mat85] *Let $e^T = [1, \ldots, 1]$ be the vector of units. For any vector s : $s^T e = 1$ and any square matrix M, define the norm $|M|_s := \| (I - es^T)M(I - es^T)^T \|_2$, where I is the identity matrix with the same dimension as M. Given a distance matrix D and any vector s, we can construct a new distance matrix D' embeddable in E_3 such that $|D - D'|_s$ is minimized.*

This construction is based on the truncation of the matrix of singular values, hence its computation is relatively fast. The drawback is that such a projection does not respect the bounds or other prior information on the entries of D.

The rest of the section presents the notions of linear algebra required; for further information see, e.g. [BP94, GV96, SS90]. Let us consider an $N \times N$ matrix M. When M is real and symmetric, its eigenvalues λ are real and its eigenvectors form an orthonormal basis of $\mathbb{R}^N$. The real symmetric eigenvalue, or spectral decomposition, problem is equivalent to solving matrix equation $M = Q^T \Lambda Q$, where $Q^T = Q^{-1}$ contains the eigenvectors as columns of Q, and diagonal matrix Λ contains the eigenvalues. Another useful matrix decomposition is the SVD (Singular Value Decomposition), which writes $M = Q_1 \Sigma Q_2^H$ where $Q_i^H = Q_i^{-1}$, $i = 1, 2$, and Q_i^H stands for the transposed conjugate matrix, and Σ is a diagonal matrix containing the *singular values* of M. The absolute values of the eigenvalues are the singular values. For a real symmetric M, both Q_i are real. Moreover, the associated left and right singular vectors are either equal or opposite to each other; the latter case occurs exactly when the corresponding eigenvalue is negative. The singular vectors are also equal to the corresponding eigenvectors within sign. The *rank* of a matrix is the number of nonzero eigenvalues, or the number of nonzero singular values. Rank computations may rely on the SVD because it is in practice faster and more stable numerically.

Our method makes use of the Moore-Penrose *pseudo-inverse* of a matrix M. This is the unique matrix M^+ satisfying $MM^+M = M, M^+MM^+ = M^+, (MM^+)^H = MM^+, (M^+M)^H = M^+M$. There are efficient and accurate public domain implementations for them. In particular, the SVD computation of the pseudo-inverse is as follows. Consider that $M = Q_1 \Sigma Q_2^T$ with $\Sigma =$

$\mathrm{diag}(\sigma_1, \ldots \sigma_\mathrm{r})$, then $M^+ = Q_2 \Sigma^+ Q_1^T$ where

$$\Sigma^+ = \mathrm{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_\mathrm{r}}, 0, \ldots, 0\right), \quad r = \mathrm{rank(M)}.$$

Let u and v be vectors of reals with $u^T v \neq 0$. Then, $\left(u^T M v\right) / \left(u^T v\right)$ is a *Rayleigh quotient*. If either u or v is (respectively, near) an eigenvector corresponding to an eigenvalue λ of M, then the Rayleigh quotient reproduces (resp. approximates) that eigenvalue. Iterative algorithms for eigenvalue computation and, in particular, spectral decompositions (like the power method), use the Rayleigh quotients to iteratively improve a numerical approximation of λ. In this paper, we shall use an iteration similar to the Rayleigh quotient iteration, for producing structured rank-reducing perturbations.

Proposition 3.2 (Extremal property of Rayleigh quotients) [GV96] *Let σ_j, v_j denote respectively the ordered singular values and (right) singular vectors of D. Then, $\sigma_n = \min_{\|x\|_2=1} x^T D x$, $x \in \mathbb{R}^N$, provided $x^T v_j = 0$ for each $n+1 \leq j \leq N$.*

Now we state the main property concerning the singular values (and eigenvalues) of matrices under small perturbations.

Theorem 3.3 (cf. [SS90, Thm.IV.2.3]) *Let R be a matrix with the same dimensions as matrix M. If $\sigma_k(M)$ denotes the k-th singular value of M, $k \leq \dim M$, then the function*

$$f(\xi) := \sigma_k(M - \xi R),$$

is differentiable with respect to real variable ξ, as long as $\sigma_k(M - \xi R)$ is distinct from all other singular values, for all ξ. Moreover, for M real and symmetric, $f'(\xi) = -u_k^T R v_k$, where u_k, v_k are the singular vectors of $M - \xi R$ associated to $\sigma_k(M - \xi R)$. An analogous result applies to eigenvalues and eigenvectors.

The theorem yields, immediately, the following Newton iteration for minimizing the smallest singular value and, eventually, driving it to zero. This makes the matrix singular. The iteration is defined as follows:

$$\xi \leftarrow \xi + [u_N^T (M - \xi R) v_N] / (u_N^T R v_N) = (u_N^T M v_N) / (u_N^T R v_N),$$

where u_N, v_N are the singular vectors of $M - \xi R$ associated with singular value $\sigma_N(M - \xi R)$. Then, the N-th singular value of $M - \xi R$ approachs zero with a quadratic rate. In [WD95], this approach is generalized to structured perturbations (and rectangular complex matrices).

In order to define the allowed structured perturbations in a general way, let R_{ij} be a *perturbation matrix* having the same dimensions as M and zeros in all entries except of units at entries (i, j) and (j, i), where $1 \leq i < j \leq N$. The number of independent R_{ij} is p, and for Cayley-Menger matrices corresponding to n points, we have $p \leq n(n-1)/2, n = N - 1$. Let $\mathcal{R}$ be a subspace of symmetric square matrices of dimension p generated by the $R_{k_i k_j}$, $1 < k_i < k_j \leq N$:

$$\mathcal{R} = \{\sum_{k=1}^{p} \xi_k R_{k_i k_j} \ : \ [\xi_1, \ldots, \xi_p] \in \mathbb{R}^p\}.$$

Algorithm 3.4 *The algorithm in [WD95], specialized to a square real symmetric $N \times N$ matrix M, consists of the following steps:*
0. Initialize the perturbation matrix $R \in \mathcal{R}$, possibly to the zero matrix.
1. Compute the SVD decomposition $M - R = U\Sigma V^T$, where the N-th singular value and vectors are denoted by σ_N, u_N, v_N respectively.
2. Let perturbation matrix $\Delta \in \mathcal{R}$ have minimum norm such that it minimizes $\|u_N^T \Delta v_N - u_N^T (M - R) v_N\|$.
3. Let $\alpha \leftarrow \|u_N^T \Delta (I - v_N v_N^T)(M - R)^+ \Delta\|$, $\gamma \leftarrow \min\{1, \|u_N^T \Delta v_N\|/(4\alpha\sigma_N)\}$, and $R \leftarrow R + \gamma\Delta$.
4. If $\gamma\|\Delta\|/\|M - R\|$ is smaller than some given tolerance, the algorithm stops; otherwise, go to step 1.

Step 2 reduces to finding vector $\xi \in \mathbb{R}^p$, which defines Δ in the basis of the R_{ij} discussed above, assuming $\|\Delta\| = \|\xi\|$. Now, $\xi = E^+ F$, where E is the p-dimensional vector $[\cdots v_N^T R_{ij}^T u_N \cdots]$, where the (i, j) range over all entries of M to be perturbed independently, and $F = u_N^T (M - R) v_N$.

Step 3 is designed so that the algorithm achieves nearly quadratic rate of convergence as it approaches the minimum. In implementing it, we can simplify the calculations by using relation $(I - v_N v_N^T)(M - R)^+ = [v_1, \ldots, v_{N-1}, 0]\ \mathrm{diag}[1/\sigma_1, \ldots, 1/\sigma_{N-1}, 0]\ U^T$. Here, $\mathrm{diag}[a_1, \ldots, a_N]$ stands for a diagonal matrix with entries $a_1, \ldots, a_N \in \mathbb{R}$.

Theorem 3.5 [WD95] *Suppose that $\|\Delta\|$ is always bounded; for this, it suffices that $\|E^+\|$ remains bounded. Then, Algorithm 3.4 makes $\sigma_N(M - R)$ approach zero, unless $u_n^T R v_N$ tends to zero for all $R \in \mathcal{R}$. The algorithm has global convergence at a nearly quadratic rate, even as the last singular value approaches its minimum value.*

4 Computing conformations

We extend the above algorithms in order to further reduce the rank of the matrix to $n - 1$, instead of $N - 1$, where n indexes henceforth the largest among the singular values that must be minimized and, eventually, reduced to zero. In practice, $n = 6$ in order for the given matrix to be perturbed to a valid Cayley-Menger distance matrix.

The main idea of our technique is the following. Suppose $N \times N$ matrix M is close enough to being embeddable in E_3. Formally, M must be in an attractive region of a valid Cayley-Menger distance matrix in terms of Newton's iteration. At each step of Newton's iteration for minimizing its n-th singular value, supposing the singular value is distinct, it suffices to compute $\Delta \in \mathcal{R}$ such that $u^T \Delta v = u^T (M - R) v$ and set the perturbation matrix $R \leftarrow R + \Delta \in \mathcal{R}$, where vectors u, v are associated to singular value $\sigma_n(M - R)$. This leads to a heuristic way of computing Δ since we have no formal manner to define a quantity like γ in Algorithm 3.4. Our method has been implemented, but it is not able to avoid certain obstacles, such as those related to the requirement of distinctness.

A much better approach uses further necessary conditions to facilitate the optimization process. Let the Dirac function be δ_{ij} such that $\delta_{ij} = 1 \Leftrightarrow i = j$ and $\delta_{ij} = 0$ otherwise. To identify a new conformation, all singular values smaller than σ_{n-1} must be close to zero. The following method uses the necessary conditions of this fact.

Algorithm 4.1 *In applying a Newton iteration for minimizing the n-th singular value (and all smaller singular values) of an $N \times N$ matrix D, it suffices to compute $\Delta \in \mathcal{R}$ such that*

$$u_n^T \Delta v_j = \delta_{nj} u_n^T (D - R) v_j, n \leq j \leq N, \tag{1}$$

and set in the next step $R \leftarrow R + \Delta$, *where singular vector* u_n *is associated to singular value* $\sigma_n(D - R)$. *Moreover, (1) should hold for each* u_i, $n \leq i \leq N$, *so the above relation becomes:*

$$u_i^T \Delta v_j = \delta_{ij} u_i^T \, (D - R) \, v_j, n \leq i, j \leq N.$$

For $n \neq j$, the above condition becomes $u^T \Delta v_j = 0$, which has the effect of keeping Δ small. Now, having fixed the basis $\mathcal{R}$ of the perturbation space, defining Δ is equivalent to finding vector $\xi \in \mathbb{R}^p$. The above conditions lead to the solution of a (dense) linear system $E\xi = F$, where

$$E = \begin{bmatrix} \vdots & \vdots \\ v_j^T R_{i_1 j_1} u_i, \ldots, v_j^T R_{i_p j_p} u_i \\ \vdots & \vdots \end{bmatrix}, \quad F = \begin{bmatrix} \vdots \\ \delta_{ij} u_i^T (D - R) v_j \\ \vdots \end{bmatrix},$$

where each pair i, j, for $i = n, \ldots, N, \ j = i, \ldots, N$ defines the corresponding row in matrix E and vector F.

The row dimension of E equals $\sum_{l=n}^{N} (N + 1 - l)$, for $N \geq n$, whereas its column dimension is p, the dimension of $\mathcal{R}$. If E is square, then LU or QR decomposition is applied for computing ξ. If the linear system $E\xi = F$ is overdetermined, we use the Moore-Penrose *pseudo-inverse*. This yields the solution $\xi = E^+ F$, optimal in a *least-squares* sense,

For example, for the cyclohexane to be examined in section 5, it is required to minimize the 6th singular value. Then F is a 3×1 vector, and E is the following 3×3 matrix:

$$E = \begin{bmatrix} v_6^T R_{25} u_6 & v_6^T R_{36} u_6 & v_6^T R_{47} u_6 \\ v_7^T R_{25} u_6 & v_7^T R_{36} u_6 & v_7^T R_{47} u_6 \\ v_7^T R_{25} u_7 & v_7^T R_{36} u_7 & v_7^T R_{47} u_7 \end{bmatrix}.$$

Proposition 4.2 *If some Cayley-Menger matrix is sufficiently close (in terms of Newton's iteration) to a given approximate distance matrix* D, *then a Cayley-Menger matrix exists and is unique if and only if the solution of (1) exists and is unique.*

Proof. Since Newton's iteration converges, matrix Δ should exist if and only if D exists. This matrix represents the direction of approaching the embeddable matrix so having a unique direction is equivalent to a unique completion, in other words a unique Cayley-Menger distance matrix. □

The uniqueness of such solution depends on the dimension of matrix E. The number of rows of E becomes $\sum_{l=6}^{N} (N - l + 1)$, the number of degrees of freedom being $N \geq 6$, for $n = 6$. The number of columns of E depends on the number of perturbation matrices. Since we seek a unique solution, the number of these columns should be at most equal to the number of rows. The common case is the number of E's columns to be exactly equal to the number of its rows, analogous to the fact that a linear system with unique solution is typically square. It is possible, of course, to have uniqueness with an overdetermined system.

Remark 4.3 *Let* p *stand for the number of all unknown or unspecified entries in* D. *If the approximate (or incomplete) distance matrix* D *leads to an embeddable Cayley-Menger matrix in* E_3, *which is sufficiently close to* D, *then at least as many as*

$$\max\{p - \sum_{l=6}^{N} (N - l + 1), \ 0\}, \quad N \geq 6,$$

of these entries can be freely perturbed; the other entries are then determined. Moreover, this number bounds from below the dimension (degrees of freedom) of the conformation manifold.

5 Cycloalkanes

The algorithm and the observations in the previous sections make no assumption about the geometry of molecular chains. Here, we consider the case of cycloalkanes since it is a problem of conformational calculations with many strong geometric constraints and furthermore it is well studied; see Section 2. For illustration, we examine molecules with 6 to 8 backbone degrees of freedom (typically carbon atoms).

The cyclohexane has an infinite number of geometrically possible conformations due to its symmetry. Besides two rigid chair conformations, it can assume any conformation in a closed one-dimensional loop manifold; this manifold contains two embedded points corresponding to boat conformations [EM99]. The Cayley-Menger matrix is

$$D = \begin{bmatrix} 0 & b & c & u_1 & c & b & 1 \\ b & 0 & b & c & u_2 & c & 1 \\ c & b & 0 & b & c & u_3 & 1 \\ u_1 & c & b & 0 & b & c & 1 \\ c & u_2 & c & b & 0 & b & 1 \\ b & c & u_3 & c & b & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix},$$

where bond lengths $b = 1.54^2$Å, bond angles remain fixed at $109.47°$ (thus, using the rule of cosines, $c = 2.51^2$Å) and u_1, u_2, u_3 represent unknown values. Once these unspecified entries are determined, we can recover the geometry of the molecule up to global translations and rotations. Thus, the starting point is to identify the symmetric perturbation matrices of D. In this case, the subspace of matrices $\mathcal{R}$ has a basis comprised of perturbations R_{14}, R_{25}, R_{36}. Assume that we already know a conformation of cyclohexane in the one-dimensional manifold. This conformation corresponds to a unique set of u_1, u_2, u_3 values. The values of u_2 and u_3 depend on u_1. Thus, by removing perturbation matrix R_{14} from $\mathcal{R}$, we can find a new *unique* Cayley-Menger matrix for each value of u_1, using our algorithm. In the cyclohexane's case, we used a fixed step value of 0.05Å, to explore the entire conformation manifold. Of course, we can remove some other matrix, in order to obtain a 1-dimensional set of solutions.

Besides computing all conformations on the manifold, our method was applied to enumerate all distinct types of conformations: By altering just dihedral angles it is impossible to pass between the boat and the chair geometries, whereas changing some angles between bonds can do it. We have applied a small perturbation (i.e., in the range of 10%), of interatomic distances in order to destroy the molecule's symmetry and produce a finite number of conformations, thus allowing us a "global" view of conformation space. Our method gives results as good as fully rigorous algebraic methods in that we obtain at most 4 solutions, as in [EM99, GS70, MZW95], where the bond lengths and angles had been perturbed by at most 10%. The 4 isolated conformations correspond to 2 chair and 2 boat conformations, which correspond to the conformations most encountered in nature and hence minimizing energy! The number of solutions upper bounds the number of connected components of the manifold, provided the input is generic (in practice, random). This procedure is not as simple for the cycloheptane and the cyclooctane, because it is harder to guarantee genericity. This shows the limitation of algebraic methods.

Now let us refer to matrix E and vector F of the previous section: Since R_{14} is removed, their dimensions are 3×2 and 3×1 respectively. Thus, the optimization involves the solution of an overdetermined system of linear equations. After 3 iterations, we have that the norm $\|E_k\xi - F_k\|_2 < 10^{-15}$, and this is zero within the precision of 16 decimal digits used. Therefore the algorithm stops.

This is an instance of a *certified answer* in the context of numerical computation.

For the cycloheptane, the Cayley-Menger matrix has 7 unknown entries. The Cayley-Menger matrix is

$$D = \begin{bmatrix} 0 & b & c & u_1 & u_2 & c & b & 1 \\ b & 0 & b & c & u_3 & u_4 & c & 1 \\ c & b & 0 & b & c & u_5 & u_6 & 1 \\ u_1 & c & b & 0 & b & c & u_7 & 1 \\ u_2 & u_3 & c & b & 0 & b & c & 1 \\ c & u_4 & u_5 & c & b & 0 & b & 1 \\ b & c & u_6 & u_7 & c & b & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix},$$

and matrices R_{14}, R_{15}, R_{25}, R_{26}, R_{36}, R_{37}, R_{47} form a basis of $\mathcal{R}$. Starting with a given conformation, we extract one perturbation matrix (e.g. R_{14}, corresponding to u_1) and proceed with the singular value optimization. In this case, E_k and F_k define a 6×6 system. We were able to completely explore each one-dimensional manifold with a step of 0.05Å, obtaining for example, more than 50 valid conformations with u_1 in the range $[8.586, 11.290]$Å. We see that the matrix entry u_1 is constrained by $u_1 < 11.29$Å. While exploring the one-dimensional manifold and after some iterations, if u_1 is increased beyond this bound, then (1) cannot be satisfied. This is a case of incompatible constraints, and matrix E_k becomes singular implying there is no possible vector ξ.

After extracting one more perturbation matrix, we could not obtain any solutions, so the dimension cannot be larger than one. Hence our method confirms what is known about the cycloheptane, i.e., that there are two one-dimensional conformation manifolds [Cri92].

For the cyclooctane, the Cayley-Menger matrix has 12 unknowns:

$$D = \begin{bmatrix} 0 & b & c & u_1 & u_2 & u_3 & c & b & 1 \\ b & 0 & b & c & u_4 & u_5 & u_6 & c & 1 \\ c & b & 0 & b & c & u_7 & u_8 & u_9 & 1 \\ u_1 & c & b & 0 & b & c & u_{10} & u_{11} & 1 \\ u_2 & u_4 & c & b & 0 & b & c & u_{12} & 1 \\ u_3 & u_5 & u_7 & c & b & 0 & b & c & 1 \\ c & u_6 & u_8 & u_{10} & c & b & 0 & b & 1 \\ b & c & u_9 & u_{11} & u_{12} & c & b & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

We are interested in the conformation manifold: A direct consequence of Remark 4.3 bounds the dimension from below by $12 - 10 = 2$. By extracting certain perturbation matrices, just as for the cycloheptane above, we could bound the dimension from above by two. Hence its dimension is 2, confirming what is known [GS70, WS99].

But we have not yet solved completely any perturbed system (as for the cyclohexane) in order to obtain isolated solutions, whose cardinality would bound the number of connected components. This number constitutes an interesting open question today; it is believed [Cri92] that there are two or three connected components, but no proof exists.

6 Computational performance

This section reports on experiments with 6 to 12 degrees of freedom in cycloalkanes, in table 1. We also apply our implementation to molecular chains of up to 20 backbone atoms, not necessarily cycloalkanes; see table 2.

Molecule	dof	Init. eigval	Final eigval	Iterations	Time [sec.]	KFlop
Cyclohexane	6	1.56e-01	3.72e-08	3	0.02	26
Cycloheptane	7	1.45e-01	1.31e-08	3	0.03	38
Cyclooctane	8	1.11e-01	4.65e-07	3	0.05	54
Cyclononane	9	1.24e-01	3.31e-08	3	0.08	80
Cyclodecane	10	1.64e-01	6.86e-07	3	0.12	119
Cycloendecane	11	2.10e-01	1.43e-06	3	0.18	183
Cyclododecane	12	1.32e-01	7.41e-08	5	0.27	281

Table 1: Method's performance for computing one cyclic conformation.

dof	Init. eigval	Final eigval	Iterations	Time [sec.]	KFlop
7	2.98e-02	6.64e-14	3	0.01	36
8	2.57e-02	4.43e-12	3	0.05	49
9	2.10e-02	6.29e-11	3	0.05	73
10	2.38e-02	2.95e-13	3	0.11	109
11	3.16e-02	2.60e-12	3	0.16	165
12	8.13e-02	1.20e-07	3	0.22	282
13	8.09e-02	8.49e-08	3	0.30	450
14	3.72e-02	6.04e-13	3	0.49	606
15	3.53e-02	2.02e-14	3	0.77	940
16	3.78e-02	1.72e-12	3	1.15	1404
17	3.83e-02	1.70e-13	3	1.54	2082
18	3.53e-02	3.93e-13	3	2.14	3039
19	3.80e-02	4.59e-14	3	2.91	4344
20	4.00e-02	7.09e-13	3	3.79	6136

Table 2: Improved implementation for computing one conformation.

Our software is in the public domain, available through the second author's webpage. It is based on MATLAB, version 5.3 [CBG99], or, alternatively, on SCILAB[Gom99]. The advantages of the latter package include its flexibility in code development, and the fact that it is freely distributed and simple to install; the two systems have almost identical syntax. We have used MATLAB to generate `C` code which, when compiled, gives faster timings, reported in table 1 and table 2. We aspire to extend the applicability of our software by implementing it in `C` or `C++`.

In the tables are shown the initial and final values for the 6th singular value. The input is created by perturbing a known conformation, then our code computes the nearest conformation. The initial perturbation is limited; this reveals the local nature of our optimization. Moreover, the 6th singular value is initially smaller than 1, otherwise we are confronted to problems of global optimization. The step size of our experiments is typically 0.05Å.

Both tables give results averaged over 3 runs, computed on a 500MHz PENTIUM-III architecture.

In Figure 1 we present a molecule with 15 degrees of freedom, as computed by our software on MATLAB. Here all bond lengths are equal to 1.5Å, as induced by table 6, which contains the Cartesian coordinates of all backbone atoms, regarded as point masses. The shown conformation

Atom	X	Y	Z
1.	2.9542	-1.4439	0.2325
2.	3.4683	-0.3184	-0.6155
3.	2.6029	0.9049	-0.5484
4.	2.0938	1.1841	0.8346
5.	1.1711	2.3655	0.8891
6.	-0.2595	2.0011	0.6237
7.	-0.5796	1.9223	-0.8397
8.	-2.0068	1.5446	-1.1051
9.	-2.7251	1.0769	0.1259
10.	-3.4377	-0.2276	-0.0748
11.	-2.8818	-1.3326	0.7738
12.	-1.9387	-2.2259	0.0239
13.	-0.6922	-1.5182	-0.4183
14.	0.4935	-1.8305	0.4457
15.	1.7375	-2.1024	-0.3472

Table 3: Atom coordinates for the molecule in Figure 1.

satisfies all bond length constraints, as well as the bond angles constraints.

Behind MATLAB and SCILAB lies the software library LAPACK [ABB+95], of which we heavily use its tridiagonal eigensolver. In particular, the orthogonalization by DSTEIN uses more than 90% of the time to compute the eigenpairs by tridiagonalization. Another bottleneck of our algorithm is dense square linear system solving, which could benefit from specialized software.

7 Further research

Our algorithm uses, in each iteration, a symmetric eigenvalue decomposition and linear system solving. This is a local optimization procedure, whereas the problem is essentially one of global optimization. To give a complete set of conformations, sampling techniques must be applied. We have also experimented with interval analysis in order to exclude regions that contain no conformation. This will also provide candidate regions which are small enough to be explored by our methods. Our preliminary tests applied the interval capabilities of MAPLE, which are rather limited, but also used package ALIAS, which has a large number of functions implemented in `C/C++` [Mer00].

We expect some improvements in accuracy and efficiency if we use the definition of a cluster in the algorithm of [DFP97]. More importantly, we have not exploited the structure of the linear systems. This should accelerate the algorithm, but may also improve accuracy since it reduces the space of perturbations sensible to roundoff error.

Bound smoothing is a standard technique in refining distance intervals obtained indirectly, namely by successively applying the triangle and tetrangle inequalities. Although their optimal use is an important open question, we may still employ this tool to reduce the size of the input intervals [EH89]. Using a local coordinate system offers better constraints propagation.

Acknowledgments

The second author thanks Gordon Crippen for many insightful discussions.

References

[ABB+95] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen. *LAPACK Users' Guide*. SIAM, Philadelphia, 2nd edition, 1995.

[BCDD90] J.M. Blaney, G.M. Crippen, A. Dearing, and J.S. Dixon. *DGEOM*. program #590, Quantum Chemistry Program Exchange, 1990. http://qcpe.chem.indiana.edu.

[BJ95] M. Bakonyi and C.R. Johnson. The euclidean distance matrix completion problem. *SIAM J. Matrix Anal. Appl.*, 16(2):646–654, 1995.

[BKWK+00] C. Bailey-Kellogg, A. Widge, J.J. Kelley, III, M.J. Berardi, J.H. Bushweller, and B.R. Donald. The NOESY jigsaw: Automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *J. Comp. Biol.*, 7:537–558, 2000.

[Blu70] L.M. Blumenthal. *Theory and Applications of Distance Geometry*, volume 15. Chelsea Publishing Company, Bronx, NY, 2 edition, 1970. (1st edition: Cambridge Univ. Press, Cambridge, 1953).

[BP94] D. Bini and V.Y. Pan. *Polynomial and Matrix Computations*, volume 1: Fundamental Algorithms. Birkhäuser, Boston, 1994.

[CBG99] T. Coleman, M.A. Branch, and A. Grace. *Optimization Toolbox for Use with MATLAB: User's Guide*. The Math-Works Inc., 1999.

[CGS89] G. Chang, W.E. Guida, and W.C. Still. An internal coordinate Monte Carlo method for searching conformational space. *J. American Chem. Soc.*, 111:4379–4386, 1989.

[CH88] G.M. Crippen and T.F. Havel. *Distance geometry and Molecular Conformation*. Research Studies Press Ltd, Taunton, Somerset, England, 1988.

[Cou95] National Research Council. *Mathematical Challenges from Theoretical / Computational Chemistry*. National Academy Press, Washington, D.C., 1995. http://www.nap.edu/.

[Cri92] G. M. Crippen. Exploring the conformation space of cycloalkanes by linearized embedding. *J. Comp. Chem.*, 13:351–361, 1992.

[Dem92] J.W. Demmel. The componentwise distance to the nearest singular matrix. *SIAM J. Matr. Anal. Appl.*, 13:10–19, 1992.

[DFP97] I. Dhillon, G. Fann, and B. Parlett. Application of a new algorithm for the symmetric eigenproblem to computational quantum chemistry. In *Proc. 8^{th} SIAM Conf. on Parallel Proces. for Scient. Comp.*, pages 383–389, 1997.

[DH88] A.W.M. Dress and T.F. Havel. Shortest-path problems and molecular conformation. *Discrete Applied Math.*, 19:129–144, 1988.

[EH89] P.L. Easthope and T.F. Havel. Computational experience with an algorithm for tetrangle inequality bound smoothing. *Bull. Math. Biol.*, 51(1):173–194, 1989.

[EM99] I.Z. Emiris and B. Mourrain. Computer algebra methods for studying and computing molecular conformations. *Algorithmica, Special Issue on Algorithms for Computational Biology*, 25:372–402, 1999.

[Exc] Quantum Chemistry Program Exchange. http://qcpe.chem.indiana.edu.

[GHHW90] W. Glunt, T.L. Hayden, S. Hong, and J. Wells. An alternating projection algorithm for computing the nearest euclidean distance matrix. *SIAM J. Matrix Analysis & Appl.*, 11(4):589–600, 1990.

[Gom99] C. Gomez. *Engineering and Scientific Computing with Scilab.* Birkhäuser, Boston, 1999.

[GS70] N. Gō and H.A. Scheraga. Ring closure and local conformational deformations of chain molecules. *Macromolecules*, 3(2):178–187, 1970.

[GV96] G.H. Golub and C.F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, Baltimore, Maryland, 3rd edition, 1996.

[Hav97] T.F. Havel. Computational synthetic geometry with Clifford algebras. In *Automated Deduction in Geometry*, volume 1360 of *LNCS*, pages 102–114. Springer, 1997.

[Hav98] T.F. Havel. Distance geometry: Theory, algorithms, and chemical applications. In *Encyclopedia of Computational Chemistry.* John Wiley & Sons, 1998.

[HKC83] T.F. Havel, I.D. Kuntz, and G.M. Crippen. The theory and practice of distance geometry. *Bull. Math. Biol.*, 45(5):665–720, 1983. Errata, 47(1):157, 1985.

[HN95] T.F. Havel and I. Najfeld. A new system of equations, based on geometric algebra, for ring closure in cyclic molecules. In *Applications of Computer Algebra in Science and Engineering*, pages 243–259. World Scientfic, Singapore, 1995.

[KNB92] J. Kuszewski, M. Nilges, and A.T. Brünger. Sampling and efficiency of metrix matrix distance geometry: A novel partial metrization algorithm. *J. Biomolecular NMR*, 2:33–56, 1992.

[KTO93] I.D. Kuntz, J.F. Thomason, and C.M. Oshiro. Distance geometry. *Methods in Enzymology*, 177:159–204, 1993.

[LEE96] S. LeGrand, A. Elofsson, and D. Eisenberg. The effect of distance-cutoff on the performance of the distance matrix error when used as a potential function to drive conformational search. In H. Bohr and S. Brunak, editors, *Protein Folds: A Distance Based Approach*, pages 105–113. CRC Press, Inc., 1996.

[Mat85] R. Mathar. The best euclidean fit to a given distance matrix in prescribed dimension. *Linear Algebra Appl.*, 67:1–6, 1985.

[MBD01] T.E. Malliavin, P. Barthe, and M.A. Delsuc. FIRE: Predicting the spatial proximity of protein residues from a 3D HSQC-NOESY. *Theor. Chem. Accts*, 106:91–97, 2001.

[MD02] T. Malliavin and F. Dardel. Structure des protéines par rmn. In *Sciences Fondamentales*, volume AF, pages 6608 (1–18). Techniques de l'Ingénieur, Paris, January 2002.

[Mer00] J-P. Merlet. ALIAS: an interval analysis based library for solving and analyzing systems of equations. In *Systèmes d'Equations Algébriques*, Toulouse, 2000.

[MW99] Jorge Mor and Zhijun Wu. Distance geometry optimization for protein structures. *J. Global Optimization*, 15:219–234, 1999. Software on: www-unix.mcs.anl.gov/ more/dgsol.

[MZW95] D. Manocha, Y. Zhu, and W. Wright. Conformational analysis of molecular chains using nano-kinematics. *Computer Applications of Biological Sciences*, 11(1):71–86, 1995.

[PL97] P.M. Pardalos and X. Lin. A tabu based pattern search method for the distance geometry problem. In F. Giannessi, Sándor Komlósi, and Tamás Rapcsák, editors, *New Trends in Mathematical Programming.* Kluwer, 1997.

[Sax79] J. B. Saxe. Embeddability of weighted graphs in k-space is strongly NP-hard. In *Proc. 17th Allerton Conf. on Communications, Control and Computing*, pages 480–489, 1979.

[SS] Conformational Searching and Analysis Software. www.netsci.org/resources/software/modeling/conf.

[SS90] G.W. Stewart and J. Sun. *Matrix Perturbation Theory.* Academic Press, Boston, 1990.

[WD95] M. A. Wicks and R. A. Decarlo. Computing most nearly rank-reducing structured matrix perturbations. *SIAM J. Matrix Anal. and Appl.*, 16(1):123–137, 1995.

[WS99] W.J. Wedemeyer and H.A. Scheraga. Exact analytical loop closure in proteins using polynomial equations. *J. Comput. Chem.*, 20(8):819–844, 1999.

[XOW+00] Y.Z. Xu, Q. Ouyang, J.G. Wu, J.A. Yorke, G.X. Xu, D.F. Xu, R.D. Soloway, and J.Q. Ren. Using fractals to solve the multiple minima problem in molecular mechanics calculation. *J. Comput. Chem.*, 21:1101–1108, 2000.