JOURNAL OF COMPUTATIONAL PHYSICS 26, 449–452 (1978)

Note

Rapid Calculation of Coordinates from Distance Matrices

1. INTRODUCTION

One of the major difficulties in the calculation of conformation by the "distance geometry" approach [1] has been that of producing a set of three dimensional Cartesian coordinates $\mathbf{v}_i$, $i = 1,..., n$ for n points, given an $n \times n$ matrix of upper bounds $\mathbf{U}$ and lower bounds $\mathbf{L}$ on the interpoint distances:

$$l_{ij} \leqslant d_{ij} = \| \mathbf{v}_i - \mathbf{v}_j \| \leqslant u_{ij}, \qquad \text{for} \quad i, j = 1,..., n. \tag{1}$$

We are not concerned here with the determination of $\mathbf{U}$ and $\mathbf{L}$, but rather with the calculation of any set of $\mathbf{v}$'s satisfying (1), assuming that a solution exists. A proposed distance matrix $\mathbf{D}$ chosen at random such that $l_{ij} \leqslant d_{ij} \leqslant u_{ij}$ for all i and j, does not correspond in general to any set of $\mathbf{v}$'s in the three dimensional Euclidean space because the corresponding $(n + 1) \times (n + 1)$ bordered matrix of squared distances, $\mathbf{C}$, is in general not of rank 5 (see [1] for proof). Here $c_{ij} = d_{ij}^2$ for $i, j = 1,..., n$ and $c_{i,n+1} = c_{n+1,i} = 1$ for $i = 1,..., n$ and $c_{n+1,n+1} = 0$. Furthermore, generating the $\mathbf{v}$'s analytically according to (6) of [1] directly from $\mathbf{D}$ (or $\mathbf{C}$) results in increasingly larger violations of (1) for larger i and j. Direct alteration of $\mathbf{C}$ (and hence $\mathbf{D}$) in order to reduce its rank to 5, permits the straightforward calculation of $\mathbf{v}$'s from the modified $\mathbf{D}$, but the modification algorithm in [1] (involving Gaussian elimination and direct alteration of individual matrix elements) required computer time proportional to n^3 and memory space proportional to n^2.

One can devise a more efficient method, time and memory requirements depending only linearly on n, by resolving the real, symmetric matrix $\mathbf{C}$ into its eigenvalues and eigenvectors:

$$\mathbf{C} = \mathbf{E}\Lambda\mathbf{E}^T \tag{2}$$

where $\mathbf{E} = (e_{ij})$ is the column matrix of eigenvectors and $\Lambda = (\lambda_i)$ is the diagonal matrix of corresponding eigenvalues. The λ's are real and may be ordered so that $|\lambda_1| > |\lambda_2| > \cdots$; the eigenvectors are real and may be taken as orthonormal [2]. The condition that $\mathbf{C}$ have rank 5 is equivalent to requiring $|\lambda_5| > \lambda_6 = \lambda_7 = \cdots = \lambda_{n+1} = 0$. Thus we may calculate a modified $\mathbf{C}$ matrix, denoted by $\mathbf{C}'$, from (2) after setting the sixth and subsequent eigenvalues to zero:

$$c'_{ij} = \sum_{k=1}^{5} e_{ik}\lambda_k e_{jk} \qquad \text{for} \quad i, j = 1,..., n + 1 \tag{3}$$

0021-9991/78/0263–0449$02.00/0

Then $\mathbf{C}'$ is the rank 5 matrix closest to $\mathbf{C}$ in the spectral sense, but some elements may violate (1), some diagonal elements may be non-zero, and some border elements may not be equal to unity. We have tried a number of iterative schemes to correct these discrepancies, but in our hands they either converged very slowly or not at all. In the examples we have tried, $\mathbf{C}$ is well-conditioned with respect to its eigenvalues but ill-conditioned with respect to its eigenvectors. By numerically minimizing the above deviations using as variables the 5 eigenvalues of largest modulus and the components of their corresponding eigenvectors, one can avoid the instability difficulties, but the constraints are awkward to handle. There are [5 variable eigenvalues] + [$5(n+1)$ variable components of the five eigenvectors] − [$(n+1)$ constraints of zero diagonal elements] − [n constraints of unity border elements] − [15 constraints from the orthonormality of the five eigenvectors] = $3n - 6$ net degrees of freedom, which is what one would expect, since we are seeking the x, y and z coordinates of n points, up to a rigid translation and rotation (and mirror inversion). C' is automatically symmetric, as can be seen from (3).

2. Algorithm

Our most successful method for large n consists of: (i) choosing $\mathbf{D}$ at random, bounded element-wise by $\mathbf{U}$ and $\mathbf{L}$ as described; (ii) calculate the first 5 eigenvalues and eigenvectors of the corresponding $\mathbf{C}$ matrix; (iii) calculate $\mathbf{C}'$ from (3); (iv) calculate the $\mathbf{v}$'s from the first four rows of $\mathbf{C}'$ according to (6) of Ref. [1]; and (v) minimize $f(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ with respect to the $3n$ Cartesian coordinates

$$f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \sum_{j>i} \begin{cases} (d_{ij}^2 - u_{ij}^2)^2, & d_{ij} > u_{ij} \\ 0, & l_{ij} \leqslant d_{ij} \leqslant u_{ij} \\ (d_{ij}^2 - l_{ij}^2)^2, & d_{ij} < l_{ij} \end{cases} \tag{4}$$

where $d_{ij} = \| \mathbf{v}_i - \mathbf{v}_j \|$, the ordinary Euclidean norm. Step (ii) is conveniently done by the method of "exhaustion," where the eigenvalue of largest absolute value, λ_1, and its corresponding eigenvector e_1 are found by iterating for $k = 1, 2, \ldots$

$$\frac{(\mathbf{C}'^k\mathbf{y}) \cdot (\mathbf{C}'^k\mathbf{y})}{(\mathbf{C}'^{k-1}\mathbf{y}) \cdot (\mathbf{C}'^k\mathbf{y})} \to \lambda_1, \qquad \text{as} \quad k \to \infty \tag{5}$$

and $\mathbf{C}'^k\mathbf{y} \to e_1$, for arbitrary $\mathbf{y}$.

Then λ_2 is found similarly by using $\mathbf{C}' - \lambda_1 \mathbf{e}_1 \mathbf{e}_1^T$ in the place of $\mathbf{C}'$ in (5), and so on [3]. Convergence of the λ's to six significant figures is regularly achieved well before $k = 100$, even when two eigenvalues are close in magnitude. The computational effort involved in the matrix-vector multiplication goes up as n^2, but since only the first 5 eigenvalues are sought, the cost otherwise depends linearly on n. Step (iv) is remarkably stable. For instance, in our calculations of conformations of the small protein, bovine pancreatic trypsin inhibitor (BPTI), $n = 58$, $d_{i,i+1} = 3.8$ Å, and 6.0 Å $\leqslant d_{5,55}, d_{14,38}, d_{30,51} \leqslant 6.5$ Å. In other words, U and L enforce a chain of 57 fixed length steps with three short crosslinks spanning nearly the whole chain.

Other distances are much less constrained. Calculation of the v's from **C** would result in $d_{5,55} \sim 100$ and so on, whereas using **C'** we regularly achieve crosslink distances between 5 and 10. Thus step (iv) initially positions the points in an overall correct fashion, although with still some substantial errors. Step (v) is a nonlinear least squares minimization of the function f, which has zero value at any solution, strictly positive values elsewhere, and continuous first derivatives everywhere. The gradient can easily be calculated analytically, so we use initially steepest descent minimization until $\| \nabla f \|^2 < 10^5$ Å^6 and then Fletcher–Reeves conjugate gradient minimization [4] until $\| \nabla f \|^2 < 10^3$ Å^6, resulting in $f < 1$ Å^4 and a largest term in (4) of 0.01 Å^4 or less. Typical initial values of f are about 10^{10} Å^4. Both minimization algorithms have storage requirements proportional to n, so large numbers of variables are easily handled. As a test of the method, we chose $u_{ij}^2 = 1.05 g_{ij}^2$ and $l_{ij}^2 = 0.95 g_{ij}^2$, where the g_{ij} are the distances between C^α atoms of BPTI taken from the x-ray crystallographic coordinates [5]. On the Lawrence Berkeley Laboratory's CDC 7600, each conformation generated according to the algorithm above took an average of 2.7 seconds. By way of comparison, minimizing from arbitrary starting points with the same **U** and **L** takes about 5 times as long [6]. The performance with looser upper and lower bounds is similar. The starting configuration in step (iv) is close enough to the minimal region of f that convergence to spurious local minima rarely occurs, in our experience. This is in sharp contrast to the difficulties experienced when using arbitrary starting points [6].

The method also performs well when the upper and lower bound matrices are much less restrictive. When **U** and **L** correspond to only fixed i to $i + 1$ distances and the three crosslinks of BPTI (plus their logical consequences regarding all other distances, as deduced from the triangle inequality [1]) then only 1 second of computer time per structure is required. Convergence is about as rapid and precise as with tight bounds, and we have encountered no spurious local minima. Of course there is a great deal of conformational variability among the resultant structures since the bounds are so loose.

3. Summary

In conclusion, we have devised and tested a practical algorithm for rapidly solving (1) in the difficult case of large n and numerous strong constraints. Computer time increases only quadratically and memory requirements increase only linearly with n, and there is little difficulty with multiple minima. We hope this will make the distance geometry approach to conformational calculation more feasible for large, highly constrained systems.

Acknowledgments

This work was supported by a grant from the Academic Senate of the University of California. The author thanks Drs. I. D. Kuntz, K. Miller, and B. N. Parlett for their stimulating and helpful discussions.

REFERENCES

1. G. M. CRIPPEN, *J. Computational Phys.* **24** (1977), 96.
2. J. H. WILKINSON, "The Algebraic Eigenvalue Problem," Oxford Univ. Press, Oxford, 1965.
3. D. K. FADDEEV AND V. N. FADDEEVA, "Computational Methods of Linear Algebra," pp. 307, 328–330, Freeman, San Francisco, 1963.
4. R. FLETCHER AND C. M. REEVES, *Comput. J.* **7** (1964), 149.
5. R. HUBER *et al.*, *Cold Spring Harbor Symp. Quant. Biol.* **36** (1971), 141.
6. I. D. KUNTZ, G. M. CRIPPEN, AND P. A. KOLLMAN, submitted for publication (1978).

RECEIVED: May 3, 1977

G. M. CRIPPEN

University of California
School of Pharmacy,
San Francisco, California 94143