# Use of Augmented Lagrangians in the Calculation of Molecular Conformations by Distance Geometry

G. M. CRIPPEN*

College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109

A. S. SMELLIE

Physical Chemistry Laboratory, South Parks Road, Oxford University, Oxford OX1 3QR, England

JEFFREY W. PENG

Biophysics Research Division, University of Michigan, Ann Arbor, Michigan 48109

Received December 30, 1985

Distance geometry is a technique widely used to find atomic coordinates that agree with given upper and lower bounds on the interatomic distances. It is successful because it chooses at random some relatively good "trial coordinates" that take into account the whole molecule and all constraints at once. Customarily, these trial coordinates must be refined by minimizing a penalty function until the structure agrees with the original bounds. Here we present an alternative to minimizing the penalty function, which has the advantage of more precisely satisfying the bounds, showing more clearly when the bounds are mutually contradictory, and simultaneously optimizing an objective function subject to precise satisfaction of the bounds.

## INTRODUCTION

The general distance geometry embedding problem is to determine atomic coordinates of a molecule such that a given set of bounds on (some of) the interatomic distances are satisfied.[1] The constraints may be derived from theory (e.g., standard bond lengths and angles) and/or from experiment (e.g., two-dimensional NMR studies), and a particular interatomic distance $d_{ij}$ may have an upper bound $u_{ij}$, a lower bound $l_{ij}$, neither, or both. In the last case, there may be a considerable range between the lower and the upper bound, or they may be equal, so that the distance is constrained to a fixed value.

$$l_{ij} \leq d_{ij} \leq u_{ij} \tag{1}$$

The standard "embedding" algorithm for solving this problem, such as DISGEO by T. F. Havel distributed by QCPE, deduces as much as it can about all interatomic distances from the given (incomplete) list of constraints and then produces a series of random sets of "trial coordinates" that agree in an overall sense with the allowed distances but violate some of the original constraints. Each set of trial coordinates is finally refined by local unconstrained minimization of a penalty function such as eq 2, where $E$ is continuous and differentiable for nonzero

$$E = \sum_{i<j} \max\left[0, \left(\frac{d_{ij}^2}{u_{ij}^2} - 1\right)\right] + \max\left[0, \left(\frac{l_{ij}^2}{d_{ij}^2} - 1\right)\right] \tag{2}$$

distances and $E \geq 0$ everywhere, with $E = 0$ only when all constraints are satisfied. In practice, although the minimization rapidly and reliably converges to rather low values of $E$, there is sometimes some lingering doubt whether the final small $E > 0$ represents a residual that can be reduced by further minimization of eq 2 or whether the constraints are mutually slightly incompatible near this conformation. The second problem is that although a remarkably wide variety of physical constraints can be expressed in terms of interatomic distance bounds,[1,2] some can be expressed only as optimization problems. Obviously, if one had some experimental constraints on the conformation of a molecule and low-energy conformations were desired, that is such a constrained optimization problem. As another example, suppose one wanted to know the most extended conformation of a molecule subject to given geometric constraints.

In this paper we introduce the use of augmented Lagrangians (AL) to solve the two classes of problem described above. The case of ascertaining complete agreement with geometric constraints is illustrated with the thorny example of [D-Pen$^2$,D-Pen$^5$]enkephalin (DPDPE), a conformationally restricted cyclic pentapeptide subject to a stringent set of bond length, bond angle, and steric constraints in addition to a preliminary set of interatomic distance constraints derived from NMR NOE experiments. The second sort of problem, constrained optimization, is illustrated by finding the most extended conformation of DPDPE subject to all of the above constraints. Not only do we obtain correct structures, but one can detect which constraints are restraining the objective function and which play no direct role, at least in the resultant conformation.

## METHOD

The augmented Lagrangians method we use comes straight from Bertsekas.[3] Let $N$ be the number of atoms, and let $\mathbf{X}$ be the $3N$-dimensional vector of $x,y,z$ coordinates of the atoms. Let $B(\mathbf{X})$ be the given (nonlinear) objective function that must be minimized, subject to $m$ (nonlinear) equality constraints of the form $h_i(\mathbf{X}) = 0$, $r$ single-sided inequality constraints $g_i(\mathbf{X}) < 0$, and $s$ double-sided inequality constraints $\alpha_i < f_i(\mathbf{X}) < \beta_i$. The various $h_i$, $g_i$, and $f_i$ may be in general any functions of the coordinates, but we have employed

$$h_i(\mathbf{X}) = d_{kj}^2 - u_{kj}^2 = 0 \tag{3}$$

when $u_{kj} = l_{kj}$ for some pair of atoms $k$ and $j$,

$$g_i(\mathbf{X}) = \begin{cases} d_{kj}^2 - u_{kj}^2 < 0 \\ l_{kj}^2 - d_{kj}^2 < 0 \end{cases} \tag{4}$$

when only one bound is given, and

$$l_{kj}^2 = \alpha_i < d_{kj}^2 = f_i(\mathbf{X}) < \beta_i = u_{kj}^2 \tag{5}$$

when both bounds are given. In addition to pairwise distance constraints, we have included chiral constraints. Let $\mathbf{a}_i$ denote the $x,y,z$ coordinates of atom $i$. Then for four atoms, $i$, $j$, $k$, and $l$, having fixed relative positions

$$(\mathbf{a}_j - \mathbf{a}_i)\cdot[(\mathbf{a}_k - \mathbf{a}_i) \times (\mathbf{a}_l - \mathbf{a}_i)] - \chi_{ijkl} = h(\mathbf{X}) = 0 \tag{6}$$

constrains their chirality, where $\chi_{ijkl}$ is the desired value of the vector triple product as calculated from some standard structure. If the four atoms are to be coplanar, $\chi = 0$, and otherwise a mirror reflection of the quartet just reverses the sign of $\chi$. Alternatively, we have also incorporated chiral inequalities where only the sign is specified, since the magnitude of $\chi$ depends on the distances among the four atoms, which may not be known in advance.

Finally, the augmented Lagrangian is defined as

$$L(\mathbf{X},\lambda,\mu,\kappa) = B(\mathbf{X}) + \sum_{i=1}^{m}\lambda_i h_i(\mathbf{X}) + (c/2)\sum_{i=1}^{m}[h_i(\mathbf{X})]^2 + \sum_{i=1}^{r}\mu_i g_i'(\mathbf{X},\mu_i,c) + (c/2)\sum_{i=1}^{r}[g_i'(\mathbf{X},\mu_i,c)]^2 + \sum_{i=1}^{s}\rho_i(f_i(\mathbf{X}),\kappa_i,c) \tag{7}$$

where

$$g_i'(\mathbf{X},\mu_i,c) = \max\left\{g_i(\mathbf{X}), -\frac{\mu_i}{c}\right\} \tag{8}$$

and

$$\begin{aligned} \rho_i(f_i(\mathbf{X}),\kappa_i,c) &= \kappa_i[f_i(\mathbf{X}) - \beta_i] + (c/2)[f_i(\mathbf{X}) - \beta_i]^2 \quad \text{if } \kappa_i + c[f_i(\mathbf{X}) - \beta_i] > 0 \\ \rho_i(f_i(\mathbf{X}),\kappa_i,c) &= \kappa_i[f_i(\mathbf{X}) - \alpha_i] + (c/2)[f_i(\mathbf{X}) - \alpha_i]^2 \quad \text{if } \kappa_i + c[f_i(\mathbf{X}) - \alpha_i] < 0 \\ \rho_i(f_i(\mathbf{X}),\kappa_i,c) &= -\kappa_i/2c \qquad \text{otherwise} \end{aligned} \tag{9}$$

The $\lambda$, $\mu$, and $\kappa$ are vectors of Lagrange multipliers, and $c$ is a weighting factor. The usual method of Lagrange multipliers would set to zero all terms involving $c$ and minimize $L$ with respect to $\mathbf{X}$, $\lambda$, $\mu$, and $\kappa$. In practice, this is frequently numerically unstable. Alternatively, the penalty function approach would neglect all terms involving the multipliers and would minimize $L$ with respect to $\mathbf{X}$ for increasingly large positive values of $c$. The difficulty in this is that until $c \rightarrow \infty$, the constraints are not well satisfied, but the minimization generally converges rather slowly when $c$ is large. Augmented Lagrangians combine the best of these two approaches by starting with coordinates $\mathbf{X}^{(0)}$, and then at "outer loop" iteration $k$, one simply minimizes $L$ with respect to $\mathbf{X}$ only, resulting in $\mathbf{X}^{(k+1)}$, and subsequently updates the multipliers and $c$ by

$$\lambda_i^{(k+1)} = \lambda_i^{(k)} + c^{(k)}h_i(\mathbf{X}^{(k+1)}) \tag{10}$$

$$\mu_i^{(k+1)} = \mu_i^{(k)} + c^{(k)}g_i'(\mathbf{X}^{(k+1)})$$

$$\kappa_i^{(k+1)} = \begin{cases} \kappa_i^{(k)} + c^{(k)}[f_i(\mathbf{X}^{(k+1)}) - \beta_i] & \text{if } \kappa^{(k)} + c^{(k)}[f_i(\mathbf{X}^{(k+1)}) - \beta_i] > 0 \\ \kappa_i^{(k)} + c^{(k)}[f_i(\mathbf{X}^{(k+1)}) - \alpha_i] & \text{if } \kappa^{(k)} + c^{(k)}[f_i(\mathbf{X}^{(k+1)}) - \alpha_i] < 0 \\ 0 & \text{otherwise} \end{cases}$$

$$c^{(k+1)} = \begin{cases} \gamma c^{(k)} & \text{if } \epsilon^{(k)} > \delta\epsilon^{(k-1)} \\ c^{(k)} & \text{otherwise} \end{cases}$$

where we have used $\gamma = 1.5$, $\delta = 0.25$, $c^{(0)} = 0.01$, and zero initial values of all multipliers. The function $\epsilon$ is simply a measure of how well all the constraints are satisfied and can be taken to be all the terms of $L$ not involving the multipliers.

The "inner loop", where $L$ is minimized, was carried out by repeatedly cycling through the list of atoms, performing a Newton minimization step with respect to the three coordinates of the single atom, until the gradient components of $L$ with respect to all $\mathbf{X}$ were adequately small. Accuracy was assured by carrying out all first and second derivative calculations analytically in double precision.

Our general experience with AL is that it is robust but probably intrinsically slower than the penalty function minimization customarily used in embedding because each iteration of the outer loop consists of a minimization of $L$. Our preliminary program, which is not optimized for speed, requires tens of hours of VAX 11/750 time for the pentapeptide examples given. The inner loop usually amounts to only a single pass through all the atoms, because the multipliers have not changed that much since the last iteration, but good progress depends on accurate minimization (hence the use of Newton's method). The outer loop updating of the multipliers and $c$ tends to "jostle" the sequence $\mathbf{X}^{(0)}$, $\mathbf{X}^{(1)}$, ..., so that overall convergence is not very smooth. The advantage of that is there is a chance of escaping minor local minima by the updating. However, AL is not a panacea for avoiding local minima in the penalty function. Whenever $E$ converges to a minimal value strictly greater than zero, AL generally will also fail to progress, cycling endlessly through ever greater values of $c$ and the multipliers. AL can be proven to converge well from a rather broad region of starting conformations if the feasible region is convex, but for conformational constraints this need not be true. Consider, for example, atoms 1, 2, and 3 in one dimension, subject to the distance equality constraints $d_{12} = 1$, $d_{13} = 2$, and $d_{23} = 1$. If initially $x_1^{(0)} = 0$, $x_2^{(0)} = 2$, and $x_3^{(0)} = 1$, then one can show analytically that in the first iteration $x_2^{(1)} - x_1^{(1)} = {}^5/_3$ and $x_3^{(1)} - x_1^{(1)} = {}^4/_3$, and for all subsequent iterations the conformation remains unchanged as $c,\lambda \rightarrow \infty$.

## RESULTS ON DPDPE, CONSTRAINTS ONLY

[D-Pen$^2$,D-Pen$^5$]enkephalin has the sequence Tyr-D-Pen-Gly-Phe-D-Pen, where the two penicillamine residues are linked by a disulfide bridge. Since it is a highly $\delta$ opioid receptor selective peptide, its conformation in solution is currently under investigation in the laboratories of Mosberg and Woodard. We represented the molecule as 57 atoms, excluding the hydrogens not bonded to the $\alpha$ carbons. Preliminary NMR studies[4] have so far yielded 20 high-confidence, nontrivial NOE interactions, allowing us to set an upper bound on the corresponding distances to 5.0 Å. The experimental work is currently being refined, so it is unimportant exactly what the constraints are and what the resulting allowed conformations might be. The aim of this paper is to simply illustrate how AL works on a realistic and demanding test case. In addition to the NOE constraints, we required standard bond lengths and angles (including sulfur–sulfur distances compatible with the bridge), planar peptide linkages, flat aromatic rings, and a conformation devoid of steric interference between the side chains. The last demand therefore represents a set of lower bound distance constraints. Altogether, we had on the order of 2000 constraints, consisting mostly of steric constraints between almost all atom pairs. Nearly all of these constraints are inactive at a feasible conformation, but which are essential is not immediately clear. Thus, we began with no steric constraints, applied AL starting from a conformation found by distance geometry embedding, examined the final conformation for van der Waals contacts, and added the required lower bound constraints. The final constraint list involved 191 equality constraints, 185 lower bound constraints, 64 double-sided constraints, and 52 chiral equality constraints. Active lower bound constraints removed the steric interference be-

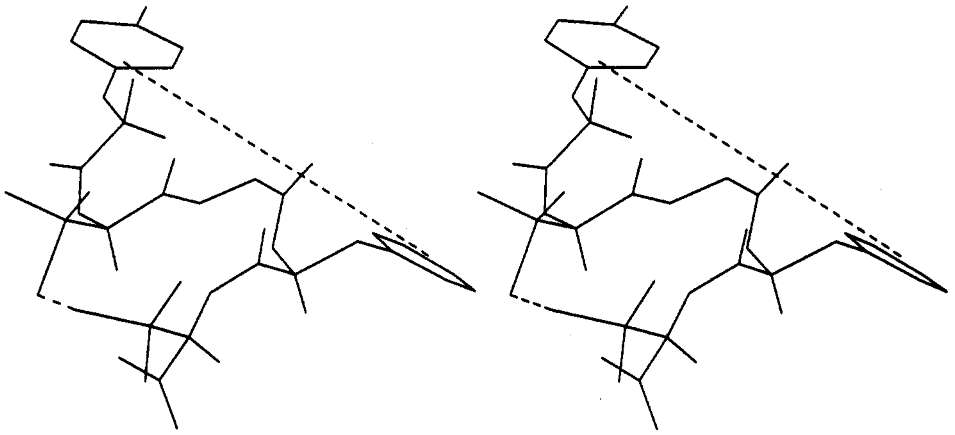

**Figure 1.** Stereoview of a conformation of DPDPE satisfying all a priori chemical structural constraints and NOE distance bounds. Long dashed line indicates the separation of the Tyr and Phe aromatic rings; the short dashed line is the disulfide bridge.

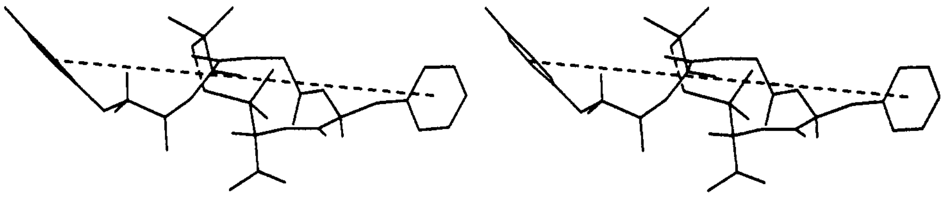

**Figure 2.** Stereoview of DPDPE satisfying all geometric constraints and in addition maximizing the separation between the aromatic rings.

tween residues Tyr[1] and Pen[2] as well as between Pen[2] and Pen[5]. The final structure is shown in Figure 1. The structure begins with Tyr[1] in the upper left corner. The ring centers are approximately 8.39 Å apart, and the long dashed line between virtual atoms near the ring centers depicts this separation. We found that AL yields a structure that complies well with all the constraints. In particular, the peptide linkages are cis–trans–trans–cis and have torsion angles deviating from planarity by less than 5°. Additionally, both of the aromatic rings are reasonably flat. The sulfur–sulfur distance of 2.11 Å, illustrated with the short dashed line, agrees well with the required 2.04 Å. The advantage of AL is that we can satisfy all the constraints very precisely and check that all the multipliers converge to zero in the end (largest value was 0.003). This indicates that there are no mutually incompatible constraints tugging against each other (or against the objective function, if there were one), even slightly. By the standard procedure of simply minimizing $E$, it is not at all clear without AL whether the minimization was not quite complete or whether there are mutually contradictory constraints active, at least near the current conformation.

## RESULTS ON DPDPE, CONSTRAINTS AND OPTIMIZATION

Doubtless, one of the major applications of AL will be to start with geometrically correct conformations of a molecule produced by embedding and then minimize the internal energy subject to maintaining those geometric constraints. We have not yet implemented such a computer program but instead have examined some geometric constrained optimization problems that also lie outside the scope of standard embedding. Suppose we want to find conformations of DPDPE that satisfy all the geometric constraints as before but in addition seek to maximize the distance between the two aromatic ring centers. The final coordinates in the previous section have a 10.32-Å separation; we attempted to extend this to 16 Å. Thus, we start at the same initial conformation as in the constraints-only case and minimize the function

$$B(\mathrm{X}) = [d^2_{\mathrm{QR_Y,QR_F}} - (16\ \text{Å})^2]^2 \tag{11}$$

where the QR's are the virtual atoms at the centers of the Tyr and Phe rings. A constrained optimization is even lengthier than the purely geometric case. Since smaller constraint lists speed the program, we once again included lower bound constraints as needed (a somewhat different set, due to reaching a different conformation). The final constraint list for the optimized structure involved 191 equality constraints, 104 lower bound constraints, 34 double-sided constraints, and 52 chiral equality constraints. Not surprisingly, the critical steric constraints were between the two pencillamine residues. The result is illustrated in Figure 2. By comparison with Figure 1, we see that the molecule has been opened up, and the aromatic rings are substantially stretched apart to a separation of 14.79 Å, indicated by the long dashed line. The two sulfurs are separated by 1.92 Å, making a shorter, but still reasonable, disulfide bridge. Additionally, planarity of the aromatic rings and peptide linkages (cis–cis–trans–trans) has been maintained as in the previous case. An exception is the first peptide bond, which has a torsion angle of 14° as well as a C–N–$C^\alpha$ bond angle of 162°. These violations of the constraints result from slow convergence and conflict between the objective function eq 11 and the constraints. All the Lagrange multipliers were larger in this case than in the unoptimized structure, ranging between 0.01 and 200. Comparing the multipliers associated with the various constraints in the optimized structure, one can easily assemble a stress map of the molecule. In particular, the multipliers are nearly an order of magnitude larger for the constraints maintaining the disulfide bridge, distances for atoms bonded to the sulfurs, the $C^\alpha$, $C^\beta$, and ring centers of the aromatic residues, and the $C^\alpha$, C, N, and O atoms of the first peptide bond. It is clear from comparing the two figures that these constraints are in opposition to increasing the ring separation since the increase was achieved mainly by movement of Tyr[1] rather than Phe[4]. On the other hand, the Lagrange multipliers for the last peptide linkage are not so great, as it is not being pulled apart. This is also evident in the illustration.

## CONCLUSIONS

The augmented Lagrangian approach is useful as a postprocessor to the standard distance geometry embedding programs in order to differentiate between conformations that have not been fully refined with respect to the constraints and mutually contradictory sets of constraints. With an energy function, one can produce conformations that have minimal energy while truly satisfying all the constraints. The customary approach of adding constraint penalty terms to an energy function merely produces a compromise between partially satisfied constraints and partially optimal energy. With various other sorts of objective functions, AL can be used to more efficiently explore the range of allowed conformations. One can answer questions such as "What is the closest approach possible between these two atoms?", "How far can they be separated?", "How nearly trans can this bond become?", or "How dissimilar can I make this conformation compared to a standard structure?". In addition to handling such optimizations subject to very general geometric equality and inequality constraints, AL gives valuable insight concerning the interactions among constraints and between the constraints and the objective at the final conformation. We believe this last feature will be very important.

## ACKNOWLEDGMENT

This work is based on some earlier efforts by R. Nambudripad and Y.-S. Sung. We acknowledge funding support by NSF Grant DMB-8705006 and NIH Grant GM37123.

## REFERENCES AND NOTES

(1) Crippen, G. M. *Distance Geometry and Conformational Calculations*; Chemometrics Research Studies Series 1, Research Studies (Wiley): New York, 1981.

(2) Crippen, G. M.; Havel, T. F. *Distance Geometry and Molecular Conformation*; Research Studies (Wiley): Chichester, England, 1988.

(3) Bertsekas, D. P. *Constrained Optimization and Lagrange Multiplier Methods*; Academic: New York, 1982; pp 20, 96–156.

(4) Mosberg, H. I.; Subramanian, P.; Sobczyk, K.; Crippen, G. M.; Ramalingam, K.; Woodard, R. W. *Combined Use of Stereospecific Deuteration, NMR, and Distance Geometry for Conformational Analysis of* [D-*Pen*$^2$,D-*Pen*$^5$]*Enkephalin*; presented at the 10th American Peptide Symposium, St. Louis, MO, May 1987.

# Canonical Numbering and Coding of Imaginary Transition Structures. A Novel Approach to the Linear Coding of Individual Organic Reactions

SHINSAKU FUJITA

Research Laboratories, Ashigara, Fuji Photo Film Co., Ltd., Minami-Ashigara, Kanagawa, Japan 250-01

Received October 9, 1987

The canonical numbering and coding of an imaginary transition structure (ITS) are described. The nodes of an ITS are partitioned partially into (pseudo)equivalent classes in light of four kinds of extended connectivities. Each of the nodes of the highest class is selected as a root, to which possible spanning trees are constructed. The nominated sets of canonical numbering are obtained from the respective spanning trees. Then the canonical code is obtained by comparing newly defined lists based on the sets of numbering. The concept of a reduced ITS is proposed. The canonical numbering and coding of the reduced ITS are also discussed.

In previous papers,[1] we presented the concept of imaginary transition structures (ITS's) for the description of organic reactions. The ITS of a given reaction is a structure that has out- (—//—), in- (—o—), and par-bonds (—) in accord with structural changes during the reaction. This formulation provides an explicit method for describing an *individual organic reaction*.[1j] The ITS contains 15 kinds of imaginary bonds, each of which is a combination of the out-, in-, and/or par-bonds.[1a] Then the ITS is stored and manipulated in terms of an ITS connection table, in which the imaginary bonds are represented by complex bond numbers. In order to construct an effective computer system, the canonical numbering of the nodes (vertices) of the ITS and the canonical coding of the ITS are remaining problems to be solved.

Many methods were reported for canonizing organic compounds (molecular graphs).[2] One of the most familiar methods is Morgan's procedure, in which (1) the nodes of a molecular graph are partially partitioned by the iterative calculation of extended connectivities and then (2) numbered after the formation of a spanning tree rooted to each of the uppermost nodes, and finally (3) the best name is selected by comparison between nominated names.[3]

The present paper deals with the canonical numbering and coding of ITS's. The resulting canonical names of ITS's (CANITS) are the first unambiguous codes for the description of *individual organic reactions*. This method is an extension of Morgan's procedure, in which (1) four kinds of extended connectivities are introduced to partition the nodes of an ITS partially, and (2) the selection of the best name is based upon a newly defined linear code. In addition, we propose the concept of reduced imaginary transition structures and their canonical coding.

## PARTIAL PARTITIONING BY MEANS OF FOUR KINDS OF EXTENDED CONNECTIVITIES

The ITS of a given reaction contains (1) *intra*string hydrogen atoms (hydrogen reaction centers), (2) implicit or explicit *extra*string hydrogen atoms (hydrogen atoms other than reaction centers), and (3) non-hydrogen atoms.[4] Among them, we consider (1) and (3) for the present coding of the ITS unless the description of stereochemistry requires the consideration of (2).

In the present method, four kinds of extended connectivities, EC1($i$), EC2($i$), EC3($i$), and EC4($i$), are computed and assigned to each node $i$ of a given ITS. Then the nodes of the ITS are partitioned into (pseudo)equivalent classes in light of these extended connectivities.[5] Figure 1 shows the flow chart of the partial partitioning of the nodes to be examined.

**Step 1.** The initial values of the extended connectivities are calculated for each node $i$ as follows: EC1($i$), the number of neighboring reaction centers (hydrogen and non-hydrogen atoms) that are linked to the current node ($i$) by in- or out-bonds; EC2($i$), the number of neighboring atoms (except extrastring hydrogen atoms) attached to the node $i$ with any kind

0095-2338/88/1628-0128$01.50/0 © 1988 American Chemical Society