JOURNAL OF COMPUTATIONAL PHYSICS 18, 224–231 (1975)

Global Optimization and Polypeptide Conformation

GORDON M. CRIPPEN*

Cardiovascular Research Institute, University of California, San Francisco, California 94143

Received January 13, 1975

The problem of calculating the conformation of a molecule by global minimization of its free energy is precisely formulated. Various bounds and estimates are derived for the number of energy evaluations necessary to perform the task, independent of the search algorithm used. The algorithm by Schubert is cited as optimal for the problem as formulated, and an improvement for starting it is presented. In light of the estimates for computer time and memory for the optimal method, ab initio global minimization is proven to be infeasible for calculating conformations of even oligopeptides.

INTRODUCTION

Numerous attempts have been made to calculate the conformation of a molecule, particularly proteins and polypeptides, by minimizing its estimated free energy as a function of conformation (see for instance [1]). This approach is sound in physical theory since at equilibrium at constant temperature and pressure, the Gibbs free energy is minimal. The system is assumed to be a dilute solution so that the energy is that of a single molecule (interacting with solvent in a sophisticated analysis) and depends upon conformation, which is customarily specified by a vector of n dihedral angles $\mathbf{x} = (x_1, ..., x_n)$ corresponding to rotation about single bonds. The difficulty is that the energy function thus calculated for all possible conformations is a very complicated, multimodal function of the dihedral angles. Hence ordinary local minimization algorithms are not appropriate [2]. A number of global methods have been proposed [3–6], but they are of limited effectiveness, and lack of a clear understanding of the problem they were to solve has clouded the issue.

From the point of view of statistical mechanics, at equilibrium the ensemble of solute molecules will be distributed throughout *all* conformation space with a density dependent upon the energy according to the Boltzmann distribution. Consequently the desired conformation is actually all regions of conformation space having energies within a few kT of the global minimum of energy. All such

* Present address: 134 Arlington Rd., Montgomery, Alabama 36105.

conformations are physically significant, and others are not. In order to make the task nontrivial as well as practical, the answers are desired with a minimum amount of computational effort.

Before the problem can be more precisely formulated, we need to examine the energy function $f(\mathbf{x})$ more closely. Since each x_i is a rotation (in, say, degrees), the domain I of f is not Euclidean space, but a toroidal manifold because $0° = 360°$; hence I is finite. For any $\mathbf{x}$, $f(\mathbf{x})$ may be computed, and such evaluations are taken to be the time limiting step. The most practical way is to approximate the energy with f as a semiempirical function [7] of the form

$$f = \sum_{i \neq j} g_{ij}(r_{ij}), \tag{1}$$

where the sum extends over all atom pairs i and j separated by a distance r_{ij}. The functions g_{ij} are minimal for, say, $r_{ij} = r_{ij}^*$. Therefore, $\sum_{i \neq j} g_{ij}(r_{ij}^*)$ is a lower bound for f. However $g_{ij} \to +\infty$ as $r_{ij} \to 0$, [7] so that there can be points or even subspaces of conformation space where $f = \infty$, and therefore f is not bounded from above. In general, $\partial f/\partial x_i$, $i = 1,\ldots, n$ is not globally bounded. Now f can be globally bounded if it is arbitrarily truncated above a certain value $y_{\max}$ chosen such that conformations of energy greater than $y_{\max}$ are physically unimportant:

$$f(\mathbf{x}) = \min \Big(\sum_{i \neq j} g_{ij}(r_{ij}), y_{\max}\Big). \tag{2}$$

Formulation

Now the global minimization task can be formulated generally and precisely. Let Φ_L be the set of *all* functions f such that f is single-valued and real, defined on a finite domain I as a function of n variables $\mathbf{x} \in I$. Further, let there be defined a metric $d(\mathbf{x}_a, \mathbf{x}_b)$ for all $\mathbf{x}_a, \mathbf{x}_b \in I$, and let f obey a Lipschitz condition

$$|f(\mathbf{x}_a) - f(\mathbf{x}_b)| \leqslant L \cdot d(\mathbf{x}_a, \mathbf{x}_b) \tag{3}$$

for some known a priori Lipschitz constant L for all $\mathbf{x}_a, \mathbf{x}_b \in I$. Assume no other conditions on Φ_L. Denote by $y^* = \min_{\mathbf{x} \in I}\{y = f(\mathbf{x})\}$ the global minimum of the function, and by Δy the chosen accuracy of function value desired (Δy comparable to kT). Then the objective is to determine $D = \{\mathbf{x} \mid f(\mathbf{x}) \leqslant y^* + \Delta y\}$ in the least number of function evaluations possible.

Remark. The above are the weakest conditions on Φ_L consistent with the objective. If I were infinite, an infinite number of evaluations would be required. If there were no metric, the Lipschitz condition would be undefined. Customarily

$d(\cdot)$ is taken to be the ordinary Euclidean distance, or more useful for a discrete I,

$$d(\mathbf{x}_a, \mathbf{x}_b) = \sum_{i=1}^{n} | x_{ai} - x_{bi} | \tag{4}$$

as will be used later. If there were no Lipschitz condition and I is continuous, an infinite number of evaluations would be necessary; for I discrete, the objective is trivial and necessitates evaluating every point in I.

For continuous I it is often useful to discretize it by choosing a step size Δx in each x_i, $i = 1,..., n$ according to the Lipschitz condition (3) given the a priori constants L and Δy:

$$\Delta x = \Delta y / L. \tag{5}$$

Then $D = \{\mathbf{x} = (k_1 \Delta x,..., k_n \Delta x) | f(\mathbf{x}) \leqslant y^* + \Delta y\}$, where the k_i are integers. The above choice of Δx is the largest permissible, since otherwise some regions which belong to D would be missed, and all regions are required. Also Δx must not be taken smaller, since then more evaluations will be made in locating essentially the same regions belonging to D, and the task was to be done in the least number of evaluations. Working with a discrete I is not only easier, but the convergence to stationary points implied by a continuous I is not physically realistic, i.e., the molecules are scattered about conformation space to some extent at temperatures greater than absolute zero. Hence only the case of discrete I will be considered in the remainder of this paper.

Bounds and Estimates

By the very nature of the objective as formulated in the previous section, there is an intrinsic range to the number of evaluations required for a global minimization. In order to derive this, we must first define "essential point."

Definition. $\mathbf{x}_a$ is essential iff there exists no $\mathbf{x}_b \in I$ such that

$$f(\mathbf{x}_b) - L \cdot d(\mathbf{x}_a, \mathbf{x}_b) > y^*.$$

That is, the evaluation of any other point plus even precise knowledge of y^* is insufficient to show that $\mathbf{x}_a$ is not a member of set D.

Theorem 1. *The number of evaluations N necessary to determine D for some $f \in \Phi_L$ has the range $2 \leqslant N \leqslant M$, where M is the number of points in I. The result is independent of search algorithm.*

Proof. (By exhibition) Upper bound: suppose $y^* + \Delta y > f(\mathbf{x}) \geqslant y^*$ for all $\mathbf{x} \in I$. Then all $\mathbf{x}$ are essential and hence $N = M$. Of course $N > M$ is impossible. Lower bound: suppose $\mathbf{x}_a$ and $\mathbf{x}_b$ are at least as distant from each other as any other pair of points in I and that

$$f(\mathbf{x}_a) - f(\mathbf{x}_b) = L \cdot d(\mathbf{x}_a, \mathbf{x}_b). \tag{6}$$

Then $\mathbf{x}_b$ is essential and $f(\mathbf{x}_b) = y^*$, so that $\mathbf{x}_b$ is the only member of D. Any other point $\mathbf{x}_c$ is not essential, and $f(\mathbf{x}_c) \geqslant f(\mathbf{x}_a) - L \cdot d(\mathbf{x}_a, \mathbf{x}_c) > f(\mathbf{x}_a) - L \cdot d(\mathbf{x}_a, \mathbf{x}_b) = f(\mathbf{x}_b) = y^*$ by Eqs. (3) and (6). Thus D is determined by evaluating $\mathbf{x}_a$ and $\mathbf{x}_b$, and therefore $N = 2$. If $M > 1$ (the case $M = 1$ is trivial), then $N = 1$ is impossible since there is nothing to compare the one function value with. Q.E.D.

It should be observed that for conformational calculations, $M = (360°/\Delta x)^n$ which can be large for large n.

The expected number of evaluations $E(N)$ can be estimated from basic information theory [8] after some simplifying assumptions. Namely, the information theory entropy of the location of the global minima, H_{gm}, is estimated as well as that of an individual evaluation, H_{ev}. Then assuming all the information contained in an evaluation can be applied to the global search problem, and assuming each evaluation to be a statistically independent experiment, then

$$E(N) = H_{gm}/H_{ev}. \tag{7}$$

Now H_{gm} may be estimated by assuming the probabilities of the various $\mathbf{x}$'s being in D are mutually independent, and then either assuming there is only one point in D (corresponding to the smallest value of H_{gm}) and hence

$$H_{gm} = M \cdot \left(\frac{-1}{M} \log \frac{1}{M} - \frac{M-1}{M} \log \frac{M-1}{M}\right), \tag{8}$$

or assuming all points have probability $\frac{1}{2}$ of being in D (maximal value of H_{gm}) and hence,

$$H_{gm} = M \log 2. \tag{9}$$

For an estimate of H_{ev}, assume that the probability that $y = f(\mathbf{x})$ is independent of adjacent function values and uniformly distributed over the range 0 to R. Then the largest value of H_{ev} is

$$H_{ev} = \log(R/\Delta y), \tag{10}$$

where maximum detail is useful; the most pessimistic value is

$$H_{ev} = \log 2. \tag{11}$$

Then combining (7), (9), and (11) for the most pessimistic estimate, $E(N) = M$, which is the upper bound of Theorem 1. Using (9) and (10) for medium pessimism, $E(N) = M \log 2/\log (R/\Delta y)$.

Algorithms

We are now prepared to discuss search algorithms to solve the global minimization problem, especially the optimal one for our purposes, but first a few words about classification and comparison of methods. With the recent proliferation of "optimal" algorithms (e.g. [9–13]), it must first be pointed out that an algorithm is best in comparison to a class of other methods. The two favorite classes are *sequential,* where the choice of $\mathbf{x}$ to be evaluated depends on the results of the previous evaluations, and *nonsequential,* where the $\mathbf{x}$'s to be evaluated are all chosen in advance. Inasmuch as nonsequential methods can be considered sequential methods that make no use of the intermediate information, only sequential methods will be considered here, although Sukharev [10] has shown that sequential is no better than nonsequential under a certain assessment of performance. The performance of an algorithm must be judged on all the members of a class of functions, in our case Φ_L. There are two common means of assessing the performance of an algorithm on a particular function: (a) for a predetermined N measure the error in the determination of D and/or y^* (preferred method for nonsequential algorithms); or (b) the approach used in our case, where one measures the N necessary to reduce the inaccuracy in determination of y^* and/or D to a predetermined level (here, large N corresponds to a large error in, say, y^* in approach (a)). Now, an algorithm is said to be optimal if it performs better when applied to an entire class of functions than any other of a class of algorithms. The traditional measure is minimax optimality [9], or the "best guaranteed result," where the maximal performance error over the class of functions is minimal for the optimal algorithm compared to that of the other algorithms. In our case, with its natural assessment (b), by Theorem 1 the maximum error always corresponds to $N = M$. Thus minimax optimality is trivial for our purposes. A more useful view is expected optimality, where an algorithm is optimal if the expected value of N, $E(N)$, taken over all Φ_L is minimal with respect to the class of algorithms. This corresponds to minimizing computational effort for solving large numbers of different problems.

The Optimal Algorithm

For our purposes, the continuous and discrete algorithm by Shubert [12, 13] is optimal. The reader is referred to his papers for a complete explanation of the method, but the following simple description will do for the moment: 1. First evaluate the function at an arbitrary point; 2. Calculate according to the known points and the Lipschitz condition (3) which point has the lowest *possible* value; 3. Evaluate at that point and return to step 2 until the lowest possible function value is no less than the lowest known value(s).

The continuous version, which is somewhat impractical to apply for $n > 1$, has been proved minimax optimal among sequential algorithms with assessment (a) with respect to accuracy in estimating y^* and D. Its convergence as measured by error in estimating y^* is $O(1/N)$ [12].

The discrete version has been proved to require smaller N than that of any other sequential algorithm with the same arbitrary starting point for *each* $f \in \Phi_L$ [13]. This is stronger than expected optimality. Clearly, Shubert's is the method of choice for conformational calculations which are formulated as in this paper.

Shubert claims the choice of starting point is arbitrary, but always picks a corner ($n > 1$) or an endpoint ($n = 1$) in examples. The following theorem shows how his algorithm can be improved by choosing the corner for starting.

THEOREM 2. *For I a discrete hypercube of n dimensions ($n \geqslant 1$), the optimal starting point for global search is a corner, $\mathbf{x}_0$. It is assumed that over the class of functions Φ_L, the likelihood $p(y)$ that $f(\mathbf{x}) = y$ is independent of $\mathbf{x}$.*

Proof. We need only demonstrate that $\mathbf{x}_0$ has the maximum probability over all $\mathbf{x} \in I$ of being essential. Then making the initial evaluation at $\mathbf{x}_0$ is the most efficient choice, since that point has maximal probability of needing to be evaluated regardless of the outcome of the other function evaluations. But from the definition we see that $\mathbf{x}_i$ being essential depends upon $f(\mathbf{x}_j)$ and the distance $d(\mathbf{x}_i, \mathbf{x}_j)$ for all $\mathbf{x}_j \in I, i \neq j$. Since as assumed, $p(y)$ is the same for all $\mathbf{x}_j$, the only difference among the $\mathbf{x}_j$ is that those with larger $d(\mathbf{x}_i, \mathbf{x}_j)$ are less likely to disqualify $\mathbf{x}_i$ from being essential, as is clear from the definition. When $\mathbf{x}_i = \mathbf{x}_0$, the d's are largest and hence $\mathbf{x}_0$ is most likely to be essential. Q.E.D.

Note. Since I for molecular conformations is an n-dimensional torus, all points are equal by the above criterion.

It is instructive to compare performance of Shubert's algorithm with the information theory estimates. He gives the example [13] of searching out the global minima of 500 functions f defined by $f(0) = 0$, $f(k) = \sum^k w$, $k = 1,..., 100$, where w is a random integer variable uniformly distributed over the interval $[-9, +9]$. Taking $L = 10$, he found an average of 27.89 evaluations were needed per function. Assuming $k = 50$ to be representative of the average behavior of each entire curve, by the central limit theorem $f(k)$ is approximately a normally distributed random variable with $\mu = 0$ and $\sigma = 36.7$. If $\Delta y = 10$, then noting $P(|f(k)| \leqslant 70) = 0.94$, we divide the range -70 to 70 into intervals i of length

10 and calculate the probability p_i that $f(k)$ lies within that interval. Then the estimated entropy of an average evaluation is $\sum_i - p_i \log_{10} p_i = 1.0635$ dits. Either $H_{gm} = 2.432$ dits as given by Eq. (8) (unique minimum) and $E(N) = 2.29$, or $H_{gm} = 30.1$ dits according to (9) (maximum entropy) and $E(N) = 28.30$, which is fortunately close to the observed 27.89. If, as seems reasonable, the global minimum is usually unique, the algorithm averaged per evaluation 0.0872 dits of information applied to the task, compared to the estimated 1.06 dits available.

The application of even this optimal algorithm to sizeable polypeptides is rather pointless, as we shall see. As a matter of experience, $\Delta x \lesssim 20°$, which means that for n variable dihedral angles $M \approx (360/20)^n$. Since the algorithm requires M words of storage for convenience, or at least N (which can approach M, by Theorem 1) even with heroic programming, then even for $n = 4$, $M =$ 100000 words, and $n > 4$ is out of the question. Unfortunately even for small proteins $n \gtrsim 200$. Furthermore, since $N \propto M$, some 10^{250} evaluations of the energy might be required for a small protein. A practical application would be for instance locating the lowest minima of the 2-variable dipeptide, glycylalanine [3]. As explained in the Introduction, this can be approached by taking a cutoff $y_{\max} = 15$ kcal/mole and $\Delta x = 20°$, and then from the energy map (see [3]), one must take $L \geqslant 7$ kcal/20°. About 100 evaluations are required, which is better than any other method tried on the same problem.

Conclusion

Straightforward prediction of conformation by energy minimization is infeasible for more than four variables, when only the minimal amount of information on the energy function is known. Only by establishing additional conditions on the energy function, on an either mathematical or physical basis, and then by employing algorithms which make use of these additional features, can conformational analysis be rigorously performed.

References

1. H. A. Scheraga, *Chem. Revs.* **71** (1971), 195.
2. D. J. Wilde and C. S. Beightler, "Foundations of Optimization," Prentice-Hall, Englewood Cliffs, New Jersey, 1967.
3. G. M. Crippen and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.* **64** (1969), 42.
4. G. M. Crippen and H. A. Scheraga, *Arch. Biochem. and Biophys.* **144** (1971), 453.
5. G. M. Crippen and H. A. Scheraga, *Arch. Biochem. and Biophys.* **144** (1971), 462.
6. G. M. Crippen and H. A. Scheraga, *J. Computational Phys.* **12** (1973), 491.
7. F. A. Momany, L. M. Carruthers, R. F. McGuire, and H. A. Scheraga, *J. Phys. Chem.* **78** (1974), 1595.

8. C. E. Shannon and W. Weaver, "The Mathematical Theory of Communication," University of Illinois Press, Urbana, Illinois, 1949.
9. J. Kiefer, *Proc. Amer. Math. Soc.* **4** (1951), 503.
10. A. G. Sukharev, *U.S.S.R. Comp. Math. and Math. Phys.* **11** (1971), 119.
11. J. H. Beamer and D. J. Wilde, *Management Science* **16** (1970), 529.
12. B. O. Shubert, *SIAM J. Numer. Anal.* **9** (1972), 379.
13. B. O. Shubert, *Management Science* **18** (1972), 687.

Printed in Belgium