Poling: Promoting Conformational Variation

ANDREW SMELLIE,* STEVEN L. TEIG, and PETER TOWBIN
Molecular Simulations, Inc., 555 Oakmead Parkway, Sunnyvale, California 94086

Received 25 January 1994; accepted 9 May 1994

ABSTRACT

This article introduces several methods of assessing the extent to which a collection of conformations represents or covers conformational space. It also describes poling: a novel technique for promoting conformational variation that can be applied to any method of conformational analysis that locally minimizes a penalty or energy function. The function being minimized is modified to force similar conformers away from each other. The method is independent of the origin of the initial conformers and of the particular minimization method used. It is found that, with the modification of the penalty function, clustering of the resulting conformers is generally unnecessary because the conformers are forced to be dissimilar. The functional form of the poling function is presented, and the merits are discussed with reference to (1) efficacy at promoting variation and (2) perturbation of the unmodified function. Results will be presented using conformers obtained from distance geometry with and without poling. It will be shown that the addition of poling eliminates much redundancy in conformer generation and improves the coverage of the conformational space.

Introduction

Conformational analysis typically attempts to represent the conformational space of a molecule with a finite collection of (low-energy) point conformations. Several approaches have been proposed to generate collections of such conformations. To assess these diverse methods and to compare different methods, we first propose an algorithm-independent measure of conformational coverage: the degree to which a set of conformers represents conformational space. Based on such a metric, we next propose poling: a general extension to conformational analysis algorithms that explicitly promotes conformational variation. Finally, we provide experimental evidence that demonstrates that poling succeeds in improving conformational coverage.

The conformational analysis algorithms proposed to date fall into two broad categories:

*Author to whom all correspondence should be addressed.

CCC 0192-8651 / 95 / 020171-17

1. Quasi-exhaustive search of the conformational space, with subsequent clustering of the generated conformers.
2. Random sampling of the conformational space, with clustering of the sampled conformers.

QUASI-EXHAUSTIVE CONFORMATIONAL SEARCH

The general approach of these algorithms is to quantize the conformational space and to search this quantized space systematically. Pioneering algorithms of this type are those of Lipton and Still[1] and Dammkoehler et al.[2] Both of these methods use the rigid rotor approximation, in which the only degrees of freedom considered are discretized rotations about rotatable bonds. It is primarily because of this approximation that the methods are quasi-exhaustive, because not all conformational states are accessible. This approximation can be partially lifted if some kind of relaxation is permitted for each candidate conformer, at the expense of longer run times. In the method of Lipton and Still, the space is quantized, and an exhaustive tree search is performed over all rotational states. The method of Dammkoehler et al. also quantizes the space but uses an efficient bump check to eliminate rotational states that would induce a van der Waals (VDW) clash. Dammkoehler's systematic search methods are employed in the context of constrained search but also contribute good technology to the systematic search problem. It is well known that both methods can produce thousands of conformers for molecules with relatively few rotatable bonds. Indeed, these methods are often cited as evidence that one needs "millions of conformers to cover conformational space."[16]

CONFORMATIONAL SEARCH BY SAMPLING

The general approach with these methods is to sample the conformational space for a period of time with the assumption that the space will be completely sampled as the time tends to infinity. There are many different sampling methods, including distance geometry,[3] stochastic search in Cartesian space,[4] stochastic search in torsion space,[5] molecular dynamics,[6] and Monte Carlo methods.[7]

Common to each of these methods is a concept of similarity between conformations as a way of getting a small number of conformations to cover the conformational space, but the use of such a similarity metric varies from method to method. One of the most common uses of a similarity metric is in clustering. A difference metric [such as root mean square (rms) difference in interatomic distances] is defined between conformers i and j, and this metric is used in a standard clustering algorithm[8] to get unique families. Then one conformer is chosen from each family and the collection of these representative conformers is taken as the sample covering the conformational space.

Promoting Conformational Variation

There are several problems with conformational analysis techniques that rely on generating large numbers of conformations and then postprocessing these conformations (e.g., by clustering) to cover the conformational space. The most significant problem is that of the coverage of space itself. No current method guarantees the complete coverage of conformational space. Quasi-exhaustive search methods are limited to points on a search grid, and sampling methods have some element of change or a systematic error that could miss some important conformations. A second problem is that of conformational redundancy. It is common to generate and, subsequently, to cluster many hundreds or thousands of candidate conformations to get different conformational families. Conformational variation between members of the same family is, by definition, small, and conformers within the same family are redundant. The problem is magnified if there is an implied insistence that each conformer lies at a local minimum on some potential surface. In this case, many candidate conformers that sample the same potential well can, after local minimization, result in the same conformation being duplicated many times. Indeed, this is used as a convergence criterion in some algorithms.[4] Yet if there is a broad shallow basin of attraction on the potential surface, conformers that are structurally far from a local minimum but are energetically near one will be missed. This is illustrated in Figure 1, in which three candidate conformers that are used to sample the potential well converge to the same conformer in an energy-minimized space. For this reason, *no method that limits itself to an exploration of local minima on a potential surface will adequately cover conformational space.*

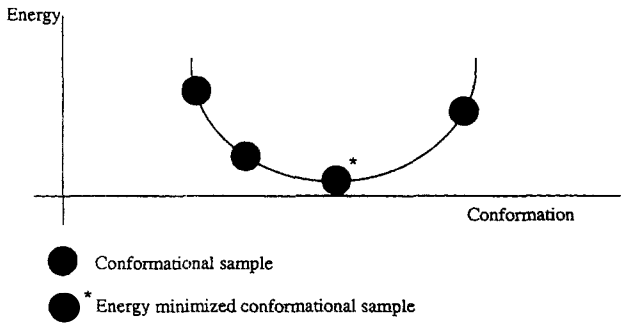


FIGURE 1. Conformational sampling and energy minimization.

Measuring Conformational Coverage

How might conformational coverage be assessed? It is necessary for any conformational analysis algorithm to have some method of assessing the quality of conformational coverage. Such measures of coverage can be divided into one-set and two-set measures. Historically, conformational coverage has been assessed in terms of a one-set measure, such as rms difference between conformations generated for a given molecule. These are measures of absolute conformational coverage and as such are flawed because it is generally not known *a priori* what the perfect coverage of conformational space is. This is illustrated by the one-dimensional example in Figure 2.

In Figure 2 we attempt to describe a one-dimensional space from $L \leq d \leq U$ by a finite number of conformers (five in this case). Suppose a conformer i was good if it differed from all other conformers j by more than D, where $D \ll (U - L)$. In this case, the conformational coverage in case I would be perfectly acceptable, even though it is obviously a suboptimal coverage. The optimal

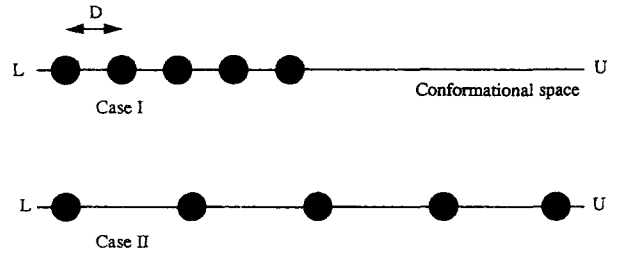


FIGURE 2. Covering a one-dimensional space with point conformers.

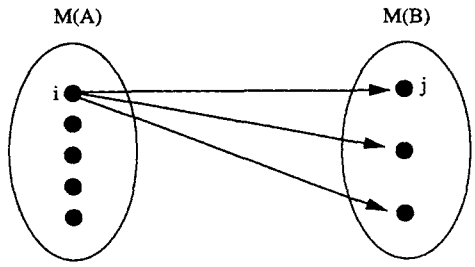


FIGURE 3. The hole-size two-set measure.

coverage is shown in case II, where the conformations are dispersed evenly over the space. So any conformational coverage metric that depends on the absolute difference between conformations can be misled because the flexibility of molecules varies greatly, and the size of the conformational space being explored is not known in advance.

Another popular one-set measure is derived from clustering.[8] Generally, clustering algorithms operate on a distance matrix $\mathbf{D}(i, j)$ that measures the distance (or difference) between conformers i and j. Several clustering parameters are required, such as the maximum intrafamily difference permitted during the clustering (which measures the difference within a single family) and the maximum number of families required. Clustering has the same basic flaw as any rms difference metric: The absolute size of the space is not known in advance, so it is difficult to choose sensible values for the required parameters.

Although it is not possible to measure the absolute coverage of conformational space with a one-set measure, the relative coverage of space can be measured with two-set measures. Two such measures are introduced here: the so-called hole-size measure and the distance bin measure.

The hole-size measure is illustrated in Figure 3. $M(A)$ and $M(B)$ are conformational models (i.e., collections of conformations) of molecule M from conformational analysis methods A and B. Let $d(M(A)_i, M(B)_j)$ be the distance between conformers i [of $M(A)$] and j [of $M(B)$] and let $H(M(A)_i, M(B)) = \min(d(M(A)_i, M(B)))$. The quantity $H(M(A)_i, M(B))$ is measuring the hole that the ith conformer of $M(A)$ finds in the entire conformational model $M(B)$. It is the minimum of distances from $M(A)_i$ to all conformers j in $M(B)$. Thus a large value of $H(M(A)_i, M(B))$ means that the ith conformer from $M(A)$ has found a large hole in the conformational model of $M(B)$. Con-

versely, a large value for $H(M(B)_i, M(A))$ means that the ith conformer from $M(B)$ has found a large hole in the conformational model $M(A)$.

One final quantity remains to be defined—namely, the largest hole that the conformational model A finds in the conformational model B for a given molecule M, denoted by $H(M(A), M(B))$. This is simply the maximum of all $H(M(A)_i, M(B))$, for every conformer i from conformational model A, and represents the largest hole that (any conformer of) A finds in the space of B.

The distance bin measure is illustrated in Figure 4. A set of interesting distances is defined topologically for the molecule. A discussion of which distances to use appears later in this article. Each distance is divided into bins of a predetermined size (say, 0.5 Å), and a bin is marked *touched* if there is a conformer in the set of interest that has the appropriate binning distance. This procedure is repeated for each distance D_i in conformer sets $M(A)$ and $M(B)$. Conformer set $M(A)$ can be compared against $M(B)$ (and vice versa) by computing the following properties:

$DB(M(A)_i, \sim M(B)_i)$ = number of extra bins touched in the ith distance by all the conformers in set $M(A)$ and not in set $M(B)$.

$DB(M(A), \sim M(B))$ = mean number of extra distance bins touched by all the conformers in set $M(A)$ and not in set $M(B)$. In Figure 4, $DB(M(A), \sim M(B)) = 2$.

$DB(M(B)_i, \sim M(A)_i)$ = number of extra bins touched in the ith distance by all the conformers in set $M(B)$ and not in set $M(A)$.

$DB(M(B), \sim M(A))$ = mean number of extra distance bins touched by all the conformers in set $M(B)$ and not in set $M(A)$. In Figure 4, $DB(M(B), \sim M(A)) = 3$.

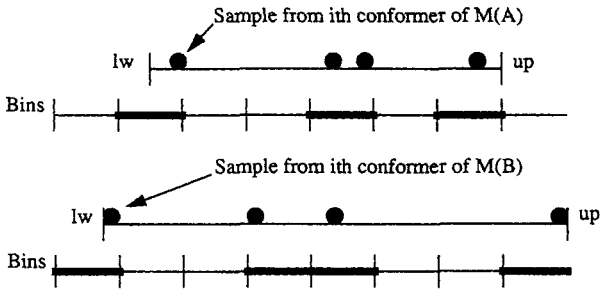


FIGURE 4. The distance bin two-set measure.

Thus, set $M(A)$ can be said to cover more space (in terms of distance bins) if $DB(M(A), \sim M(B)) > DB(M(B), \sim M(A))$. If this condition is true, then conformer set A has touched more extra bins in the space of conformer set B than vice versa. Set $M(A)$ can be said to cover set $M(B)$ completely if $DB(M(A), \sim M(B)) = 0$ and $DB(M(B), \sim M(A)) = 0$.

A two-set measure can be used to compare two conformational analysis methods. To use such a two-set measure to assess a conformational analysis method A on molecule M, denoted by $M(A)$, it is necessary to use a reference method of conformational analysis that is quasi-exhaustive. This is denoted by M(ref). Two examples of quasi-exhaustive methods against which to compare are systematic search (SS) and repetitive use of a sampling algorithm, such as distance geometry (DG). Thus, the hole-size metrics described earlier can be used to measure the largest hole that the quasi-exhaustive method finds in the test method ($H(M(\text{ref}), M(A))$) or the mean number of extra distance bins that the quasi-exhaustive method finds in the test method ($DB(M(\text{ref}), M(A))$).

Poling: Promoting Conformational Variation

What are the goals of promoting conformational variation? A technique was sought that possessed the following properties:

1. It should complement many existing conformational analysis methods in promoting conformational variation.
2. It should be applicable to a broad range of conformer generation methods.
3. It should require minimal parameterization or tuning.
4. It should be relatively efficient.

It was observed that many conformational methods require some form of geometry optimization, so a penalty function was added to this optimization that penalizes any conformer that is too close to another conformer in the set. We call this function a poling function, drawing on the notion of poles in systems science and complex analysis:[14] points at which an otherwise well-behaved (e.g.,

analytic) function goes to infinity. The generalized functional form of the poling function used is given in eqs. (1) and (2).

$$F_{\text{pole}} = W_{\text{pole}} \sum_i \frac{1}{(D_i)^N} \tag{1}$$

$$D_i = \left(\frac{\left(\sum_{j=1} (d_j - d_{ij})^2 \right)}{N_d} \right)^{1/2} \tag{2}$$

Here D_i = rms difference between the poling distances of the current conformation being poled (d_j) and the corresponding distances in the ith previous conformation (d_{ij}), computed over all N_d poling distances. The exact nature of these poling distances will be discussed subsequently. The steepness of the poles is controlled by raising D_i to an arbitrary positive power N. What effect does eq. (1) have on the overall function being minimized? In this article, eq. (1) is added to a regular molecular mechanics force field, and minimization is performed on the entire system. This is summarized by eqs. (3) and (4).

$$\text{Minimize}(F_{\text{total}}) \tag{3}$$

where

$$F_{\text{total}} = F_{\text{bond}} + F_{\text{angle}} + F_{\text{dihed}} + F_{\text{oopl}} + F_{\text{linbend}} + F_{\text{VDW}} + F_{\text{pole}} \tag{4}$$

The poling function has the effect of modifying the potential surface being minimized. As illustrated in the one-dimensional example in Figure 2, the addition of a poling term has the effect of penalizing the conformational space surrounding any conformer already in the set. The closer the conformer being minimized comes to any other conformer in the set, the higher the poling function.

Consider the one-dimensional potential surface in Figure 5 depicted by a graph of energy (E) versus an arbitrary coordinate (R). There are two local minima shown. If conventional conformational analysis techniques were used to cover this space with finite conformers, only two unique conformers would be found: one at each minimum. Introducing a poling term into a final minimization step ensures that the space is more thoroughly covered. In the example shown, the conformer generated in the first iteration (C_1) will be located

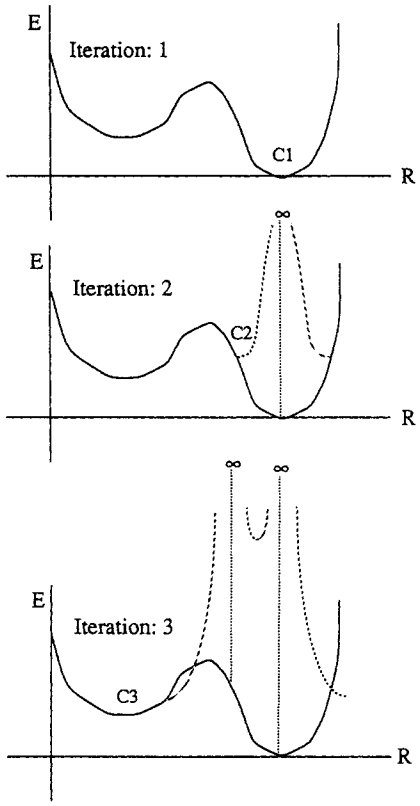


FIGURE 5. The effect of poling on conformational spaces.

at one of the minima (because there is nothing to pole against). At iteration 2, a second conformer (C_2) is generated, but there is a pole at the place where conformer 1 is located. This pole is symmetrical and tends to infinity as the difference between the conformer being generated (C_2) and the previous conformer (C_1) tends to zero. In Figure 5, this second conformer is located at a local minimum on the modified potential surface. At the start of iteration 3, one of the local minima is almost completely covered by conformers 1 and 2, so conformer 3 (C_3) is generated in the remaining local minimum.

We have considered three possible paradigms for the use of a poling function: sequential poling, single simultaneous poling, and multiple simultaneous poling. Sequential poling is summarized in algorithm 1, which is outlined in pseudo C code. Conformations are generated sequentially, by any suitable algorithm, and the current conformer (i.e., the one that has just been generated) is minimized with respect to all previously generated conformations.

Algorithm 1: *Sequential Poling*

```
i = 0
finished = FALSE
while (not finished) {
    generate conformer i
    if (i > 0) {
        minimize conformer i, poling against
          conformers 0 through (i − 1)
    }
    i = i + 1
    if (sufficient conformers) finished = TRUE
}
```

Single simultaneous poling is identical to sequential poling, with the exception that after generating a conformer i and minimizing it with respect to the poles of all other conformers (0 through $[i - 1]$), a previous conformer j is randomly selected and minimized with respect to the poles of all other conformers. This is algorithm 2, and it has the advantage of permitting a conformer that is trapped by surrounding conformers, in the sense that it is surrounded by poles in the conformational space, to escape. In this scheme, a conformer once generated and minimized) is not locked into a conformation. It has the potential to minimize away from a crowded region of conformational space (pushed by any neighboring poles).

Algorithm 2: *Single Simultaneous Poling*

```
i = 0
finished = FALSE
while (not finished) {
    generate conformer i
    if (i > 0) {
        minimize conformer i, poling against
          conformers 0 through (i − 1)
        select a random conformer 0 ⇐ j < i
        minimize conformer j, poling against
          all other conformers
    }
    i = i + 1
    if (sufficient conformers) finished = TRUE
}
```

Multiple simultaneous poling involves the minimization of several conformers simultaneously. Instead of generating a conformer i and minimizing it with respect to all previous conformers, several conformers are generated and minimized simultaneously. This permits several conformers to be poled simultaneously, potentially permitting them to pole against other conformers and to cover different regions of conformational space. Algorithm 3 describes multiple simultaneous poling.

Algorithm 3: *Multiple Simultaneous Poling*

```
i = 0
N = Number of conformers to be generated
      simultaneously
finished = FALSE
while (not finished) {
    generate conformers i to i + N − 1
    if (i > 0) {
        minimize conformers i to i + N − 1,
          poling against all other conformers
    }
    i = i + N
    if (sufficient conformers) finished = TRUE
}
```

Implementation of the Poling Function

The poling term, from eq. (1), is simply an extra term in the minimization of a regular molecular mechanics force field. It is intended to promote conformational variation in any poling distances that are used to define the poling function. These distances are arbitrary and can be applied at the discretion of the user. For example, a set of atom-based chemical features (e.g., heteroatoms) could be identified and poling distances defined between all such features. This would promote conformational variation between those regions of interesting space. As another example, if a series of molecules has a large common substructure, poling distances can be defined between this substructure and the rest of the molecule to promote conformational variation in those parts of the molecules that differ from each other.

In the sections that follow, we describe one such application of the poling function that involves a linear set of poling distances.

A Prototypical Poling Experiment

METHODOLOGY

The molecules used in this study are shown in Figure 6 and have been chosen as a representative sample of structural types that might be encountered. The molecules are submitted to a quasi-random sampling conformational analysis method, with and without sequential poling. The two methods are compared and contrasted on the grounds of the (1) energetics of the resulting conformers and (2) coverage of the space.

Compound Report 6 **18 Compounds**

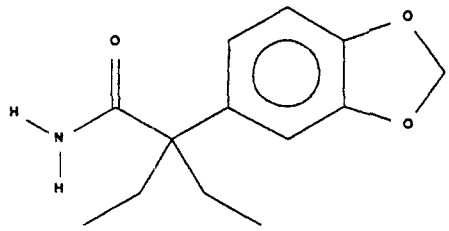
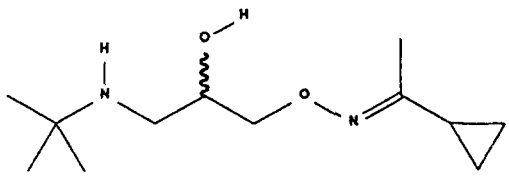
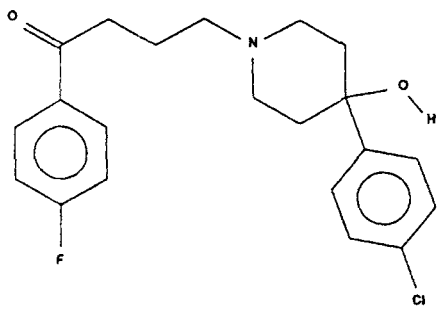
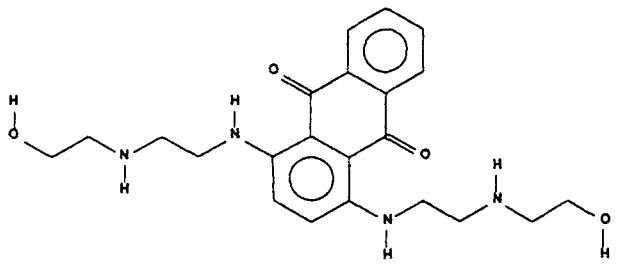
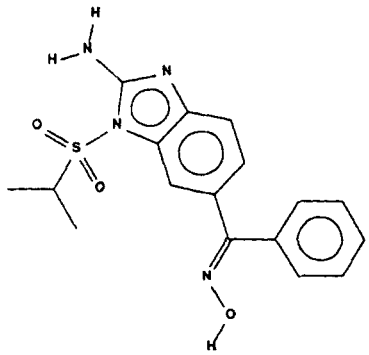
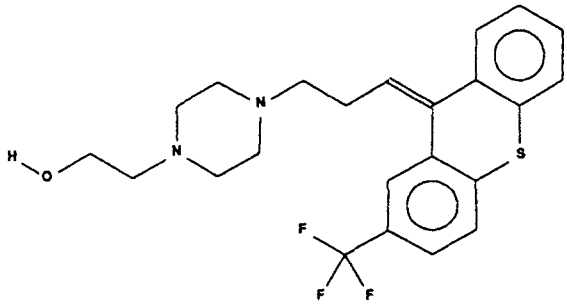
CAS:	Compd 1	CAS:	Compd 2
			
MW:	235.28	MW:	228.33
Formula:	C13H17NO3	Formula:	C12H24N2O2
Name:	AADIETHYLPHENYLACETAMIDE34DIOXYMET-	Name:	FALINTOLOL
Comments:		Comments:	
CAS:	Compd 3	CAS:	Compd 4
			
MW:	375.87	MW:	412.49
Formula:	C21H23ClFNO2	Formula:	C22H28N4O4
Name:	HALOPERIDOL	Name:	AMETANTRONE
Comments:		Comments:	
CAS:	Compd 5	CAS:	Compd 6
			
MW:	358.41	MW:	434.52
Formula:	C17H18N4O3S	Formula:	C23H25F3N2OS
Name:	ENVIROXIME	Name:	FLUPENTIXOL
Comments:		Comments:	

FIGURE 6. Report sheet for the molecules studied.

Compound Report 6 **18 Compounds**

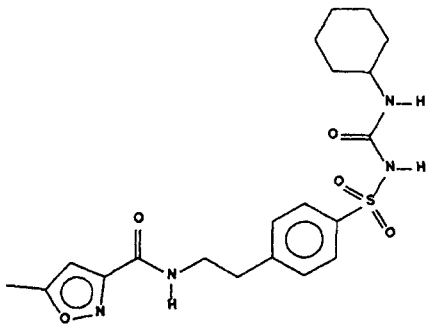
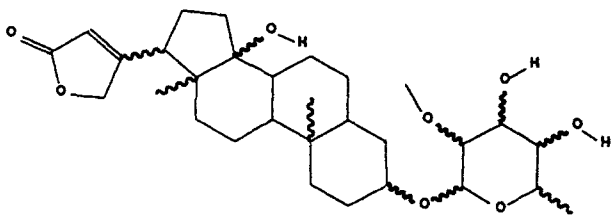
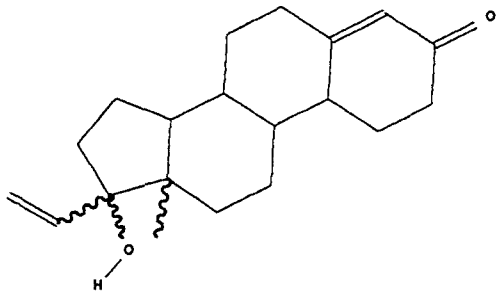
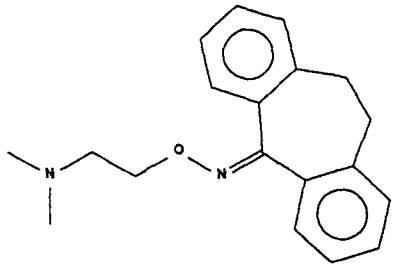
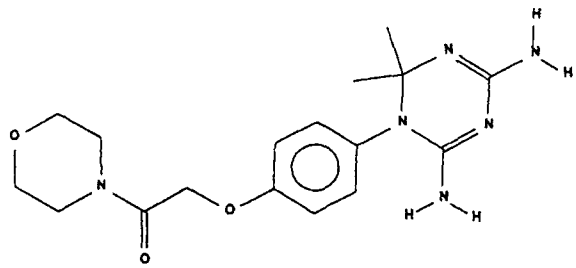
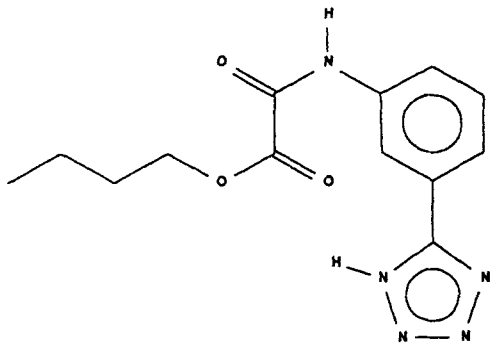
CAS: ..... Compd 7	CAS: ..... Compd 8
	
MW: 434.51	MW: 534.69
Formula: C20H26N4O5S	Formula: C30H46O8
Name: GLISOLAMIDE	Name: NERIIFOLIN
Comments:	Comments:
CAS: ..... Compd 9	**CAS: ..... Compd 10**
	
MW: 300.44	MW: 294.40
Formula: C20H28O2	Formula: C19H22N2O
Name: NORVINISTERONE	Name: NOXIPTILIN
Comments:	Comments:
CAS: ..... Compd 11	**CAS: ..... Compd 12**
	
MW: 360.42	MW: 289.29
Formula: C17H24N6O3	Formula: C13H15N5O3
Name: NSC140020	Name: TAZANOLAST
Comments:	Comments:

FIGURE 6. (Continued)

Compound Report 6 **18 Compounds**

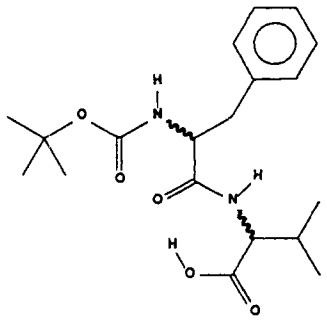
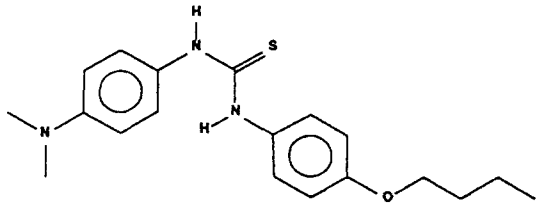
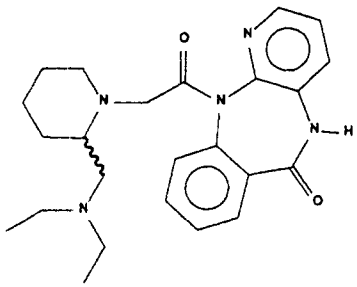
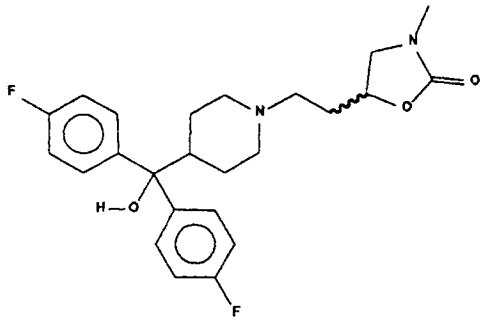
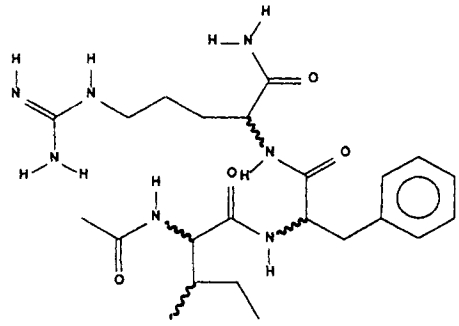
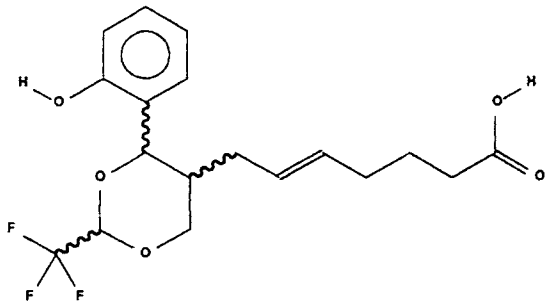
CAS: ..... Compd 13	CAS: ..... Compd 14
	
MW: 364.44	MW: 343.49
Formula: C19H28N2O5	Formula: C19H25N3OS
Name: TBOCPHEVAL	Name: THIAMBUTOSINE
Comments:	Comments:
CAS: ..... Compd 15	**CAS: ..... Compd 16**
	
MW: 421.54	MW: 430.49
Formula: C24H31N5O2	Formula: C24H28F2N2O3
Name: AFDX116	Name: AHR14310C
Comments:	Comments:
CAS: ..... Compd 17	**CAS: ..... Compd 18**
	
MW: 475.59	MW: 374.36
Formula: C23H37N7O4	Formula: C18H21F3O5
Name: ACILEPHEARGN	Name: ICI185282
Comments:	Comments:

FIGURE 6. (Continued)

QUASI-RANDOM SAMPLING CONFORMATIONAL ANALYSIS, NO POLING

The algorithm for the quasi-exhaustive conformational analysis method is outlined in algorithm 4. Conformers are generated using a distance geometry algorithm with full distance metrization[3] to improve the conformational sampling. An embedding is considered successful when the maximum distance error is less than 0.5 Å. These initial conformers are submitted to energy minimization in a CHARMm[10]-derived force field using conjugate gradient minimization[11] far from a local minimum and quasi-Newton minimization[12] in the region of a local minimum. Energy minimization is deemed to have converged when the rms gradient on any atom is less than 0.1 kcal/Å or the function changes by less than 0.001 kcal over 10 iterations. A large number of conformers (255) is generated for each molecule and defines a raw quasi-exhaustive coverage of the space for that molecule. Clearly this set is not truly exhaustive and contains many redundant conformers, but it will be contrasted against a conformational set that is generated under the exact same conditions, but with the addition of poling.

Algorithm 4: Quasi-Exhaustive Conformational Analysis

```
i = 0
maxTries = 10
finished = FALSE
numDesiredConformers = Number of desired
  conformers (255)
while (not finished) {
   for (nTries = 0; nTries < maxTries; ++nTries) {
      embed comformer i
      if (embedding successful) {
         energy minimize conformer i
         (converge when gradient norm
            < 0.1 kcals/Å or
         when function changes by  < 0.01 kcals
            in 10 iterations).
         break
      }
   }
   if (nTries = = maxTries) break
   i = i + 1
   if (i = = numDesiredConformers) {
      finished = TRUE
   }
}
```

QUASI-RANDOM SAMPLING CONFORMATIONAL ANALYSIS, WITH POLING

Initial conformers are generated using distance geometry, as in the quasi-exhaustive algorithm just described. However, each candidate conformer is energy minimized in the presence of poling.

The poling function of eq. (1) is based on a set of interatomic distances. It would be inefficient to use every interatomic distance, because this would make each energy (and gradient) evaluation expensive. To address this, a prior distribution of chemically significant features was selected topologically for each molecule. For this study, chemically significant features were defined as H-bond acceptors, H-bond donors, and hydrophobes.[15] Perhaps the most obvious set of poling distances to select now are all interfeature distances. However, this means that the number of poling distance rises quadratically with the number of features. Because the poling function (and derivative) must be evaluated at least once per iteration of minimization, a linear poling scheme was devised. In this scheme, a poling distance was defined from every feature to the centroid of all the features. By doing this, ensuring conformational variation in the poling distances will promote conformational variation in the shape of the molecule while using only one poling distance per feature.

The energy of the poling function of eq. (1) must be scaled relative to other energy contributions from the force field. If the poling energy is too high, conformers will repel each other to such an extent that it will quickly become impossible to find new (and valid) low-energy conformers. If the poling energy is too low, conformers will not repel each other enough, and many similar conformers will be produced. The scaling factor was estimated by taking advantage of the overall scaling of the energetics of the molecule. It was felt that, because molecular mechanics energies have been calibrated against empirical data, any poling energy contribution should be scaled against the energy of the molecule. A simple algorithm that does this scales the poling energy to be a known value (F_k) at a known interconformer distance (D_k). Returning to the general poling functional form of eq. (1),

$$F_{\text{pole}} = W_{\text{pole}} \sum_i \frac{1}{(D)^N}$$

Let $N = 2$, so the poling contribution from the ith

previous conformer is given by

$$F_{\text{pole}_i} = \frac{W_{\text{pole}_i}}{(D_i)^2}$$

For an rms interconformer distance D_k, the poling energy should scale to F_k, so

$$W_{\text{pole}_i} = F_k(D_k)^2 \quad (5)$$

In this experiment, the poles were scaled so that the poling energy from the ith previous conformer to the current conformer (F_i) = 3.0 kcal at a poling distance (D) of 1.0 Å. The same initial conformers were used as in the quasi-exhaustive study described earlier, but two convergence criteria for the analysis were introduced that took advantage of the poling information. The algorithm used in this sequential poling study is summarized in algorithm 5. The user supplies an energy threshold (which defines the energy spread of the resulting conformers), a failure count (the maximum number of consecutive failures to generate a novel conformer before conformational analysis terminates), and a minimum interconformer distance (which defines a distance below which the conformer is rejected). Conformers can fail for two reasons: (1) The energy of the new conformer is too high, or (2) the new conformer is too close to any previous conformer.

Algorithm 5: *Prototypical sequential Poling*

```
user enters: energyThreshold
user enters: maxFailCount
user enters: minDist
failCount = i = 0
finished = FALSE
select a set of poling distances (from features)
compute a poling weight from Eq. [5]
minConformerEnergy = LARGE NUMBER
while (not finished) {
    maxConformerEnergy = minConformerEnergy +
      energyThreshold
    embed conformer i
    if (embedding successful) {
        if (i > 0) {
            energy minimize conformer i, poling
              against conformers 0 through (i - 1)
            converge when gradient norm
              < 1.0 kcal/A or
            when function changes by  < 0.01 kcals
              in 10 iterations.
        }
    }
    i = i + 1
    oldFailCount = failCount
    if (distance to closest conformer < minDist) {
        failCount = failCount + 1
    }
    else if (conformer energy > maxConformerEnergy) {
        failCount = failCount + 1
    }
    else {
        if (conformer energy < minConformerEnergy) {
            minConformerEnergy = conformer energy
            prune any high energy conformers
        }
        failCount = 0
    }
    if (failCount = = maxFailCount)
      finished = TRUE
}
```

A sample set of 18 molecules, representing a range of sizes and structural motifs, was chosen from the BioByte database.[13] Because no stereochemical information is specified in the BioByte database, each molecule was taken to represent the *union of all of the stereoisomers consistent with the specified molecular topology*. By doing so, we hoped to challenge the poling technique to cover large, complicated conformational spaces. For example, molecule 17 is a functionalized tripeptide, but with four unspecified stereocenters it represents 16 molecules, each with over a dozen rotatable bonds.

Conformer sets were generated using each strategy described earlier, and statistics were collected for each set. Let set P_i be the conformer set for molecule i derived from the sequential poling strategy, and let NP_i be the conformer set for molecule i derived from the nonpoling strategy. For each molecule i in each set, we report the (1) minimum, maximum, and mean energies found from each set; (2) maximum and mean hole size in coordinate and distance space found by set P_i in set NP_i; (3) maximum and mean hole size in coordinate and distance space found by set NP_i in set P_i; (4) mean number of new distance bins found by set P_i in set NP_i; and (5) mean number of new distance bins found by set NP_i in set P_i.

RESULTS: POLE DISTANCE HOLE SIZE

In Table I, we report the maximum and mean hole sizes in pole distance space, where the distance metric used in the generic hole size two-set measure described earlier is the distance from each feature to the centroid of features as used in the poling function. Specifically, the distance between

TABLE I.
Pole Distance Hole Sizes for Generated Conformations.

Mol.	Hole size found in set *P* by set *NP*			Hole size found in set *NP* by set *P*		
	#Confs.	Max	Mean	#Confs.	Max	Mean
1	255	0.4	0.2	23	0.5#	0.2
2	255	0.6	0.3	59	0.6	0.4#
3	255	0.6	0.3	178	2.4#	1.1#
4	255	0.8	0.6	255	2.0#	1.0#
5	255	0.3	0.1	33	0.6#	0.4#
6	255	0.8	0.5	255	1.8#	0.7#
7	255	1.0	0.5	255	1.7#	0.9#
8	255	1.1	0.7	161	2.5#	1.1#
9	255	0.6	0.3	37	1.2#	0.4#
10	255	0.5	0.3	47	0.5	0.3
11	255	0.7	0.3	40	1.4#	0.6#
12	255	0.7	0.3	94	1.0	0.5#
13	255	0.8	0.5	64	0.8	0.5
14	255	0.4	0.2	36	0.9#	0.4#
15	255	0.8	0.4	52	0.8	0.4
16	255	0.7	0.4	129	1.5#	0.6#
17	255	1.0	0.7	255	1.7#	1.0#
18	255	1.0	0.9	223	1.4#	0.9

Units are Å.
Poled set found a larger hole in the unpoled set.

conformers is defined as the rms difference in the respective poling distances (from each conformer). The superscript * is used in the tables to mean that the unpoled set found a larger hole in the poled set than vice versa. The superscript # is used in the table to mean that the poled set found a larger hole in the unpoled set than vice versa. This notation was not used when the absolute magnitude of the hole sizes was less than 0.1 Å.

There was no example of the unpoled set finding a larger hole (in pole distance space) in the poled set than vice versa. In other words, the poled set covered the conformational space of poling distances better than the unpoled as it was designed to do, with fewer conformers on average. In addition, the poled set found a larger hole in the unpoled set a total of 26 times, not reporting ties or any hole size less than 0.1 Å. Reassuringly, poling improves the coverage of conformational space with respect to its own metric.

RESULTS: POLE DISTANCE BINS

Following the distance bin metric described previously, each poling distance was divided into 0.5 Å bins, and a bin was marked *touched* if there was a conformer in the set that had this particular distance. This procedure was repeated for every poling distance used in the molecule (of which there was one for every feature). Thus, distance bins were generated for every poling distance for every molecule in both the poled and the unpoled sets. Table II reports the mean number of extra

TABLE II.
Mean Number of Extra Distance Bins Touched.

Mol.	Set *NP* and not set *P* *DB(NP, ~ P)*	Set *P* and not set *NP* *DB(P, ~ NP)*
1	0.0	1.18
2	0.1	1.5
3	0.0	5.67
4	0.0	3.67
5	0.0	2.07
6	0.0	2.82
7	0.0	3.10
8	0.0	6.19
9	0.37	2.12
10	0.0	1.11
11	0.06	2.59
12	0.0	1.74
13	0.06	1.24
14	0.0	2.08
15	0.0	1.0
16	0.06	2.94
17	0.0	2.73
18	0.0	2.05

distance bins touched by the unpoled set (i.e., that were not in the poled set) and the mean number of extra distance bins touched by the poled set (i.e., that were not in the unpoled set).

In every molecule, the poled set touched more extra distance bins per poling distance than the unpoled set, showing that conformational variation was promoted in the distances that were used in poling. In almost every molecule, all distance bins touched by the unpoled set were touched by the poled set ($DB(NP, \sim P) = 0$). In other words, in the poled distance space, the poled conformers touched all the space covered by the unpoled conformers. The distance bins had a width of 0.5 Å, so touching one extra bin per poling distance means that on average each poling distance is spanning or sampling one extra 0.5-Å distance range. The poled set managed to touch more unique bins (i.e., bins not in the unpoled set) than vice versa, even with a median number of conformers one fourth the number of unpoled conformers, demonstrating that poling has performed well; it has promoted conformational sampling in the user-defined poling distances.

RESULTS: ENERGETICS

A legitimate concern about the use of the poling function is that it directly modifies the molecular mechanics force field and that a poorly scaled set of poles may result in the final conformers having energies that "float" too far above the unmodified molecular mechanics energy surface, especially if the poles are too high. Table III shows the minimum, maximum, and mean energies of the conformers generated with and without poling. There are several observations to be made from this table:

- The energies of the conformers with poling are generally higher than without poling. This is to be expected because conformers generated with poling are not generally situated at local minima on the molecular mechanics surface.
- Generally, the difference in global minimum energies between the poled and unpoled conformers differs by no more than about 2 kcal, suggesting that the global minimum for poled conformers is a reasonable estimate of the unpoled global minimum, despite the poles. Of course, the estimate of the (unpoled) global minimum can be straightforwardly improved at the expense of a small increase in runtime by minimizing each poled conformer with respect to the force field without the poles. The unpoled, minimized conformers can be thrown away once

TABLE III.
Energetics for Generated Conformations.

	Without poling				With poling			
Mol.	#Confs.	Min e.	Max e.	Mean e.	#Confs.	Min e.	Max e.	Mean e.
1	255	65.30	73.84	67.94	23	67.37	85.87	76.93
2	255	3.33	10.68	6.37	59	4.93	24.74	14.18
3	255	27.21	35.14	31.06	178	28.07	47.95	39.02
4	255	54.53	66.41	58.22	255	54.53	73.87	63.66
5	255	38.88	44.00	39.94	33	39.78	58.65	47.83
6	255	52.98	65.28	55.48	255	53.42	72.35	62.25
7	255	17.70	33.34	24.04	255	18.45	38.31	28.52
8	255	59.02	71.89	62.77	161	55.66	75.35	68.65
9	255	25.35	45.91	35.67	37	30.08	48.70	42.17
10	255	51.08	67.40	53.49	47	52.12	71.88	62.09
11	255	44.54	51.45	45.67	40	44.54	64.12	52.81
12	255	28.37	40.31	32.23	94	29.10	48.85	38.30
13	255	14.92	30.26	19.50	64	17.13	36.31	25.84
14	255	44.69	52.62	47.63	37	45.14	65.04	55.83
15	255	68.64	88.28	75.22	52	71.20	90.67	81.10
16	255	45.72	63.34	49.97	129	46.21	65.85	55.97
17	255	16.04	30.67	22.38	255	17.49	37.24	27.17
18	255	20.55	40.39	35.51	223	23.20	43.13	45.83

Units are kcal.

their energy has been evaluated, because the poled conformers provide better coverage per unit conformer.

- Molecule 9 is an example of a molecule for which the poled estimated global minimum energy is unusually far away from the unpoled, estimated global minimum. Upon further investigation, this molecule proved to be highly constrained, and poling was pushing hard to try to ensure conformation variation. The poled set did have a conformer that was within 0.1 Å (rms of heavy atom coordinates) of the global minimum energy conformer of the unpoled set, so the shape of that conformer was represented in the poled collection and, once again, minimizing the poled conformers without the poles would provide an excellent estimate of the globally minimal energy.
- The maximum and mean energies of the conformers of the poled set were almost always greater than their unpoled counterparts, but the poled set always explored the full range of conformational space permitted from the user-supplied energy threshold (in this case, 20 kcal) obtained by taking the $\Delta(\max\ E - \min\ E)$. Often the unpoled set repeatedly fell into local minima and did not explore the full range of the specified conformational space.
- In one example, molecule 8, the poled set actually found a lower global minimum energy conformer than the unpoled set. In this case, poling was able to push the conformational search to an area of space that contained a lower energetic minimum.

RESULTS: FEATURE COORDINATE HOLE SIZE

In an attempt to relate the poling results to more traditional rms metrics, we report the maximum and mean hole size in feature coordinate space in Table IV. This is computed using a hole-size metric based on the rigid body superposition of all heteroatoms and centroids representing hydrophobic points: the features for which poling distances were defined. During the development of the poling method, the sizes of the poles, which were tuned by running poled conformational analysis against quasi-exhaustive search methods on a large number of molecules, were chosen so that the mean (or median) feature coordinate hole size that any method of conformational analysis would

TABLE IV.
Feature Coordinate Hole Sizes for Generated Conformations.

	Hole size found in set *P* by set *NP*			Hole size found in set *NP* by set *P*		
Mol.	#Confs.	Max	Mean	#Confs.	Max	Mean
1	255	1.79*	0.80*	23	0.58	0.33
2	255	1.15	0.61	59	1.16#	0.81#
3	255	0.87	0.58	178	2.28#	1.04#
4	255	1.37	0.95	255	2.09#	1.13#
5	255	0.80	0.43	33	0.91#	0.52#
6	255	1.10	0.68	255	2.00#	0.84#
7	255	1.46	0.99	255	1.99#	1.21#
8	255	1.75	0.94	161	2.77#	1.37#
9	255	1.33*	0.62*	37	1.03	0.48
10	255	0.73	0.42*	47	0.74#	0.38
11	255	1.11	0.55	40	1.38#	0.67#
12	255	1.33*	0.84*	94	1.19	0.71
13	255	1.70*	1.09*	64	1.39	0.80
14	255	1.19	0.67	36	1.34#	0.68#
15	255	1.45*	0.88*	52	1.08	0.57
16	255	1.21	0.72	129	1.83#	0.80#
17	255	2.08	1.33	255	2.19#	1.40#
18	255	1.68	1.15	223	1.70#	1.17#

Units are Å.
#Poled set found a larger hole in the unpoled set.
*Unpoled set found a larger hole in the poled set.

find in the set of poled conformers is smaller than about 1.25 Å.

Despite the often modest number of conformers generated, the mean hole size that the unpoled conformers found in the poled conformers is under 1.25 Å for all molecules except molecule 17, where the mean is a borderline 1.33 Å, and the maximum hole found is a somewhat large 2.08 Å. Molecule 17 is a functionalized tripeptide with four unspecified stereocenters (i.e., actually 16 molecules), which is pushing the limit of what distance geometry with poling can represent with only 255 conformers. The unpoled set found a larger hole (either maximum or mean hole size) in the poled set than vice versa 11 times, but the poled set found a larger hole in the unpoled set 25 times. Even though poling took place in distance space, the generated conformers form a respectable covering of feature coordinate rms space.

RESULTS: COORDINATE HOLE SIZE

Using the hole-size metrics previously described, the maximum and mean hole sizes in coordinate space (after superposition of heavy atoms) were computed and are reported in Table V. In this case, the abstract distance used to describe the hole-size metric is the rms difference in coordinates after rigid-body superposition of all heavy atoms.

Even though poling focuses only on distances to user-specified features, the mean hole sizes that the unpoled sets find in the poled sets using rms over all heavy atoms are all less than or equal to 1.3 Å. The maximum hole sizes have gone up, of course, because the generated sets of poled conformations have made no attempt to minimize rms differences over all heavy atoms. Of course, all heavy atoms may be specified as features, in which case poling should reduce the sizes of coordinate holes to values nearer the feature coordinate hole sizes.

The unpoled set found a larger hole (either maximum or mean hole size) in the poled set than vice versa 19 times, but the poled set found a larger hole in the unpoled set 17 times. Note that we are poling on distances from each feature to the centroid of the features and are not promoting variation in every atom position. Thus, although the centroid of a phenyl ring (a feature we pole on) might stay at the same location in space, the phenyl ring itself might rotate, so the results of a coordinate rms fit might show a large difference in the conformers even if the poling distances are similar. Even though we were not poling directly on every atom position, the poled set still performed credi-

TABLE V.
Coordinate Hole Sizes for Generated Conformations.

	Hole size found in set *P* by set *NP*			Hole size found in set *NP* by set *P*		
Mol.	#Confs.	Max	Mean	#Confs.	Max	Mean
1	255	1.68*	0.84*	23	0.60	0.33
2	255	1.40*	0.97*	59	1.15	0.82
3	255	1.28	0.88	178	2.27#	1.13#
4	255	1.42	0.98	255	1.89#	1.04#
5	255	1.08*	0.71*	33	0.79	0.51
6	255	1.18	0.74	255	1.86#	0.88#
7	255	1.66	1.12	255	1.99#	1.21#
8	255	1.76	0.87	161	2.42#	1.20#
9	255	1.40*	0.72*	37	0.87	0.46
10	255	1.62*	0.73*	47	0.82	0.50
11	255	1.33	0.94*	40	1.38#	0.75
12	255	1.58*	0.88*	94	1.14	0.74
13	255	1.85*	1.30*	64	1.62	0.86
14	255	1.56	1.12*	36	1.67#	0.95
15	255	1.63*	1.01*	52	1.13	0.68
16	255	1.43	0.93*	129	1.74#	0.88
17	255	2.06	1.25	255	2.07#	1.33#
18	255	1.61	1.13	223	1.68#	1.16#

Units are Å.

bly, beating (in terms of finding larger hole sizes) the unpoled set almost half the time.

Each molecule had 255 conformers generated without poles, but the median number of conformers generated with poles is only 64. Even with fewer conformers, the poled set was able to perform well at finding holes in the unpoled set.

Conclusions

This article has presented two novel quantitative metrics for assessing conformational coverage —the hole-size and distance bin metrics. These metrics were used to assess the efficacy of a new technique, poling, which promotes conformational variation. In this technique, poling distances are defined topologically and are used to construct a penalty function that is added to a potential energy function during energy minimization. The poling function penalizes conformers that are too close together and is general—it can be applied to any collection of poling features in molecules of arbitrary size.

A prototypical poling experiment was performed in which a poling function was used to promote conformational variation of a test set of 18 molecules from the BioByte database. Conformational analysis was performed twice for each molecule using the same underlying algorithm, except that one set was generated with poles and the other set was generated without poles.

The following results were presented:

- A well-scaled poling function does not severely distort the molecules energetically, even though the molecular mechanics surface is being modified by the poling function. It was seen that the global minimum energies of the two sets were in close agreement. The mean energies of the poled set were somewhat higher than those of the unpoled set, consistent with our goal of covering conformational space with respect to a user-defined energy threshold rather than just elucidating local minima. Poling exploits the additional conformational freedom implied by such an energy threshold to explore a wider range of conformations, as evidenced by poled conformers that span the permitted energy range.
- Using poling promotes conformational variation in the distances that are being poled. The improved variation was especially noticeable when hole sizes were examined in pole distance space as opposed to heavy atom coordinate space.
- Poling is a highly effective technique for varying the conformational arrangement of a collection of molecular features (i.e., the poled features). To a lesser, but measurable, extent, poling also varies conformational characteristics that correlate with poling distance—for example, the absolute positions of features in space (as measured by the hole-size metrics in feature coordinate rms space).
- Poling forces a conformational analysis algorithm to explore more of conformational space than the unpoled version of the algorithm would, as shown by the distance bin metric. Poling prevents conformers from dropping into the same local minimum on the molecular mechanics surface and encourages conformers to be dissimilar.

Acknowledgments

The authors would like to thank Jonathan Greene and Patrick Coffey for their valuable input and stimulating discussions.

References

1. M. Lipton and W. C. Still, *J. Comp. Chem.*, **9**, 343 (1989).
2. R. A. Dammkoehler, S. F. Karasek, E. F. Berkley-Shands, and G. A. Marshall, *J. Comp. Aid. Mol. Des.*, **3**, 3 (1989).
3. G. M. Crippen and T. F. Havel, *Distance Geometry and Molecular Conformations*, Research Studies Press, Ltd., John Wiley & Sons, New York, 1988.
4. M. Saunders, *J. Am. Chem. Soc.*, **109**, 3151 (1987).
5. See, for example, P. F. Shenkin, D. L. Yarmush, R. M. Fine, H. Wang, and C. Levinthal, *Biopolymers*, **26**, 2053 (1987).
6. See, for example, J. A. McCammon and S. C. Harvey, *Dynamics of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK, 1987.
7. See, for example, N. Go and H. A. Scheraga, *Macromolecules*, **11**, 552 (1978).
8. H. C. Romesburg, *Cluster Analysis for Researchers*, Lifetime Learning Pub., Wadsworth Inc.,
9. *Conformational Analysis in Catalyst*, Technical White Paper, 1991, BioCAD Corporation, 555 Oakmead Parkway, Sunnyvale, CA 94086.

10. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaninathan, and M. Karplus, *J. Comp. Chem.*, **4**, 187 (1983).
11. R. Fletcher, *Practical Methods of Optimization* 1: *Unconstrained Optimization*, John Wiley & Sons, New York, 1980.
12. C. G. Broyden, *J. Inst. Math. Appl.*, **6**, 76, 222 (1970).
13. BioByte Corporation, 4070 Olive Knoll Pl., Claremont, CA 91711. The distributed product is named Masterfile.
14. See, for example, J. Millman, *Microelectronics*: *Digital and Analog Circuits and Systems*, McGraw-Hill, New York, 1979, p. 493.
15. J. Greene, S. Kahn, H. Savoj, P. Sprague, and S. Teig, submitted *J. Chem. Inf. Comp. Sci.*, July 1994.
16. Tad Hurst, *J. Chem. Inf. Comp. Sci.*, **34**, 190 (1994).