Conformation of an Ideal Bucky Ball Molecule by Genetic Algorithm and Geometric Constraint from Pair Distance Data

Genetic Algorithm

ABSTRACT
A genetic algorithm is proposed with real value variables, spatially based crossover operator, a small mutation, large scale mutation, vector sum local search and geometric only based objective function to generate candidate molecule conformations from atomic pair distance data. To better simulate experimental data only information from the pair distance data is used as constraints. Ideal Bucky ball with 60 atoms is used as the test case with both perfect pair distance data and Gaussian noise perturbed pair distance data. The GA generated result shows molecules close to ideal Bucky balls but with some defects. A description of the spatially based crossover operator is provided along with a local search based on vector summed error for each atom.

Categories and Subject Descriptors
F.2.2 [Nonnumerical Algorithms and Problems]: Geometrical problems and computations; G.1.6 [Optimization]: Global optimization; J.2 [PHYSICAL SCIENCES AND ENGINEERING]: Physics

Keywords
Genetic algorithm, molecular conformation, NP-hard search, spatial crossover, Bucky ball

1. INTRODUCTION
Molecular conformation is the process of using experimental data to infer the structure of atoms in a molecule. This problem is critical to many areas of current research such as pharmacology, biology, chemistry, and physics. The typical sources for the experimental data include x-ray scattering and NMR studies. These sources provide a histogram or distribution of inter-atomic distances between atoms in the molecule commonly call pair distance function or PDF. The problem is to determine the locations of atoms that will satisfy this set of distances.

The study of the molecular conformation problems have been extensive with entire books on the subject such as the Crippen and Havel 1988 text[2]. The question of uniqueness for a given conformation is clearly an issue along with the fact that experimental pair distance data will have a certain number of flaws and uncertainties. As computer power has grown some of the more traditional search methods have been able to significantly extend the size of the molecules solved. However the size of the molecules that need to be identified has increased. For example, some of the proteins and organic molecules of interest range in the thousands of atoms and more.

The Bucky ball is a unique molecule formed from carbon atoms. The basic shape is similar to a soccer ball and is a difficult example for molecular conformation because of its high degree of symmetry. This symmetry produces a limited number of unique distances in the pair distance function. The molecule shape is made of 10 hexagons and 20 pentagons on the surface of a sphere. The total number of atoms is 60 and in terms of the scale this presents a large search problem if no chemical or symmetry properties are used to reduce the space. A major goal of this work is to use the minimum amount of domain knowledge, i.e. only pair distance data to produce candidate molecules.

1.1 Computation size
In simple terms, assignment of pair distance data to atom pairs can be expressed as $\frac{1}{2}(N - 1)!$ permutation search problem. In the case where the problem is cast as a combinatorial problem, the bond lengths are fixed and the rotation of each bond is quantized into $X$ discrete steps. This search can be expressed in worst-case terms as $X^N$ permutations where $N$ is the number of atoms. The size of these searches makes the use of GA’s a good approach to find effective solutions. This class of problems was proven NP-Hard by Saxe in 1979 [13]. When $N$ is small the problems are solvable by brute force approaches but when $N$ is moderate in value such as the Bucky ball with $N = 60$ the search space is very large.
Many of the current application programs utilize domain knowledge to reduce the conformation problem to a size that is more feasible. Examples of the domain knowledge used includes known structures, type of bonds, number of bonds, bond angles, and previously discovered low potential energy configurations.

A further complication of the problem is the fitness landscape which is densely populated with local minima. The estimate for the number of local minima, in general, is on the order \( e^{N^2} \). Many of the current applications programs have extensive mechanisms for mapping the landscape to localize the area of search or escape from local minima. The use of the Lennard-Jones term of the energy functions has been used to find new global minima for atomic clusters.

A pure geometric method of solving the conformation problem was presented by Hendrickson [6]. In this method an iterative approach using smaller groups of atoms then combining them is used. This method has significant computational cost as the number of molecules increases.

1.2 GA applications
The work by Deaven and Ho [3] used genetic algorithm coupled with a minimum energy function to produce Bucky balls. That work also used a spatially based crossover operation. In their algorithm they made extensive use of escape mechanisms to prevent being locked into a local minima. This work was popular because of its description appearing in a Nature article by Maddox [10]. A more recent paper by Rivera-Gallego describes a genetic algorithm applied to the distance matrix completion problem [4]. In this problem some of the distance assignments are known to specific molecules and the problem is to finish the assignment.

A brief review of the basic applications of genetic algorithms to molecular conformation was prepared by Moscato [9]. In this review several key points are highlighted along with the correct classification of GA’s that include local searches as Memetic algorithms. A far more extensive review of evolutionary algorithms for clustering with potential energy functions by Hartke [5] describes in detail the methods and approaches used historically, leading up to the current trends. In all the work described in this review, energy functions were the dominate means of identifying clusters for conformation. Hartke also reports that using these current methods it is possible to find molecules with the computational cost on the order of \( N^3 \) incorporating an evolutionary algorithm, local searches, and other operators.

Sastry’s work on cluster optimization using energy functions shows a significant reduction in the computational cost of the molecular conformation using seeded values [12].

1.3 Objectives
Develop a method that relies only on experimental data to derive geometry constraints and utilizes a geometric based objective function with a genetic algorithm to produce candidate molecules. This work will present a variety of GA experiments using a geometric objective function and constraints based only on the pair distance data. The objective function was calculated using the pair distance data provided as the target and each candidate molecule. Two pair distance data target sets were used. One set consisted of ideal distances and the second set perturbed ideal distances with a Gaussian noise. Using these target data sets one group of experiments was conducted to determine the range of operational parameters. The group of experiments was conducted in actual search for the ideal Bucky ball.

A description of the genetic algorithm will be presented with additional information about the small mutation, crossover operator, a large scale mutation and a vector summed search. The experimental results includes a set of range experiments for determining the robustness of operating set points along with the results of seeded GA runs using the best-of molecules from many single level runs as the initial genetic material. The results will show candidate molecules with less than 5% average error compared to the ideal pair distance data and less than 1% average error compared to the Gaussian noise perturbed data. The visual displays of the molecules will show a fair resemblance to the ideal Bucky ball but with significant defects.

2. REPRESENTATION & OPERATIONS
Overall operation of the genetic algorithm is based on a replacement population each generation. The chromosome representation uses real number values. All operations are performed on selected parents from the last generation to produce the new members of a fixed size population. Real numbers will be used to represent each atoms location. The genetic operations will include mutations and crossover operations on this representation.

2.1 Representation
The basic representation of each chromosome or member of the population is a column vector of 3D points referenced to the center of mass for each molecule consisting of real valued coordinates in \( x, y, z \). All the operations will be based on this chromosome structure. The geometric objective function uses the inter-atomic distances in an ascending sorted order. The calculation outline is described by equation (1) though equation (4) where the final results is the ascending sorted list of distances. This set of distances is then used to calculate the objective function as shown in equation (5).

The target for the conformational search can come from a variety of experiments. The X-ray diffraction experiment results on crystals and NMR studies produce a histogram of distances that represent the distances between atoms. In the case of NMR studies there is sometimes additional 2D and 3D location information that may be derived about groups of atoms or individual atoms. New advances in the NMR research are using multiple orientations, new frequency bands and pulses to improve the information available. Once the set of distances between all the atoms in a structure is obtained in this histogram form, then using a method like the one proposed in this work would enable production of candidate molecules that have the same set of distances.

A geometric objective function was used during the experiments in contrast to an objective function based on potential energy of the molecule. Several of the published works on conformation use an potential energy based objective func-
tation. The same data from the $D$ array shown in equation (2) is used to calculate the energy. In this case the minimum energy or objective function implies a stable molecule that is most likely to represent the experimental data. One of the main objectives of this work is to show that experimental distances alone can achieve good results and the potential energy calculations without local minima escape mechanisms were used only for reference.

$$d_{ij} = \begin{cases} i \neq j & : |p_i - p_j| \\ i = j & : 0 \end{cases}$$

(1)

$$D = \begin{bmatrix} 0 & d_{1,2} & \cdots & d_{1,n-1} & d_{1,n} \\ d_{2,1} & 0 & \cdots & d_{2,n-1} & d_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ d_{n-1,1} & d_{n-1,2} & \cdots & 0 & d_{n-1,n} \\ d_{n,1} & d_{n,2} & \cdots & d_{n,n-1} & 0 \end{bmatrix}$$

(2)

$$d = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{m-1} \\ d_m \end{bmatrix}$$

(3)

$$\forall \; d_k, \; 1 < k \leq m \; |d_{k-1} \leq d_k$$

(4)

$$objd = \sum_{k=1}^{m} (t_k - d_k)^2$$

(5)

The target $t_k$ is simply the sorted list of the pair distance data from the experiment or in this case the ideal Bucky ball with and without perturbations.

2.2 GA Operation

The operation of the GA utilizes an integrated vector summed search along with two mutation operators, replication and a unique crossover operator. Each of the operators was applied with a probabilistic distribution including the copy operation making a totaled 1.0. The population sizes varied from a few as 800 molecules to 28,000 molecules. The selection process was binary tournament selection for parent(s) as needed. A replacement population was created on each generation using the selection and operators. Runs were conducted only keeping children that were better than the parents and keeping all children. The impact of this Machiavellian strategy was very notable on both the rate of convergence and the best objective function results obtained.

A typical set of parameters is show in the Table 1. During all operations where atoms are relocated a check for violation of minimum distance is conducted. If an atom is in violation, the location of the atom is moved in a direction that will relieve the violation. This minimum distance constraint is completely derived from the pair distance data.

The crossover operator is unique in that it would take a random plane through the center of mass of two molecules. Then for each parent it adjusts the location of the random plane along an axis normal to the plane until half of the atoms are on each side of the plane. The crossover operator then creates a single child from the two parents combining a half from each parent. In the sense of building blocks, atoms that were closer to each other in space have much greater impact on each other than those farther away. Hemispheres with lower values of objective function evaluations will tend, when combined, to create better molecules. Figure 1 shows a crossover operation in progress where the random plane divided the original two molecules in half. This Figure shows the left half from one and the right from the other prior to being combined into a new child.

The vector summed search tries to minimize the distance error for each atom choosing the atom with the greatest distance error to manipulate each pass. Equation (7) shows the vector summation process where the individual error contributions are summed using the normalized vector between the two atoms as the directional contribution. In equation (7) the term $\|p_i - p_j\|$ is the unit vector aligned between two atoms. Once all the error vectors have been calculated the error vector with the maximum magnitude is chosen and applied to the corresponding atom. Note that the scaling factor $sf$ from equation (8) is in a range of 0.5 to 0.001. It is possible to create unstable results with factors over 0.5 and small factors take longer to converge. The largest factor that is stable is desired. The correction to the location of the atom is based on the sum of all the distance error contributions or some subset. The variables $a$ and $b$ in equation (6) set the contiguous range of distances to used.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crossover</td>
<td>0.35</td>
</tr>
<tr>
<td>Mutate</td>
<td>0.20</td>
</tr>
<tr>
<td>Mutate2</td>
<td>0.05</td>
</tr>
<tr>
<td>Vec. Sum</td>
<td>0.15</td>
</tr>
<tr>
<td>copy</td>
<td>0.25</td>
</tr>
<tr>
<td>population</td>
<td>8000</td>
</tr>
<tr>
<td>max generation</td>
<td>3000</td>
</tr>
</tbody>
</table>

Table 1: Typical parameters of GA runs

![Figure 1: Left and Right hemisphere combined in crossover operation](image)
in the calculation. By variation of the contiguous subsets, a robust improvement was obtained. The alternating subsets and full sets of distances achieved the best objective function results in an efficient manner often two or three times faster than either the energy gradient or straight geometric gradient on the objective function. The final version of the genetic algorithm incorporates this vector summed search as an operator with a specific probability of selection.

\[ \forall k \mid a \leq k \leq b \]  
\[ E_i = \sum_{k=a}^{b} \| p_i - p_j \| (t_k - d_k) \mid d_k = |p_i - p_j| \]  
\[ p_i = p_i - sf \ast E_i \mid \max(E_i) \]  

Two mutation operators are used in the GA for small and large scale mutations. The small mutation operator takes some number of atoms and moves them each with a different vector that has random direction and magnitude. The magnitude is a linear random value within a specific range. The large scale mutation operator, often referred to as mutate2 takes a random selected atom and relocates it by selecting a second random atom and placing the first atom a random direction and the minimum distance from that second atom.

The initialization method for the molecules used a random scheme where an existing atom was chosen at random and new atoms were added at minimum distance in a random direction as long as the new atom did not violate a minimum distance constraint. The average had a typical raw objective function of 500 as compared to the resulting best molecule with a raw objective function value of under 10.

When a genetic algorithm run was conducted from seeded material such as the best-of molecules from other runs, the new population was created by random crossovers from the seed population. This prevented the best of the seed molecules from dominating the population immediately and allows for better convergence.

Diversity measurements were also conducted for convergence using a random sample from the population and the basic geometric objective function. When the population converged the average objective function from the best-of molecule to random members of the population were extremely low indicating dominance of that molecule.

3. EXPERIMENTAL RESULTS

A variety of experiments were conducted. Figure 2 shows a typical resulting molecule. In many of the experiments the results of the genetic algorithm were further refined using a vector summed local search. This procedure lowered the value of the objective function by 10%, however there was little change in the visual appearance of the molecule. The purpose of the range experiments was to determine if any specific settings of mutation and crossover operation would yield a significant improvement in the overall operation of the G.A. The outline for the various experiments conducted are listed as follows:

1. Range experiments were conducted with Gaussian perturbed distance as the target data
2. 100+ runs with ideal distance data
3. 6 runs with seeded molecules from best of previous runs
4. Post processed best-of molecules using energy based gradient functions
5. Post processed best-of molecules using geometric gradient functions
6. Post processed best-of molecules using vector summed functions
7. Incorporated the vector summed function in the genetic algorithm as a type of mutation operator

3.1 Range experiments and Gaussian Perturbed Data Target

This series of experiments was used to determine the parameter settings of the basic crossover, mutate and population sizes needed. A single parameter was varied over a range while all the other parameters were kept constant. The distance based objective function was used in all the range runs. For these runs the Gaussian perturbed target data was used. It should be noted that the objective function had better values when perturbed target data was used but the molecules were not significantly better than those generated with ideal data. Table 2 shows the results of the range runs. The best of run is raw objective function or the total error of the distances squared per equation (5).
3.2 Seeded runs
The seeded runs were conducted using between 10 and 50 best-of molecules. These runs were conducted with both high and low mutation rates but with low crossover rates. The results of these runs produced visually better molecules but not significantly better objective function values.

3.3 Post run processing
The best of molecules were subjective to 5 different post processing methods. These included an energy based gradient search, a geometric based gradient search, a vector summed error search, and mixtures of the three basic methods. While all the methods would improve the objective function for the best-of molecules, the improvement was not sufficient to remove visual defects or make break through reductions in the objective function.

4. DISCUSSION OF RESULTS
The range experiments did not produce any significant trends with respect to the best set of operating parameters. The population size did show some impact on the overall operation but with little improvements for very large populations. It was clear from many of the experiments that crossover was the dominate factor and higher rates would produce faster convergence when population sizes from 5000-8000 did provide a consistently better objective function.

The large scale mutation operator was able to find a better visual molecule but did not necessarily improve the objective functions. Attempts to increase the heuristic selection of the atoms to be affected by the large scale mutation did not yield a useful drop in objective function. The problem with this mutate operator was that the temporary impact on the objective function made it difficult to compete for reproduction. When used as a post process with the vector summed operation, good results were obtained.

As is common the objective function does not discriminate enough near the final convergence to help promote the better molecules. Further, the difference between the distance based and energy based objective functions are not significant. Both have terms that are similar and are both incapable of escaping the local minima. While the energy objective functions moved the atoms around, the visual differences were small. The vector summed search produced faster results and better results in a shorter time by some small amount than the two pure gradient searches.

The use of multiple best-of results to seed a hierarchical run did produce a visually better molecule. While this molecule still has visual defects and an objective function only 80% the value of the single level runs it is clearly a better result. Figure 2 as produced with the program Atomeye [8] shows the resulting molecule from the seeded GA.

The analysis of the crossover operators impact produce several unexplained results. First, the use of high selection pressure by only keeping offspring better than the parents produced an unusual profile of constructive crossovers as a function of generation. A more typical operation that keep all the offspring produced the expected profile of constructive crossover operations. The other issue is that any improvement of objective function was expected to become near zero in later generations for the crossover operator. Instead it leveled off in both cases about 7% of operations. While the amount of improvement was very small at the end of the run it would continue to converge for several thousand of generations. The mutation operators did in fact show a much smaller ratio of constructive improvement to the objective functions. The ratio was calculated as the number of operations of a given type (crossover, mutation, etc) that improved on the objective function divided by the total number of operations of a given type per each generation.

A diversity measurement was made on the population using a limited number of samples compared to the best-of molecule for the current run. In these measurements, it was clear that diversity of the Machiavellian approach decreased rapidly while keeping all offspring provided greater diversity for more generations but did not produce better results. Figure 3 shows the impact of the scaled fitness as a function of generation. Equation (9) shows the conversion of the objective function to scaled fitness. There was a significant reduction in the fitness when all offspring were kept. The visual differences were not as major as the numeric results would suggest. The diversity measurement when keeping all offspring was significant for many generations and with the high rates of mutation and crossover never really reduced to that of the Machiavellian method of offspring retention.

$$fitness = \frac{5}{objd^2} + \frac{36}{objd} \quad (9)$$

5. CONCLUSIONS
The use of the spatially based crossover operator provides a means to construct progressively better building blocks for the genetic algorithm. The limit of the crossover methods to improve on the objective function is reached once the building blocks obtain a given size and it is not possible to exclude poorly placed atoms. The population sizes in the range 5000

<table>
<thead>
<tr>
<th>run</th>
<th>rd1</th>
<th>rd2</th>
<th>rd3</th>
<th>rd4</th>
<th>rd5</th>
<th>rd6</th>
<th>rd7</th>
<th>rd8</th>
<th>rd9</th>
<th>rd11</th>
<th>rd12</th>
</tr>
</thead>
<tbody>
<tr>
<td>crossover</td>
<td>0.35</td>
<td>0.45</td>
<td>0.25</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td>mutate</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.4</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>mutate2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mutate mag</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>350</td>
<td>150</td>
<td>250</td>
<td>250</td>
</tr>
<tr>
<td>population</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1500</td>
<td>1500</td>
<td>2000</td>
<td>500</td>
<td>1000</td>
<td>5000</td>
<td>5000</td>
</tr>
<tr>
<td>max generation</td>
<td>3000</td>
<td>3000</td>
<td>3000</td>
<td>3000</td>
<td>3000</td>
<td>3000</td>
<td>3000</td>
<td>3000</td>
<td>3000</td>
<td>3000</td>
<td>3000</td>
</tr>
</tbody>
</table>
to 8000 showed a significant improvement in the overall objective function. Population sizes greater than this did not improve this objective functions final convergence. An estimate of probability for atoms in the correct location showed that populations on the order of 15,000 would be needed. This conflicts with the experimental data runs which shows little improvement in convergence or objective function with increases in size beyond 8000. The fact that the crossover operator continues to make small improvements to the objective function after convergence is based on the elastic nature of the recombination process for the production of the child molecule from two parents. These improvements are small and do not address the issue of the visual defects.

The ability of the GA’s with distance based objective function was able to produce molecules with less than 1% average error on inter-atom distances for the Gaussian perturbed ideal data. The resulting molecules have defects that only a potential energy function objective could remove based on in-depth domain knowledge.

6. FUTURE WORK
Continuing development of the geometric based objective functions. Exploration of the possible variations for the vector summed local search. From the analysis of the resulting molecules the improvement of the large scale mutate operator will be required to explore the near optimal space better. The construction of building blocks from subsets of the distance constraints could be used in both the mutation and initialization operations of the GA. The current local search based on the vector summed error could be expanded to present a wave front approach to the local search. In addition, it is clear that some of the assignments of atom pair distances to target distances are incorrect, preventing the local search from providing a better result. Using the GA and some form of operator it may be possible to determine a better fitness by perturbation of the ordered list. Continue to develop the theoretical basis for the use of spatially based crossover and mutation function. These functions have been used in many real world applications of GA’s and will be more significant as more GA’s applications use real value representation. Modeling physical entities shows a natural building block relationship of points close to each other.

7. REFERENCES