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

Genetic Algorithm

David M. Cherba
Dr. William Punch
Computer Science Department
Michigan State University
3105 Engineering Building
East Lansing, MI 48823 USA
{cherbada, punch} @cse.msu.edu

Dr. Phil Duxbury
Dr. Simon Billinge
Dr. Pavol Juhas
Department of Physics and Astronomy
Michigan State University
East Lansing, MI 48823 USA
{duxbury, billinge, juhas}@pa.msu.edu

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

General Terms

Algorithms

Keywords

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

GECCO'05, June 25–29, 2005, Washington, DC, USA.
ACM 1-59593-010-8/05/0006.

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 distance data include x-ray scattering and NMR studies. The problem is to determine the locations of atoms that will satisfy this set of distances.

Crippen and Havel [1] book on molecular conformation sets the foundation work for with distance geometries. The upper bound on the computational cost is $[\frac{N}{2}(N-1)]!$ for assignment of distances to atom pairs or X^N for protein folding where N is the number of atoms. The solution to embedded graphs used to represent molecule structure was proven NP-Hard by Saxe in 1979 [9]. Using energy functions in an optimization form of the problem was shown to have on the order e^{N^2}[5] local minima's.

2. GA CLUSTER CONFORMATION

The work by Deaven and Ho [2] used genetic algorithm and the Lennard-Jones energy function to derive Bucky ball structure using energy minimum criterion for sixty atoms. That work introduced a spatially based crossover operator and used a $(\mu+\lambda)$ population strategy. Hartke [4] expanded on these methods and was able to find all the low energy configurations up to 250 atoms. A recent paper by Rivera-Gallego describes a genetic algorithm applied to the distance matrix completion problem [3]. A survey by Moscato [7] describes the application of Memetic algorithms to cluster conformation. All of the work cited above used real-valued genes. The work by Sastry [8] uses binary encoded atom locations and Minimum Descriptive Length measures for fitness with energy functions.

3. CURRENT WORK

This work uses a real-valued representation for atom location reference to the center of mass. Equation 1 shows the distance calculation and the equation 2 shows the labeled correspondence between atoms and distance.

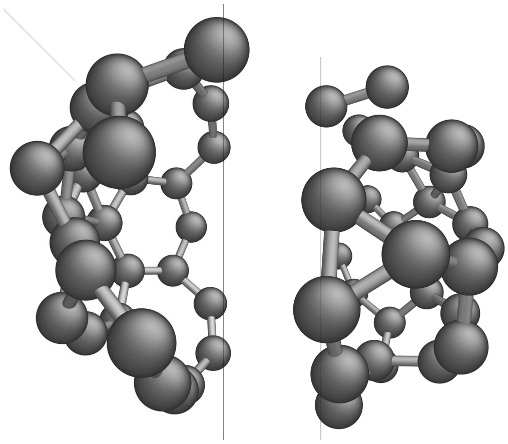

Figure 1: Left and Right hemisphere combined in crossover operation

$$d_{i,j} = \begin{cases} i \neq j & : \quad |p_i - p_j| \\ i = j & : \quad 0 \end{cases} \tag{1}$$

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

$$objd = \sum_{k=1}^{m} (t_k - d_k)^2 \tag{3}$$

The target t_k is the sorted list of distances from experimental data. The equation 3 shows the fitness measure used by this work. The index shift from $d_{i,j}$ to d_k indicates the unlabeled nature of the data. The crossover operators using two halves of parent molecules is shown before final merger in figure 3. The local search was performed using the vector summed method that assigns each contribution of the objective function to a correction vector for each atom. This is described in equation 4 with the correction to each atom shown in equation 5.

$$\mathbf{E}_i = \sum_{k=a}^{b} \|p_i - p_j\| \; (t_k - d_k) \quad | \quad d_k = |p_i - p_j| \tag{4}$$

$$p_i = p_i - sf * \mathbf{E}_i \quad | \quad \max(\mathbf{E}_i) \tag{5}$$

The best results was achieved using a two level GA run. The best of molecules from twenty runs were used as seed material for a second level GA run. This second level run was able to achieve the Bucky ball shown in Figure 2 using atomeye program [6] to construct the figure.

4. CONCLUSIONS

This work was able to derive molecular structure from only distance data using a two level population and a Genetic Algorithm. This method incorporated a spatial based crossover and mutation with addition of a local search using corrective vectors based only on distance. The calculational cost was on the order of $O(n^{3.8})$ and used no seeds or known good configurations of low energy molecules.

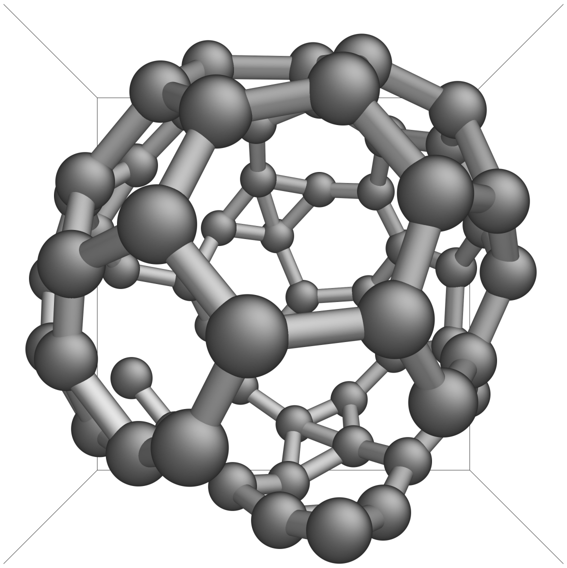

Figure 2: Typical Bucky ball

5. REFERENCES

[1] G.M. Crippen and T.F. Havel, *Distance Geometry and Molecular Conformation.* Studies Press Ltd., Taunton, Somerset, England, 1988.

[2] D. M. Deaven and K.O.Ho. Molecular geometry optimization with a genetic algorithm. *Physical Review Letters,* 75(2), 288-291 Jul 10,1995.

[3] Wilson Rivera-Gallego. A Genetic Algorithm for Solving the Euclidean Distance Matrices Completion Problem, *SAC 1999*, San Antonio, Texas, pg286-290, 1999.

[4] Bernd Hartke. Application of Evolutionary Algorithms to Global Cluster Geometry Optimization, *Structure and Bonding*, Vol. 110 (2004): 33-53, 2004.

[5] M.R. Hoare. Structure and Dynamics of Simple Micro Clusters. *Adv. Chem. Phys.*, 40:49-135, 1979.

[6] J. Li, Modelling Simul. Mater. Sci. Eng. 11 (2003) 173.

[7] P. Moscato, Memetic algorithms for molecular conformation and other optimization problems, *Newsletter of the Commission for Powder Diffraction*, no. 20, 1998.

[8] Kumara Sastry, Efficient Cluster Optimization Using ECGA with Seeded Population, Optimization by Building and Using Probabilistic Models, *(OBUPM) 2001*, Jul., pg222-225.

[9] J.B. Saxe, Embeddability of weighted graphs in k-space is strongly NP-hard, *In Proc. 17th Allerton Conference in Communications, Control and Computing*, pp. 480-489 1979.