MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **55** (2006) 305-314

ISSN 0340 - 6253

Generating Molecules with Specific Properties: Acyclic Hydrocarbons

Steven A. Alexander
Department of Physics
Southwestern University
Georgetown, TX 78626

(Received June 16, 2005)

We describe a Monte Carlo algorithm that can generate molecules with specific properties. For testing purposes we confine this initial investigation to acyclic hydrocarbons (C_NH_{2M}) and use a simple formula to evaluate their magnetic susceptibility.

1. Introduction

A huge number of molecules are used today for industrial or medicinal purposes. Identifying a compound that has a specific set of properties often requires an elaborate trial and error search through a vast collection of potential candidates. In principal if one could define a problem to be solved in sufficient detail, theoretical chemical methods could be used to generate a molecule that meets the required specifications. Such an algorithm would consist of three parts: (1) A method for changing an existing molecule into a new one and for discarding unphysical candidates e.g. one in which a carbon atom has five bonds. (2) A method for calculating the desired properties of each molecule. (3) A method for selecting those molecules which optimize the properties of interest. Earlier attempts to generate molecules with specific properties have documented some impressive successes. These methods generally arrange a set of basic building blocks (atoms, chemical groups, etc.) using either combinatorial methods [1-3] or genetic algorithms [4,5]. In this paper we describe how molecules can be constructed from a Markov-chain-like walk which is guided by a simulated annealing optimization. For the sake of simplicity we choose to limit this initial investigation to acyclic hydrocarbons (C_NH_{2M}) and a simplified formula for the magnetic susceptability.

2. Generating random molecules

Neutral acyclic hydrocarbons can be graphically represented by a set of carbon atoms on a square lattice. An adjacency matrix records which atom is connected to which and with what type of bond (single, double or triple). Only the carbon atoms in these molecules need to be recorded since the number of hydrogens attached to each carbon is directly related to the number of bonds attached to that carbon. Given some initial molecule there are two actions that can be performed on the adjacency matrix that will create a new molecule:

1. Randomly select a carbon atom in the initial molecule and attach a new carbon atom to it. Both the type of bond given this new atom and its direction on the lattice are chosen at random. If the new atom is placed on an occupied site or if the selected atom has more than four bonds attached to it, the original molecule is left unchanged.

2. Randomly select a carbon atom on the initial molecule and delete it. All bonds with this atom are severed. If this step leads to the molecule being split into two unconnected pieces, the original molecule is left unchanged.

The transformation of the initial molecule to some final structure is similar to a random walk through a parameter space. Since there are only two actions, the steps in this walk can be defined by the probability with which new atoms are added. When the probability of adding a new atom is set at 0.55, more atoms are likely to be added than deleted. After 100 steps the final molecule in Figure 1 contains three times more atoms than the initial molecule. When the probability of adding a new atom is set at 0.5, the chances of adding a new atom is exactly equal to the chance of deleting an existing atom. Not surprisingly, the final molecule in Figure 1 is almost the same size as the initial one. When the probability of adding a new atom is set at 0.45, more atoms are likely to be deleted than added. In Figure 1 the final molecule is half the size of the initial one.

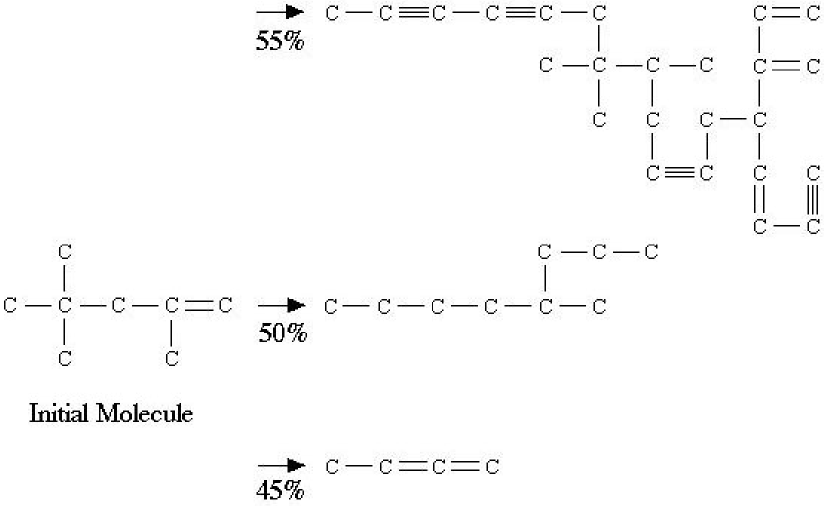


Figure 1. Change in an initial molecule after 100 undirected steps and three different probabilities of adding a new atom.

3. Generating molecules with specific properties

A number of numerical methods have been developed to search for the global minimum of a complicated multidimensional function. Finding a molecule with a specific set of properties is an optimization problem in which the variables are discrete rather than continuous. Simulated annealing is one method that has been successfully applied to problems with both kinds of variables [6-9]. Starting from some initial point, x_{old}, and its value at that point, $f(x_{old})$, this algorithm generates a new point in the multidimensional space, x_{new}, and calculates its value at that point, $f(x_{new})$. If $f(x_{old}) > f(x_{new})$ this step is accepted and x_{new} becomes the starting point for the next step. If $f(x_{old}) < f(x_{new})$, an acceptance function determines whether x_{new} is accepted or rejected. Although several different acceptance functions have been described in the literature [10], a popular choice is the Metropolis function [11]

$$A = \min(R, \exp(-f(x)/T)) \qquad (1)$$

Here R is a random number between 0.0 and 1.0, T is a parameter known as the temperature and f(x) is the function to be minimized. When the temperature is high, most moves are accepted and the whole parameter space is covered. As the temperature is reduced, some moves begin to collect in minima. The idea is to lower the temperature slow enough so that a point can escape from a local minimum and eventually end up in the global minimum. Because each minimization function has different requirements, a variety of cooling schedules have been examined [12,13]. A common choice is the simple geometric relation

$$T_{k+1} = C\ T_k \qquad (2)$$

where C is a constant between zero and one.

To illustrate how simulated annealing can be used to find a molecule with a specific property we set

$$f(x_i) = | \square_i - 200 | \qquad (3)$$

where $\square_i$ is the magnetic susceptibility of the current molecule in magnetic units (1 m.u. = 10^{-6} erg G^{-2} mol^{-1}) and 200 m.u. is the desired value. In Ref. 14 Bytautas, Klein and Schmalz describe how the molecular graph of an acyclic hydrocarbon can be used to calculate its magnetic susceptibility. Their formula assigns a value to each atom depending on how many single, double and triple bonds it has. The molecular magnetic susceptibility may be simply approximated as the sum of these contributions. Figure 2 shows how the minimization function, Eqn. 3, changes over a set of 1000 steps when the probability of adding a new atom is set at 0.5. Unlike the random walk described in the last section, simulated annealing creates a directed walk that quickly converges to the desired value. Starting from the same initial molecule as in Figure 1 and an initial magnetic susceptibility of 90.69 m.u., this walk gets to within 0.5 of the desired value after only 25 steps. The fluctuations that appear after the initial drop are the result of uphill moves that this temperature (T=10) accepts as it searches for the global minimum. During this walk a molecule with 21 carbon atoms and a magnetic susceptibility of 199.99 m.u. gave the lowest value. Because the magnetic susceptibility in this model indirectly depends on the size of the final molecule, it is

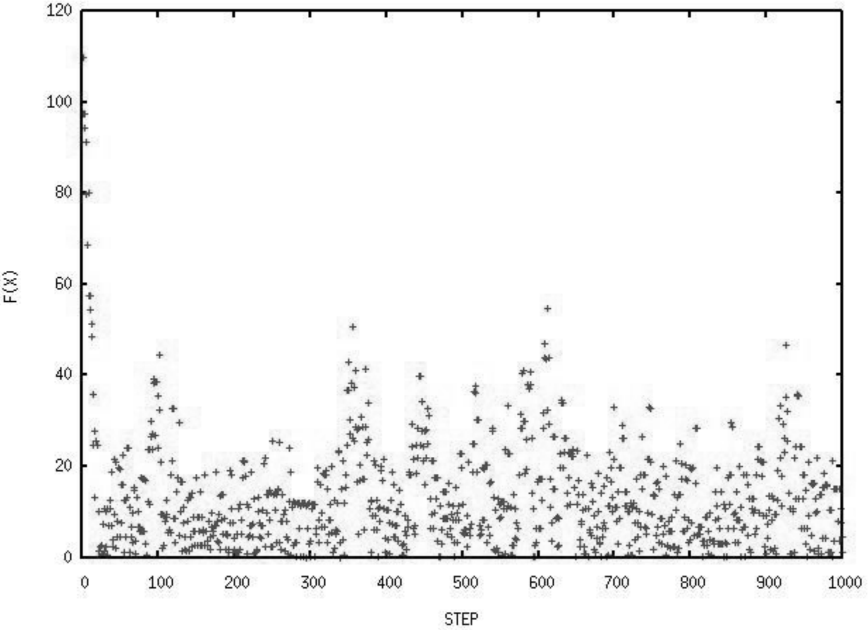


Figure 2. Change in the minimization function, Eqn. 3, during a directed walk at a temperature of 10. This walk is searching for a molecule with a magnetic susceptibility of 200 m.u..

the difference in size between the initial molecule and the final molecule that primarily determines the speed with which the directed walk converges.

In Figure 3 we show the convergence of a directed walk that optimizes both the value of the magnetic susceptibility and the number of carbon atoms

$$f(x_i) = | \square_i - 200 | + | N_i - 25 | \qquad (4)$$

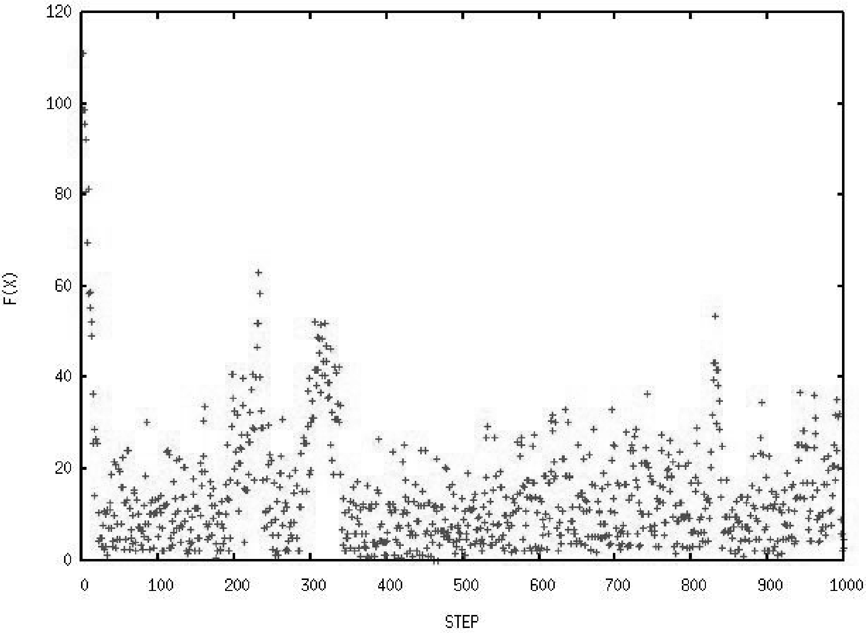


Figure 3. Change in the minimization function, Eqn. 4, during a directed walk at a temperature of 10. This walk is searching for a molecule with a magnetic susceptibility of 200 m.u. and 25 carbon atoms.

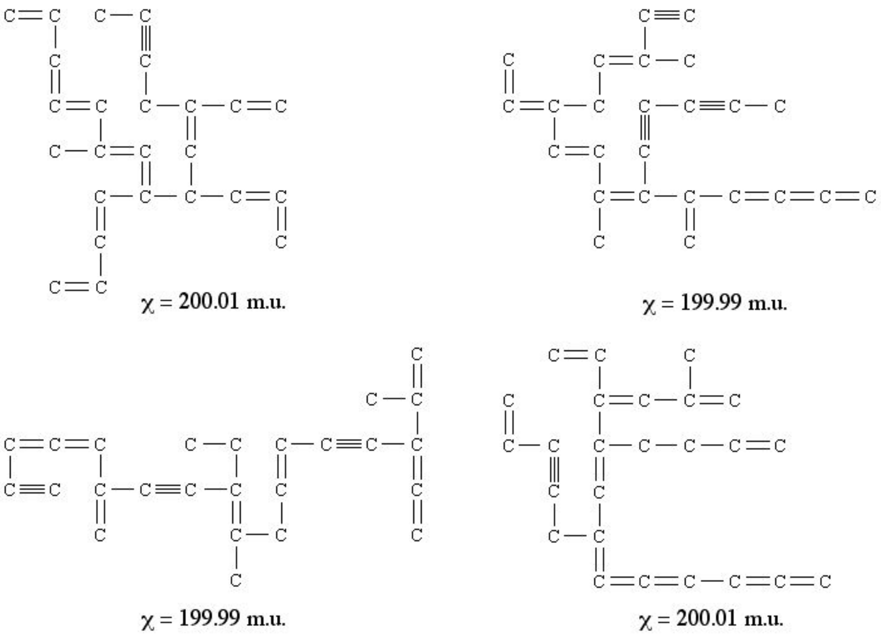
Although the initial molecule and the temperature are the same as in the last test, our optimization routine now has to find a region in this parameter space that satisfies both requirements. As a result, the convergence is slightly slower and this walk needs 175 steps before it gets to within 0.5 of the minimum. During this particular walk a molecule with 25 carbon atoms and a magnetic susceptibility of 200.04 m.u. gave the lowest value. If we want to try and improve this value, we can reduce the temperature using Eqn. 2 and C = 0.75. At each temperature a 1000 step directed walk is performed and the molecule that minimizes T_{iter} becomes the initial point for T_{iter+1}. After 10 reductions of the temperature no new minima were identified. For this test a molecule with 25 carbon atoms and a magnetic susceptibility of 200.01 m.u. gave the lowest value.

Because simulated annealing depends on a series of random numbers, different initial molecules or different random number seeds could lead to different results. In Figure 4 we show the final results of four full runs using Eqn. 4 and different random number seeds. All of these molecules come very close to having the required properties yet they are all structurally different. This ability to identify multiple solutions to a particular problem is yet another advantage of our algorithm.

If we had kept the magnetic susceptibility to be 200 m.u. but set the desired number of carbon atoms to be 5

$$f(x_i) = | \square_i - 200 | + | N_i - 5 | \qquad (5)$$

there would be no molecule in this model that would completely satisfy both requirements. In this case our optimization algorithm would have to balance these two demands as best as possible. The result of this test, obtained after 20 temperature iterations, is a molecule with 17 carbon atoms and a magnetic susceptibility of 197.96 m.u. Since each property in Eqn. 5 is equally weighed, it is understandable that a molecule with a reasonably good magnetic susceptibility and the lowest possible number of carbon atoms would minimize this function. If we had assigned different weights to each property, the final result could easily be changed to a molecule with a smaller number of atoms and a worse magnetic susceptibility.



χ = 200.01 m.u.

χ = 199.99 m.u.

χ = 199.99 m.u.

χ = 200.01 m.u.

Figure 4. Final molecules obtained after a full optimization of Eqn. 5. The goal of this walk was to find a molecule with a magnetic susceptibility of 200 m.u. and 25 carbon atoms.

4. Conclusions

In this paper we have shown how acyclic hydrocarbons can be generated with specific properties. The first part of this algorithm is a method for changing an existing molecule into a new one by adding or deleting individual carbon atoms. This method could easily be generalized to molecules containing any atom or rings. Another possibility would be to use entire chemical groups as the basic building blocks. This flexibility would allow us to easily restrict the generated compounds to specific chemical families such as acids, alcohols or phenols.

The second part of our algorithm is a method for calculating the desired properties of a molecule. Bytautas, Klein and Schmalz describe how the molecular graph of an acyclic hydrocarbon can be used to calculate its magnetic susceptibility via a substructural cluster expansion [14]. Similar formulas have been developed to estimate a wide variety of other properties [15,16]. There is no reason, however, why semiempirical, ab initio, density functional theory or molecular mechanics could not be used to compute these properties directly.

The final part of our algorithm is an optimization method that minimizes a multidimensional function. This function specifies not only the desired properties the final molecule should have but also the importance of these properties. Since the simulated annealing algorithm makes few assumptions about the function to be optimized, it is quite robust. In this paper we have used a simple version of this program. Several modifications to the simulated annealing algorithm have been described which may improve its convergence [17,18]. It should also be possible to adapt this algorithm to multiple processors [19]. At the moment we are exploring all of these options. The results of these tests will be reported in the near future.

Acknowledgements

I want to thank Dr. Doug Klein for many valuable discussions and Ms. Lindsay Jones for helping prepare the figures.

References

1. K.G. Joback and G. Stephanopoulos in: “Advances in Chemical Engineering” vol. 21, edited by G. Stephanopoulos and C. Han (Academic Press, San Diego, 1995).
2. P.M. Harper, M. Hostrup and R. Gani in: “Computer Aided Molecular Design: Theory and Practice”, edited by L.E.K. Achenie, R. Gani and V. Venkatasubramanian (Elsevier, Amsterdam, 2003).
3. A. Apostolakou and C.S. Adjiman in: “Computer Aided Molecular Design: Theory and Practice”, edited by L.E.K. Achenie, R. Gani and V. Venkatasubramanian (Elsevier,

Amsterdam, 2003).

4. P.R. Patkar and V. Venkatasubramanian in: “Computer Aided Molecular Design: Theory and Practice”, edited by L.E.K. Achenie, R. Gani and V. Venkatasubramanian (Elsevier, Amsterdam, 2003).
5. S. Sundaram, V. Venkatasubramanian and J.M. Caruthers in: “Computer Aided Molecular Design: Theory and Practice”, edited by L.E.K. Achenie, R. Gani and V. Venkatasubramanian (Elsevier, Amsterdam, 2003).
6. S. Kirkpatrick, C.D. Gerlatt Jr. and M.P. Vecchi, Science **220**, 671 (1983).
7. V. Cerny, J. Optim. Theory Appl. **45**, 41 (1985).
8. D. Vanderbilt and S.G. Louie, J. Comput. Phys. **56**, 259 (1984).
9. W.L. Goffe, G.D. Ferrier and J. Rodgers, J. Econometrics **60**, 65 (1994).
10. K.H. Hoffmann, A. Franz and P. Salamon, Phys. Rev. E **66**, 046706 (2002).
11. N. Metropolis, A.W. Rosenbluth, A.H. Teller and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
12. B. Hajer, Math. Oper. R **13**, 311 (1988).
13. Z.B. Zabinsky and G.R. Wood in: “Handbook of Global Optimization” vol. 2, edited by P.M. Pardalos and H.E. Romeijn (Kluwer, New York, 2002).
14. L. Bytautas, D.J. Klein and T.G. Schmalz, New J. Chem. **24**, 329 (2000).
15. W.J. Lyman, W.F. Reehl and D.H. Rosenblatt, “Handbook of Chemical Property Estimation Methods” (McGraw-Hill, New York, 1982).
16. L. Constantinou and R. Gani, AIChE Journal **40**, 1697 (1994).
17. P.J.M. Van Laarhoven and E.H.L. Aarts, “Simulated Annealing: Theory and Applications” (Reidel, Dordrecht, 1987).
18. G. Ruppeiner, J.M. Pedersen and P. Salamon, J. Phys. I **1**, 455 (1991).
19. K.W. Chu, Y. Deng and J. Reinitz, J. Comput. Phys. **148**, 646 (1992).