APPLICATIONS OF ARTIFICIAL INTELLIGENCE METHODS IN MACROMOLECULAR CHEMISTRY - COMPUTER-AIDED MOLECULAR DESIGN♦

Teodora Rusu[1*], Hugh Cartwright[2]

[1] "Petru Poni" Institute of Macromolecular Chemistry, Iasi, Romania
[2] Chemistry Department, Oxford University, Physical and Theoretical Chemistry Laboratory, Oxford, England

*Corresponding author: rusu_teodora@yahoo.com

Received: 15/01/2008
Accepted after revision: 26/02/2008

Abstract: This paper describes the implementation of the Tabu Search (TS) algorithm in concert with the Computer-Aided Molecular Design (CAMD) scheme. Although other optimization approaches have been applied to CAMD with properties predicted using group contribution techniques, the TS algorithm implemented with novel neighbor-generating operators and combined with property prediction via connectivity index-based correlations provides a powerful technique for generating lists of near-optimal molecular candidates for a given application. In addition, the tabu lists help TS search the solution space both in a diversified way, to cover the entire search space, and in an intensified manner, to locate the final solution precisely. Moreover, TS is able to locate a large number of near optimal solutions within a short time.

Keywords: *Artificial Intelligence methods, Macromolecular Chemistry, Computer-Aided Molecular Design, Tabu Search algorithm*

♦ Paper presented at the fifth edition of: "Colloque Franco-Roumain de Chimie Appliquée – COFrRoCA 2008", 25 – 29 June 2008, Bacău, Romania.

INTRODUCTION

As shown by Venkatasubramanian [1], the process of designing new molecules that have defined desirable properties is very important, especially in the formulation of such materials as polymers, polymeric composites, blends, paints and varnishes, refrigerants, solvents, drugs, and pesticides. The traditional approach to such a problem may involve a combinatorial search through a large number of potential candidate molecules. This approach is an expensive and time-consuming iterative process, during which the scientist or engineer proposes and synthesizes a compound, tests it for the desired properties, and then repeats the procedure if the desired properties are not found. Property prediction is usually based on either group contribution methods or topological indices. Of these methods, the group contribution approach has been the more widely reported [2-4], and incorporates a two level group contribution method which utilizes molecular structure information to estimate the physical and thermodynamic properties of pure components.

This paper describes the implementation of the Tabu Search (TS) algorithm in concert with the CAMD scheme. Although other optimization approaches have been applied to CAMD with properties predicted using group contribution techniques, the TS algorithm implemented with novel neighbor-generating operators and combined with property prediction via connectivity index-based correlations provides a powerful technique for generating lists of near-optimal molecular candidates for a given application. In addition, the tabu lists help TS search the solution space both in a diversified way, to cover the entire search space, and in an intensified manner, to locate the final solution precisely. Moreover, TS is able to locate a large number of near optimal solutions within a short time.

CAMD METHODOLOGY

Computer-aided molecular design (CAMD) is an attractive alternative to the traditional synthesize and test methodology. Venkatasubramanian showed that the CAMD approach requires the solution of two problems (Figure 1) [1]:

- the *forward* problem in which macroscopic properties are calculated given a candidate molecular structure;
- the *backward* problem in which a molecular structure is proposed that should display the desired properties.

A variety of methods including molecular models, group contribution methods, empirical models, correlations etc. can be used to address the forward problem; however, there has been considerably less progress in the backward problem which poses a real challenge for CAMD.

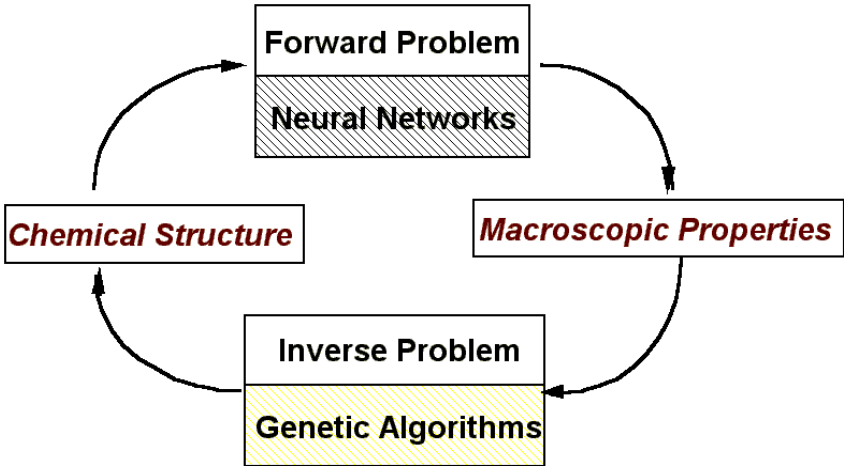

Figure 1. *Venkatasubramanian CAMD Scheme*

Venkatasubramanian proposed the use of Neural Networks for the *forward* problem and the genetic algorithm for the *inverse* problem. For complex molecules such as macromolecular compounds (polymers or copolymers), these approaches have limitations due to combinatorial complexity, nonlinear search spaces with local minima traps, difficulties in knowledge acquisition, difficulties in dealing with the nonlinear structure-property correlations, and problems in incorporating higher level chemical knowledge and reasoning strategies. Thus, there exists a critical need to explore alternate strategies for molecular design that can circumvent these problems.
Combinatorial and heuristic-based enumeration approaches have been reported by Gani and Brignole (1983) [5], and by Joback and Stephanopoulos (1989) [6]. Kier, Lowell, and Frazer (1993) [7] used a graph reconstruction approach to determine feasible molecular structures with bounded physical property values, while Vaidyanathan and El-Halwagi (1996) [8] described an interval analysis approach for the computer-aided synthesis of polymers and blends. CAMD problems have recently been formulated as mixed-integer linear/non-linear programming (MILP/MINLP) problems.
Stochastic optimization approaches have also been developed as alternate strategies for rigorous deterministic methods. Marcoulaki and Kokossis (1998) [9] described a simulated annealing (SA) and Wang and Achenie (2002) [10] presented a hybrid global optimization approach that combines the SA algorithms for several solvent design problems.

TABU SEARCH (TS) METHOD

Tabu search (TS) is a heuristic approach for solving combinatorial optimization problems by using a guided, local search procedure to explore the entire solution space while largely avoiding the danger of trapping in local optima. It differs from other stochastic optimization techniques by maintaining lists of previous solutions (usually termed "memory") that help guide the search process. These lists are useful for CAMD since they provide a direct method to track the discovery of near-optimal solutions
TS starts from an initial randomly generated solution. A set of neighbor solutions, $N(x)$, is constructed by modifying the current solution, x. The best one among them, x', is

selected as the new starting point, and the next iteration begins. Memory, implemented with tabu lists, is employed to escape from locally optimal solutions and to prevent cycling. At each iteration, the tabu lists are updated to keep track of the search process. This memory allows the algorithm to adapt to the current status of the search, so as to ensure that the entire search space is adequately explored and to recognize when the search has become stuck in a local region. Intensification strategies are employed to search promising areas more thoroughly, while at the same time diversification strategies are employed to broadly search the entire feasible region, further helping to avoid becoming stuck in local optima. Finally, aspiration criteria are employed to override the tabu lists in certain cases. Figure 2 presents a typical flow chart of the TS algorithm.

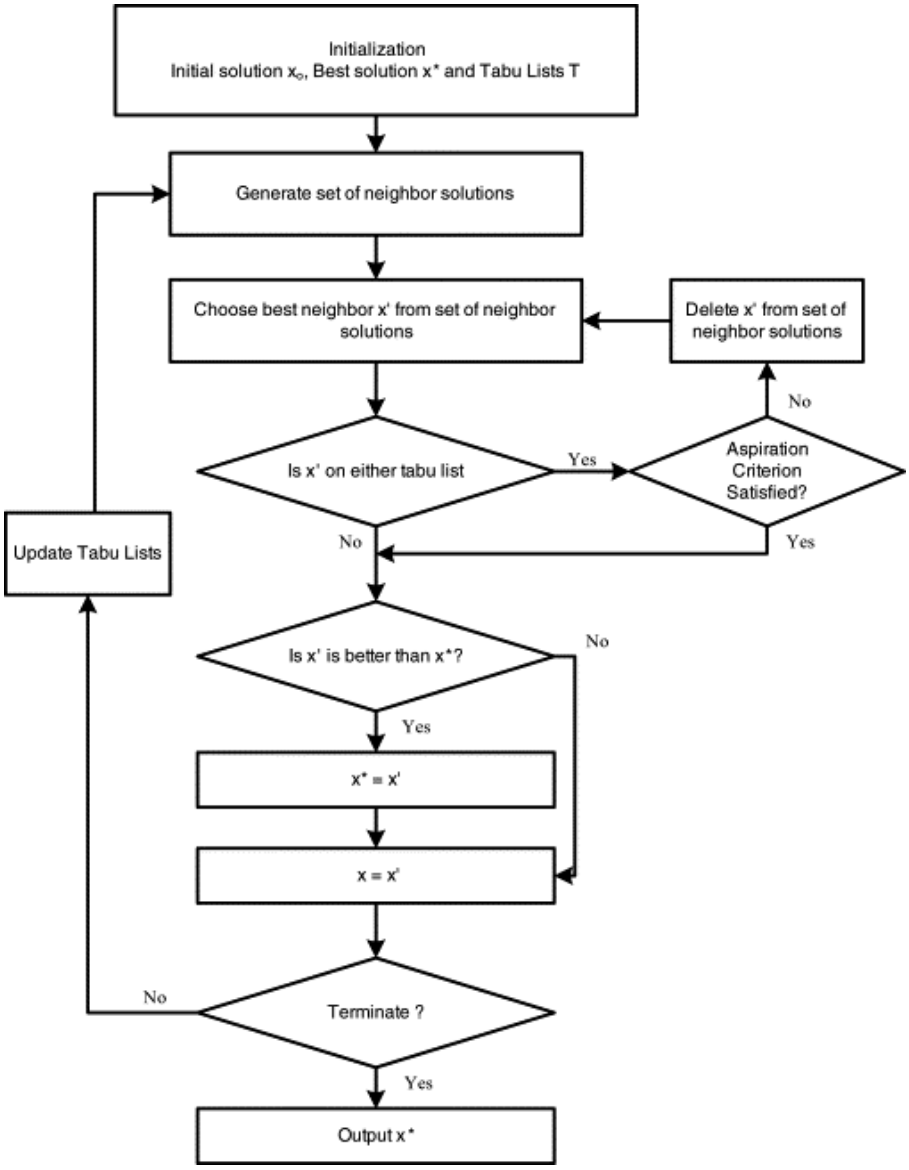

Figure 2. *Flow chart of the TS algorithm*

The standard TS algorithm developed by Glover & Laguna [11] can be adapted to represent the molecule efficiently. Each solution is a molecule consisting of a series of fully connected groups. Therefore, the initial solution is constructed by connecting basic groups together. In generating neighbor solutions, operators are also designed to handle basic groups. In the CAMD framework, a mainchain is defined as the list within a molecule with the largest number of groups. It is determined during the process of constructing a molecule. The length of sidechains (list of groups connecting to the mainchain) is constrained to be less than or equal to that of the mainchain. While the search is proceeding, several parts of the search space are classified as tabu according to either recency or frequency information. The length of the tabu lists determines how many solutions are forbidden during the search process.

If the length of the tabu lists is small, only a few solution areas are forbidden. This saves time in updating and maintaining tabu lists; however, it minimizes the advantage of the memory, since TS can only prevent cycling for a limited time and may spend a lot of time within areas that have already been visited. Consequently, the solution space cannot be fully searched and the probability of being trapped in local optima is high. On the other hand, if tabu lists are long, they keep many former moves from being revisited. This may also prevent promising areas from being explored more effectively, which results in final solutions with low precision.

As a starting point, the length of tabu lists should be set equal to the number of variables, N_{var}. If cycling in the process of finding solutions is observed in 10 tests, a larger value shall be assigned, increasing the length by [0.25 N_{var}] and testing for 10 additional trials. This should be repeated until cycling is eliminated. If new solutions are successively selected as the best neighbor for the initial 10 runs, assigning a smaller value to the length of the tabu lists will promote efficient searching around a specific solution. The length of the tabu lists should be reduced by [0.25 N_{var}] for each set of 10 runs, until the final solution is consistently found.

For small optimization problems, the recency-based tabu list can be sufficiently effective to guide the search process; however, for multi-dimension optimization problems with many local optima, the frequency-based tabu list will be helpful. Since this tabu list tracks the frequency with which a solution appears in a certain area, it is able to identify when the search becomes stuck so that the search can be restarted elsewhere.

Table 1 summarizes the computational requirements of several other methods that have addressed the same problem. Based on this comparison, it is clear that TS compares favorably with the most popular methods.

CONCLUSIONS

Although other optimization approaches have been applied to CAMD with properties predicted using group contribution techniques, the TS algorithm implemented with novel neighbor-generating operators and combined with property prediction via connectivity index-based correlations provides a powerful technique for generating lists of near-optimal molecular candidates. The Tabu lists help TS search the solution space both in a diversified way, to cover the entire search space, and in an intensified manner,

to locate the final solution precisely. Moreover, TS is able to locate a large number of near optimal solutions within a short time.

Table 1. *Comparison of TS with Other Approaches for the Global Minimum Search*

Method	Reference	N_{var}	CPU hr	Function Evaluations 10^5
Monte Carlo minimization	[13 a, b]	19*	2 – 3	1.0
Monte Carlo with minimization	[16 a]	24	1.5 – 4	-
Simulated annealing	[14]	24	2.5	2.5
Threshold accepting	[15]	24	1.5	2.0
Multicanonical algorithm	[17]	19*	6	1.5
Conformational space annealing	[18]	24	0.75	1.7
Diffusion equation	[19 a, b]	19*	0.33	-
Mean field theory	[20]	10*	1.6	-
αBB	[21]	24	1.3	3.9
TS algorithm	[12]	24	0.07	1.7

* - With this number of variables the corresponding methods reached an apparent global minimum.

TS is well suited to conformational searches, even if one works with discrete variables. To achieve this, one has to experiment with different values for the tabu parameters, such as the tabu distance, the length of the tabu list, the length of the long-term memory, the frequency criterion; in addition, it may be beneficial to fine tune the diversification and the intensification processes.

Acknowledgement

We are pleased to acknowledge support from NATO Research Grant CBP.EAP.CLG 982787.

REFERENCES

1. Venkatasubramanian, V., Chan, K., Caruthers, J.M.: Computer-aided Molecular Design Using Genetic Algorithm, *Computers & Chemical Engineering,* **1994**, **18** (9), 833-844;
2. Gani, R., Nielsen, B., Fredenslund, A.: A Group Contribution Approach to Computer-Aided Molecular Design, *AIChE Journal,* **1991**, **37** (9), 1318-1332;
3. Maranas, C.D.: Optimal Computer-aided Molecular Design: A Polymer Design Case Study, *Ind. Eng. Chem. Res.,* **1996**, **35**, 3403-3414;
4. Harper, P.M., Hostrup, M., Gani, R.: *A Hybrid CAMD Method in Computer Aided Molecular Design: Theory and Practice*, (Achenie, L.E.K., Gani, R., Venkatasubramanian, V. – Eds.), Elsevier, Amsterdam, **2003**, **2**, 122-169;
5. Gani, R., Brignole, E.A.: Molecular design of solvents for liquid extraction based on UNIFAC, *Fluid Phase Equilibria,* 1983, **13**, 331-340;

6. Joback, K.G., Stephanopoulos, G.: Designing molecules possessing desired physical property values, in: *Proceedings of the Foundations of Computer-Aided Process Design (FOCAPD)* (Siirola, J.J., Grossmann, I., Stephanopoulos, G. – Editors), July 12–14, **1989**, Snowmass, CO, CACHE-Elsevier, 363-387;
7. Kier, L.B., Lowell, H.H., Frazer, J.F.: Design of molecules from Quantitative Structure-Activity Relationship Model. 1. Information Transfer between Path and Vertex Degree Counts, *Journal of Chemical Information and Computer Sciences,* **1993, 33**, 142;
8. Vaidyanathan, R., El-Halwagi, M.M.: Computer-Aided Synthesis of Polymers and Blends with Target Properties, *Industrial and Engineering Chemistry Research,* **1996**, **35** (2), 627-634;
9. Marcoulaki, E.C., Kokossis, A.C.: Molecular design synthesis using stochastic optimization as a tool for scoping and screening, *Computers and Chemical Engineering,* **1998**, **22**, 11-18;
10. Wang, G., Milne, W.A.: Graph theory and group contributions in the estimation of boiling points, *Journal of Chemical Information and Computer Science,* **1994**, **34**, 1242-1250;
11. Glover, F., Laguna, M.: *Tabu Search*, Kluwer Academic Publishers, Boston, **1997**;
12. Morales, L.B., Garduño-Juárez, R., Aguilar-Alvarado, J.M., Riveros-Castro, F.J.: *Journal of Computational Chemistry*, **1999**, **21** (2), 147-156;
13. (a) Li, Z., Scheraga, H.A.: *Proc. Natl. Acad. Sci. USA,* **1987**, **84**, 6611; (b) Li, Z., Scheraga, H.A.: *J. Mol. Struct. (Theochem),* **1988**, **179**, 333.
14. Morales, L.B., Garduño-Juárez, R., Romero, D.: *J. Biomol. Struct. Dynam.*, **1991**. **8**, 721;
15. Morales, L.B., Garduño-Juárez, R., Romero, R.: *J. Biomol. Struct. Dynam.*, **1992**, **9**, 951;
16. (a) Nayeem, G.A., Vila, J., Scheraga, H.A.: *J. Comput. Chem.,* **1991**, **12**, 594; (b) Vásquez, M., Meirovitch, E., Meirovitch, H.: *J. Phys. Chem.,* **1994**, **98**, 9380;
17. Hansmann, U.H.E., Okamoto, Y.: *J. Comp. Chem.*, **1993**, **14**, 1333;
18. Lee, J., Scheraga, H.A., Rackovsky, S.: *J. Comp. Chem.*, **1997**, **18**, 1222;
19. (a) Kostrowicki, J., Piela, L., Cherayil, B.J., Scheraga, H.A.: *J. Phys. Chem.,* **1991**, **95**, 4113; (b) Kostrowicki, J., Scheraga, H.A.: *J. Phys. Chem.,* **1992**, **96**, 7442;
20. Olszewski, K.A., Piela, L., Scheraga, H.A.: *J. Phys. Chem.*, **1992**, **96**, 4672;
21. Androulakis, I.P., Maranas, C.D., Floudas, C.A.: *J. Global Opt.*, **1997**, **11**, 1;