CHEMICAL ENGINEERING TRANSACTIONS, Volume 3, 2003
Edited by Sauro Pierucci
 ISBN 88-900775-2-2

A Combinatorial Approach for Generating Environmentally Benign Solvents and Separation Agents

B. Holenda[1], A. Dallos[2], A.B. Nagy[1], F. Friedler[1] and L.T. Fan[3]

[1]Department of Computer Science and [2]Department of Physical Chemistry, University of Veszprém, Veszprém, Egyetem u. 10, Hungary H–8200

[3]Department of Chemical Engineering, Kansas State University, 105 Durland, Manhattan, Kansas 66506–5102, U.S.A.

A novel method based on the mixed-integer non-linear programming algorithm proposed herein resorts to complex multivariate methods. It estimates the mixture's properties by non-linear functions of atomic fragments and predicts the possible molecular structures of the desired solvents or separation agents. The algorithm systematically generates the collections of functional groups that can constitute structurally feasible molecules or compounds satisfying the constraints on the given properties. The applicability and efficacy of the method are demonstrated by designing selective agents for sustainable separation processes.

1. Introduction

Designing new molecules possessing desired properties is an important activity in the chemical and pharmaceutical industries, encompassing the design of various chemicals and materials such as polymers, blends, paints, solvents and drugs. Although a number of methods has been developed for this purpose, much remains to be done to devise new methods and improve available ones for efficiently generating candidate molecular structures for such compounds.

The traditional approach to this problem often requires a search involving a combinatorially large number of potential candidate molecules. It is expensive and iterative and requires that the target compound be hypothesized, synthesized and tested in light of the designed properties. Computer-aided molecular design (CAMD) is an attractive alternative to the traditional synthezis-and-test methodology. The CAMD requires the solution of two problems: the forward problem requires the computation of macroscopic properties for a given molecular structure, while the backward problem requires the identification of the appropriate molecular structure satisfying the desired properties. A variety of methods, including molecular modelling, group contribution methods, empirical modelling, and correlations, has been developed to address the forward problem; however, little progress has been made in solving the backward problem.

The available methods for solving the backward problem can be divided into two major classes. In the first class, structures are composed exhaustively, randomly or heuristically by resorting to expert systems (artificial intelligence) (Joback and Stephanopoulos, 1995; Venkatasubramanian *et al.*, 1996) from a given set of groups. The resultant compound is subsequently examined to determine if it is endowed with the specified properties. This "generate-and-test" methodology is usually capable of taking into account only a small number of feasible molecular structures of the compound of interest. While yielding promising results in some

applications, the chance of reaching the target structure by the strategy can indeed be small for any complex problem, e.g., that involving a large number of groups. In the second class, a mathematical programming method is applied to a problem in which the objective function expresses the "distance" to the target (Macchietto *et al.*, 1990). Since the method for estimating the properties of the structure generated, e.g., group contribution, is not sufficiently precise, the assessment of the results on the basis of this objective function may be precarious. While all these methodologies have certain appeal and advantages, they also suffer from some serious drawbacks. For complex and industrially relevant molecules, such drawbacks are attributable to combinatorial complexity, nonlinear search spaces with local minima-traps, and difficulties of incorporating higher level chemical knowledge and reasoning strategies. Consequently, a critical need exists to explore alternate strategies for molecular design that can circumvent these drawbacks.

The present work proposes a combinatorial approach, or method, for generating all feasible candidate molecular structures whose properties determined by group contributions fall within the given intervals. The final selection of the most appropriate structure or structures is carried out by further analysis of such candidate structures with available techniques.

2. Problem formulation

Suppose that the following information is given.

1. Set G of n groups of which a molecular structure can be composed;
2. The lower bounds, p_j's, and the upper bounds, P_j's ($j = 1, 2, \ldots, m$), of the properties to be satisfied;
3. The lower limit, l_i, and the upper limit, L_i ($i = 1, 2, \ldots, n$), for the number of appearances of group i in a molecular structure to be determined; and
4. Function f_k ($k = 1, 2, \ldots, m$) representing the value of property k estimated by the group contribution method as $f_k(x_1, x_2, \ldots, x_n)$

In statement 4, $x_1, x_2, \ldots, x_n$ are, respectively, the numbers of groups #1, #2,..., #n contained in the molecular structure or compound. For convenience, a functional group in set G with one bond is called the terminator, and that with three or more bonds, the brancher group.

The problem can be formulated as follows: Search all molecular structures formed from the given groups, #1, #2, ..., #n, whose numbers are $x_1, x_2, \ldots, x_n$, respectively, under the condition that the property constraints given below are satisfied; thus,

$$p_j \leq f_j(x_1, x_2, \ldots, x_n) \leq P_j \quad (j = 1, 2, \ldots, m) \tag{1}$$

Throughout this work, the molecular structures are expressed by simple connected graphs whose vertices and edges represent, respectively, the functional groups from set G and the associated bonds. As a result, the set of such connected graphs needs to be generated from the set of functional groups G satisfying the property constraints, by considering multiplicities of the functional groups. In any of the conventional generate-and-test approaches, all or some of the connected graphs, i.e., molecular structures, are generated from the available functional groups and then tested against the property constraints. This usually yields an excessively large number of structures.

3. Algorithmic generation of candidate molecules

Let us consider function $f_k(x_1, x_2, \ldots, x_n)$, which is monotonous in x_i $(i = 1, 2, \ldots, n)$. A backtracking algorithm is adopted to generate all the candidate molecules satisfying the property and structural constraints. This can be represented by an enumeration tree in which each node of the tree represents a partition of the search space defined by constraints (3); this partition is called a partial problem.

Suppose that variables $x_1, x_2, \ldots, x_k$ $(k \leq n)$ are fixed a priori at an intermediate phase of the procedure. Then, the problem is treated according to the following two cases.

Case 1: $k \leq n-1$. We compute the value of $f_j(x_1, x_2, \ldots, x_{k-1}, x'_k, x'_{k+1}, \ldots, x'_n)$, where $x'_i = L_i$ $(i = k, k+1, \ldots, n)$, if the increasing number of group $\#i$ increases the value of f_j; otherwise, $x'_i = l_i$ $(i = k, k+1, \ldots, n)$. Hence, we have an upper bound on the value of the property. If $f_j(x_1, x_2, \ldots, x_{k-1}, x'_k, x'_{k+1}, \ldots, x'_n) < p_i$, i.e., if it is lower than the lower bound of the property constraint, then a solution does not exists for this subproblem; therefore, the problem is not branched into subproblems. Similarly, we can compute a lower bound which is $f_j(x_1, x_2, \ldots, x_{k-1}, x'_k, x'_{k+1}, \ldots, x'_n)$, where $x'_i = L_i$, if the increasing number of group $\#i$ reduces the value of f_j; otherwise, $x'_i = l_i$ $(i = k, k+1, \ldots, n)$. If $f_j(x_1, x_2, \ldots, x_{k-1}, x'_k, x'_{k+1}, \ldots, x'_n) > P_j$, the problem is also not branched into subproblems.

Case 2: $k = n$, i.e., the partial problem belonging to a leaf of the tree. For this case, a test must be performed by simple substitution to determine the following constraints for $x_1, x_2, \ldots, x_n$.

Condition 1.

$$p_j \leq f_j(x_1, x_2, \ldots, x_n) \leq P_j$$

Condition 2. If the molecule specified by $x_1, x_2, \ldots, x_n$ contains functional groups with different types of bonds, e.g., single and double bonds, then, there must be a group contained in the molecule, which has at least two different types of bonds, and each type belongs to at least one functional group in the molecule containing another type of bonds.

Condition 3. The number of bonds identical in type is even.

Condition 4.

$$\sum_{i \in I_1} x_i - \sum_{i \in I_2} x_i$$

is an even number and not less than -2. If the partial problem under consideration is proven to be valid by the test, it represents a feasible partition of the candidate molecules. In other words, each molecular structure composed of $x_1, x_2, \ldots, x_n$ numbers of functional groups $\#1, \#2, \ldots, \#n$, respectively, satisfies the property constraints.

Suppose that three functional groups OH, CH_3CO and CH_2 are available to compose molecular structures satisfying the following constraints on the logarithm of the octanol-water partition coefficient ($\log P^{ow}$);

$$0 \leq \log P^{ow} \leq 0.5$$

The value of $\log P^{ow}$ is one of the key parameters for predicting the environmental fate of organic chemicals (bioaccumulation, soil adsorption, etc.) and can be predicted from the limiting activity coefficients of a compound in the aqueous (w) and organic (o) phases. The octanol-water partition coefficient of the compound is calculated by the UNIFAC group-contribution method as follows:

$$P^{ow}_{i,unifac} = \frac{x^w_w M_v + x^w_o M_o}{x^o_w M_w + x^o_o M_o} \cdot \frac{d^o_{sol}}{d^w_{sol}} \cdot \frac{\gamma^{w,\infty}_{i,unifac}}{\gamma^{o,\infty}_{i,unifac}}, \tag{2}$$

where $\gamma^{j,\infty}_{i,UNIFAC}$ is the limiting activity coefficient of compound i in phase j; M_k, the molar mass of the solvent k; x^j_k, the mole fraction of solvent k in phase j; and d^j_{sol}, the density of the saturated phase j in equilibrium.

It is assumed that OH and CH_3CO may appear at most twice, and CH_2 may appear at most once. Moreover, we know that the increasing number of OH's reduces the octanol-water coefficient, and the increasing number of CH_3CO and CH_2 magnifies the octanol-water coefficient. The enumeration tree (Figure 1) illustrates the working of the algorithm. First, the problem is branched into 3 subproblems according to the number of functional group OH. Incorporating functional group OH in the molecule significantly decreases the octanol-water coefficient, even if functional groups CH_3CO and CH_2 are incorporated into the molecule to their maximal number (2 and 1, respectively). Consequently, the lower bound of the property constraint is not reached, i.e., there are infeasible partial problems represented by X's in the enumeration tree. Second, the only feasible partial problem is branched into 3 subproblems according to the number of functional group CH_3CO. After calculating the lower and upper bounds for these partial problems, we see that each of the new partial problems is feasible. Finally, all feasible partial problems are divided into two subproblems according to the number of functional group CH_2. Since they belong to leaves of the tree, they should be examined in view of conditions 1 through 4. Solid circles at certain leaves of the tree represent the partial problems where the conditions are not fulfilled. This leaves one feasible partial problem satisfying the property and structural constraints, which contains 2 CH_3CO and 1 CH_2 functional groups.

4. Application

The proposed combinatorial methodology is applied to two examples: selection of a single compound matching specified values of the octanol-water partition coefficient and selection of solvents for extractive destillation. Both selections are based on mixture properties, namely, the activity coefficients of the components, which are not linear, but monotone functions of the contributions of their atomic groups as estimated by the UNIFAC method.

The first example is concerned with the determination of compounds with a general structure of $X(CH_2)_nY$ ($0 \leq n \leq 4$), each of which has the octanol-water partition coefficient in the range of : $-10 \leq \log P^{ow} \leq 1.0$. In this example, groups X and Y for the structure generation include OH, CH_3CO, CH_3O, CH_2CH, and C_6H_5. The feasible molecular structures that satisfying the constraint at $T = 298$ K are listed in Table 2 with their $\log P^{ow}$ values highlighted in bold.

The second example is concerned with the selection of solvents for separating the components of the azeotrope-mixture of benzene (1) and cyclohexane (2) by

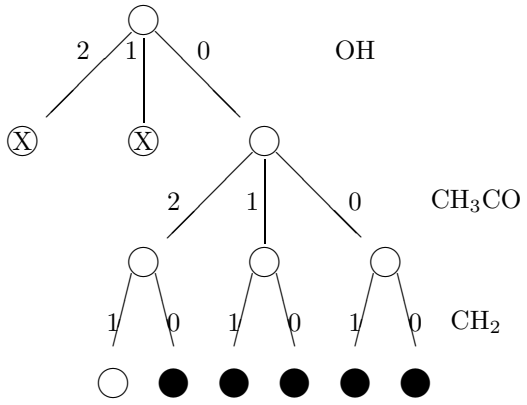

Figure 1: *Enumeration tree for the illustration.*

Table 1: *List of selected model compounds in example 1 with their estimated* $\log P^{ow}$ *values*

X	Y	n=0	n=1	n=2	n=3	n=4
OH	OH	**-1.7718**	**-1.3381**	**-0.9030**	**-0.4687**	**-0.0343**
CH_3CO	CH_3CO	**-0.9079**	**-0.4734**	**0.0389**	**0.3954**	**0.8298**
CH_3COO	CH_3COO	**-0.0165**	**0.4178**	**0.8521**	1.2865	1.7209
CH_3O	CH_3O	**0.2965**	**0.7309**	1.1652	1.5996	2.0340
CH_2CH	CH_2CH	1.8709	2.3052	2.7397	3.1737	3.6084
C_6H_5	C_6H_5	5.3298	5.4673	5.7643	6.1986	7.0674
C_6H_5	OH	1.7790	2.2134	2.6473	3.0821	3.5165
C_6H_5	CHO	2.8909	3.3255	3.7597	4.1943	4.6287
C_6H_5	CH_3CO	2.2095	2.6454	3.0795	3.5141	3.9486
C_6H_5	CH_3COO	2.6560	3.0909	3.5254	3.9600	4.3942
C_6H_5	CH_3O	2.8129	3.2474	3.6819	4.1163	4.5507
C_6H_5	CH_2CH	3.600	4.0347	4.4692	4.9035	5.3379

extractive distillation. The solvent structures are evaluated in terms of the selectivity. The selectivity can be estimated by calculating the ratio of the limiting activity coefficients of the components to be separated, 1 and 2, in solvent S as follows:

$$S^{\infty} = \frac{\gamma^{S,\infty}_{1,UNIFAC}}{\gamma^{S,\infty}_{2,UNIFAC}}. \tag{3}$$

For simplicity, only monosubstituted benzene- and cylohexane-derivatives have

been chosen as candidate solvents in the example. This gives rise to C_6H_5–X and C_6H_{11}–X with the pre-selected first-order groups, including NO_2, CH_3O, Cl, NH_2, OH, and CH_3 for X, and the feasibility criterion, $2.5 \leq S^{\infty} \leq 10.0$. Table 2 lists the solvents satisfying this selectivity constraint at $T = 340$ K as well as infeasible candidates.

Table 2: *List of potential solvents in example 2 with the estimated selectivity values*

		X	S^{∞}
5ACH	1AC	NO_2	**3.2881**
5ACH	1AC	Cl	1.3651
5ACH	1AC	NH_2	**2.9839**
5ACH	1AC	OH	**2.6295**
5ACH	1AC	CH_3	1.3546
$5CH_2$	1CH	$1CH_3$	1.4110
$5CH_2$	1CH	1OH	1.3231
$5CH_2$	1CH	$1CH_3$	0.7300
$5CH_2$	1CH	1Cl	1.0578
$5CH_2$	1CH	$1NH_2$	1.2930

5. Acknowledgement

The authors wish to acknowledge financial support received from the Hungarian Scientific Research Foundation (OTKA T35220).

References

Bärbel K., J. Gmehling, U. Onken, 1979, Ber. Bunsenges Phys. Chem., **83**, 1133–1136, Auswahl von Lösungmitteln für die Extraktiv-Rektifikation mittels vorausberecheneter Gleichgewichtsdaten

Briggs, B. G., 1981, J. Agr. Food Chem., **29**, 1050–1059, Theoretical and Experimental Relationship Between Soil Adsorption, Octanol-water Partition Coefficients, Water Solubilities, Bioconcentration Factors and the Parachor

Dallos A., J. Liszi, 1995, J. Chem. Thermodyn., **27**, 447–448, (Liquid + liquid) equilibria of (octan-1-ol + water) from 288.15 K to 323.15 K.

Fredenslund, Aa., R.L. Jones, J.M. Prausnitz, 1975, AIChE J., **21**, 1086, Group-Contribution Estimation of Activity Coefficients of Nonideal Liquid Mixtures

Friedler, F., L. T. Fan, L. Kalotai, A. Dallos, 1998, Comp. Chem. Eng., **22**, 809–817, A combinatorial approach for generating candidate molecules with desired properties based on group contribution

Joback, K. G., Stephanopoulos, G., 1995, Advances in Chemical Engineering, **21**, 257–311, Searching spaces of discrete solutions: The design of molecules possessing desired physical properties

Nolker K., M. Roth M., 1998, Chem. Eng. Sci., **53**, 2395, Modified UNIFAC parameters for mixtures with isocyanates

Sanster, J., 1989, J. Phys. Chem. Ref. Data, **18**, 1111–1229, Octanol-water Partiton Coefficients of Simple Organic Compounds

Venkatasubramanian, V., Chan., K., Caruthers, J., 1994, Comp. Chem. Eng., **16**, 833–844, Computer-aided molecular design using genetic algorithms