Evolutionary Design of Extractants for the Separation of Organic Compounds from Aqueous Streams by Liquid Extraction

Juan Carlos Serrato
National University of Colombia
Chemical Engineering Laboratory
Bogotá, Colombia
057-1-3165000 Ext 14301
jcserratob@unal.edu.co

Jonatan Gómez
National University of Colombia
System Engineering Department
Bogotá, Colombia
057-1-3165000 Ext 14029
jgomezpe@unal.edu.co

Luis A. Caicedo
National University of Colombia
Chemical Engineering Laboratory
Bogotá, Colombia
057-1-3165000 Ext 14053
lacaicedom@unal.edu.co

ABSTRACT
Industry poses different separations problems for diverse compounds recovery. Separation stage is very important on the product cost, especially for the biotechnological products. Liquid-liquid extraction is a widely used separation operation; however it requires separation agents that must fulfill conditions such as low toxicity, high distribution coefficient and selectivity, etc. Therefore, different approaches have been developed to tackle this problem, being computer aided molecular design the most promising alternative. It is an optimization process looking for the best group combination to successfully perform some objective in a given process, including some physical, feasibility and process restrictions. To solve this problem we present a new genetic algorithm that does not show some limitations present in early works.

Categories and Subject Descriptors
J.6 [**Computer-aided Engineering**]: Computer Applications – Computer-aided engineering – Computer-aided design.

General Terms
Algorithms, Experimentation.

Keywords
Genetics algorithms, molecular design, Unifac, CAMD, extraction agents.

1. INTRODUCTION
Chemical and biochemical processes are composed basically by pretreatment, synthesis and product recovery operations. The recovery stage is an important part of the process, due to cost and technical difficulties. For example, separation operations in biotechnological processes may be as much as 50% of the total process cost. Streams with low organic concentration are among the most common separation problems for industry. These are a big challenge due to the low available potentials for mass transfer and the great amount of contaminants in the streams under treatment.

GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA.
ACM 978-1-60558-131-6/08/07.

Liquid-liquid extraction is a simple and cheap separation operation ; it is used in the recovery of diverse products such as phenols, organic acids (lactic, citric, propionic, butyric, etc), among others. However, that operation requires separation agents with some conditions such as low toxicity, high distribution coefficient and selectivity, etc. These limitations impose a great complexity on the molecule type to be used; therefore, this problem has been addressed using different approaches. The first approach is the knowledge-based selection methods such as the search on databases; it is an easy method but it is restricted to the relatively low number of compounds on these. Another approach is the enumeration method, its principal problem being the combinatorial explosion caused by the group contribution methods used (GANI & BRIGNOLE, 1983).

The most promising alternative and then the most used, is the computer aided molecular design (CAMD). This inverse engineering methodology has been successfully used by some authors for the generation of different molecules as refrigerants (Duvedi & Achenie, 1997), polymers and polymer blends (Buxton*, et al.*, 1999), gas absorption solvents (Stefanis*, et al.*, 1996), liquid-liquid extraction solvents (Marcoulaki & Kokossis, 2000), blanket wash solvents (Achenie & Sinha, 2004), crystallization solvents (Karunanithi*, et al.*, 2006), reaction and supercritical solvents (Gani*, et al.*, 2005), among others. The molecular design could be understood as a optimization process: $\min/\max_{x,j} J(x,y,z,...)$, where $J(x,y,z,...)$ represents the molecule performance and $(x,y,z,...)$ are the contributions of the groups constituting the molecule. The design has different kinds of constraints: physical, structural properties, etc, defined as function of the structural groups. In other words, the molecular design methodology tries to find the best group combination (the best molecule) to perform satisfactorily some objective in a given process. The process imposes some restrictions on the molecule, for example: boiling point, melting point, vaporization latent heat, octanol-water distribution coefficient, cost, etc.

The CAMD methodology is based on different property estimation methods coupled with an optimization method. There are four alternatives for properties estimation: group contribution methods (the more used model is Unifac), quantum chemistry based methods (Sheldon*, et al.*, 2006), group contribution methods with some characteristics of quantum calculations as Cosmo-RS (Klamt*, et al.*, 2001), and others methods (Meniai & Newsham, 1999). A large

number of the works published has used group contribution methods, in spite of some limitations such as the insufficient precision in properties estimation, no isomers distinction and no inclusion of proximity effects in the model. Although molecular simulation based on quantum chemistry can give solutions to the limitations of the group contribution methods aforementioned, such simulation is not straightforward, and requires a lot of computational resources (time and memory). Hybrid methods combine the group contribution methods low level of complexity and the quantum calculations rigorousness.

The fundamental idea of the group contribution methods is that a molecular group makes the same contribution to the evaluated property regardless the type of molecule. Therefore, the summation of the group's contributions constituting the molecule, gives as result the property on evaluation. Different group contribution methods have been used for properties estimation.

The optimization problem involved in molecular design is formulated as a mixed integer non linear programming model (MINLP) (Stefanis, *et al.*, 1996, Hostrup, *et al.*, 1999), but some authors use non linear programming. To solve this problem, different methods are used: deterministic ones such as branch and bound (Xu & Diwekar, 2005), interval analysis (Achenie & Sinha, 2003), outer approximation algorithm (Sheldon, *et al.*, 2006), generalized benders decomposition (Karunanithi, *et al.*, 2005) and stochastic methods such as simulated annealing (Marcoulaki, *et al.*, 2000) and genetic algorithms (Van Dyk & Nieuwoudt, 2000, van Dyk & Nieuwoudt, 2002, Lehmann & Maranas, 2004, Xu & Diwekar, 2005).

Due to the high level of complexity, deterministic methods not always find the global optimum and are very susceptible to get trapped in local minima. In contrast, stochastic methods have a greater possibility finding the global optimum and their form of use is very flexible. These are the reasons that lead to the use of genetic algorithms in this work.

Thanks to the advantages of the genetic algorithms over the other stochastic methods, authors as van Dyk (van Dyk & Nieuwoudt, 2002), Venkatasubramanian (Venkatasabrumanian, *et al.*, 1994) and Xu (Xu & Diwekar, 2005) has used them for solvent design. However, their implementations have some limitations: a) the molecules representation is diverse and often allow the construction of molecules with a low complexity (because of the representation and the low maximum number of groups that form the molecule); additionally, almost in all cases the representation permits multiple molecules b) there is a non uniform use of the restrictions of the problem, especially the structural feasibility restrictions, and it is unknown the effect of the problem formulation on the results, c) some researchers use gaussians or sigmoidals fitness functions that need an a priori knowledge of the function, limiting its application d) in spite of the genetic operators used are similar, their probability of use in the algorithm is a parameter to be adjusted by the user, no warranting convergence nor optimality, and e) the use of different property estimation methods makes difficult the results evaluation and introduces a high uncertainty level in the problem that could change the results.

This work presents an alternative software that aims to alleviate the aforementioned difficulties.

2. SOLVENT DESIGN

The CAMD software for solvent design is divided in the following way: a routine to construct molecules structurally feasible, another to estimate properties, and a final that encompasses the others and makes the optimization. Every part are described in the following lines.

2.1 Molecule's Construction

The required molecules to initialize the software are constructed in the following way: once the user has defined how many groups are allowed in a molecule (Np), the software chooses the number of terminal groups randomly. That number varies between 1 and Np/2. Terminal groups have a valence of 1. Then, other group with a valence higher than 1 is selected randomly and is connected to some terminal groups according to its valence, leaving a number of bonds free. These connected groups are joined with other random selected groups up to the molecule valence is zero.

For aromatic and cyclic compounds a special routine is used. The software determines the cycle size randomly and it is constructed using only special cyclic groups. For aromatics construction the cycle size is fixed to 6 and it is constructed using only aromatics groups. When appears a cyclic-aliphatic or aromatic-aliphatic group, the aliphatic branch is constructed by the same random selection used before and it is finished with a terminal group.

In all cases the maximum number of functional groups is restricted to three, considering easy of synthesis heuristic. Constructed molecules must fulfill the Odele-Machietto restrictions (GANI & BRIGNOLE, 1983).

2.2 Properties Estimation

The principal properties needed for the estimation of solvent performance are distribution coefficient and selectivity. Boiling point, Gibbs free energy and density are also included in the design.

2.2.1. Activity coefficient

Distribution coefficient and selectivity are functions of activity coefficients. These coefficients are calculated using the Unifac-Dortmund model (Gmehling, *et al.*, 1993, Gmehling, *et al.*, 2002). In this method, the activity coefficients are estimated as the sum of residual and combinatorial contributions.

$$\ln \gamma_i = \ln \gamma_i^C + \ln \gamma_i^R$$

The combinatorial contribution considers compounds with very different sizes and forms, and it is calculated as:

$$\ln \gamma_i^C = 1 - V_i' + \ln V_i' - 5q_i\left(1 - \frac{V_i}{F_i} + \ln\left(\frac{V_i}{F_i}\right)\right)$$

The parameters V'_i, V_i, q_i, r_i y F_i are calculated by the following equations and are functions of the relative van der Waals volumes and areas (R_k y Q_k) for the different groups:

$$V_i' = \frac{r_i^{3/4}}{\sum_j x_j r_j^{3/4}} \qquad V_i = \frac{r_i}{\sum_j x_j r_j}$$

$$r_i = \sum \upsilon_k^{(i)} R_k$$

$$q_i = \sum \upsilon_k^{(i)} Q_k$$

$$F_i = \frac{q_i}{\sum_j x_j q_j}$$

The residual contribution is calculated as:

$$\ln \gamma_i^R = \sum_k \upsilon_k^{(i)} (\ln \Gamma_k - \ln \Gamma_k^{(i)})$$

$$\ln\Gamma_k = Q_k\left(1-\ln\left(\sum_m \theta_m \Psi_{mk}\right)-\sum_m \frac{\theta_m \Psi_{km}}{\sum_n \theta_n \Psi_{nm}}\right)$$

The group area fraction (Φ_m) and the group mole fraction (X_m) are given by the following equations:

$$\theta_m = \frac{\theta_m X_m}{\sum_n \theta_n X_n}$$

$$X_m = \frac{\sum_j \upsilon_m^{(j)} x_j}{\sum_j \sum_n \upsilon_n^{(j)} x_j}$$

The group interaction parameters are temperature dependent by the following expression:

$$\Psi_{nm} = \exp\left[-\frac{\left(a_{nm} + b_{nm}T + c_{nm}T^2\right)}{T}\right]$$

The relative van der Waals volumes and areas, and some group interaction parameters can be founded in literature (Weidlich & Gmehling, 1987, Gmehling, *et al.*, 1993, Gmehling, *et al.*, 2002).

2.1.2 Other estimated properties

Boiling point, density and Gibbs free energy are used as restrictions in the design process. For its estimation the last version of the GCVOL method for density (Ihmels & Gmehling, 2003) and the Constantinou and Gani methods for boiling point and Gibbs free energy are used. To include these restrictions in the design, the method of fitness function penalization is used.

2.3 Genetic Algorithm

For the solution of the problem, an adaptive genetic algorithm called HAEA [42] was used. Genetic search methods have their basis in Darwinian models of natural selection and evolution. Introduced by Holland, the general idea behind genetic algorithms is the evolutionary creation of a new population of individuals from an earlier generation through genetic processes, such as crossover and mutation, and by passing on the better offsprings to the next generation. In general, individuals that are better adapted to their environment will have a better chance of survival and thus pass on their genetic material to the succeeding generations. This approach is expected to lead to generations that become more and more fitted through evolution thus achieving the desired design objective.

2.3.1. Fitness Function

Distribution coefficient is the fundamental property for the study of the liquid-liquid extraction; it shows how a compound is distributed between two phases. However, a solvent with good extraction capacity shows low selectivity, and the inverse relation is also true. Therefore in this work, the fitness function for solvent design is the product of the distribution coefficient and the selectivity. Selectivity is calculated with the equation suggested by Pretel [41]:

$$\beta = \frac{\gamma_{B,S}^{\infty}}{\gamma_{A,S}^{\infty}} \frac{MW_A}{MW_B}$$

The distribution coefficient is calculated by the following equation:

$$k = \frac{\gamma_{A,B}^{\infty}}{\gamma_{A,S}^{\infty}} \frac{MW_B}{MW_S}$$

Both properties are functions of activity coefficients and molecular weights. The former property is calculated by the Unifac method described above.

2.3.2. Representation

Every population individual or molecule is represented in two ways, genotype and phenotype. Molecule genotype is a tree representation, where every node represents a group and the branches, the bonds of this group with others groups, it is used for the genetic operators. Phenotype is an array of groups of the same molecule; it is employed for properties estimation. For example, lactic acid whose structure is CH3CH(OH)COOH, is represented by its genotype (figure 1):

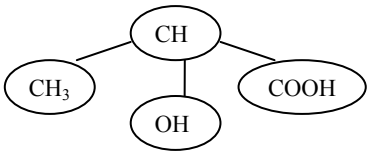

Figure 1. Lactic acid representation

Its phenotype is given by the array (1, 3, 15, 43), where the numbers are the corresponding groups codes (Gmehling, *et al.*, 1993).

2.3.3. Population

An initial population of 100 individuals was used. The molecules have variable length and a maximum length of 30 groups. In order to achieve a good diversity of the initial population, 10 % of the molecules created are cyclic, 10 % aromatics, and the others aliphatic.

2.3.4. Reproduction

For the generation of new individuals four genetic operators were used: mutation, crossover, group insertion y deletion. The

operator's probabilities were adjusted with the problem using the adaptive algorithm HAEA.

2.3.4.1. *Mutation*

Initially the program chooses a parent individual randomly, then choose the mutation point and the group to insert according to the residual valence of the fragment (randomly) and make the interchange. Figure 2 shows an example.

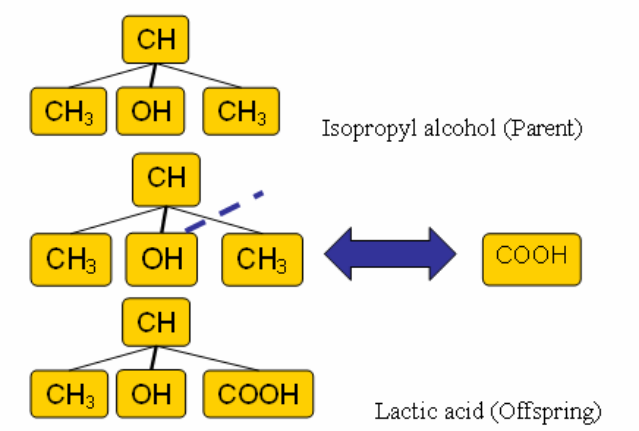

Figure 2. Molecule's mutation example.

2.3.4.2. *Cross*

To do a cross, the program chooses randomly two parent individuals by their fitness and two operation points, one for molecule. Then it makes the fragment interchange, to construct the new molecules. An example is given in figure 3:

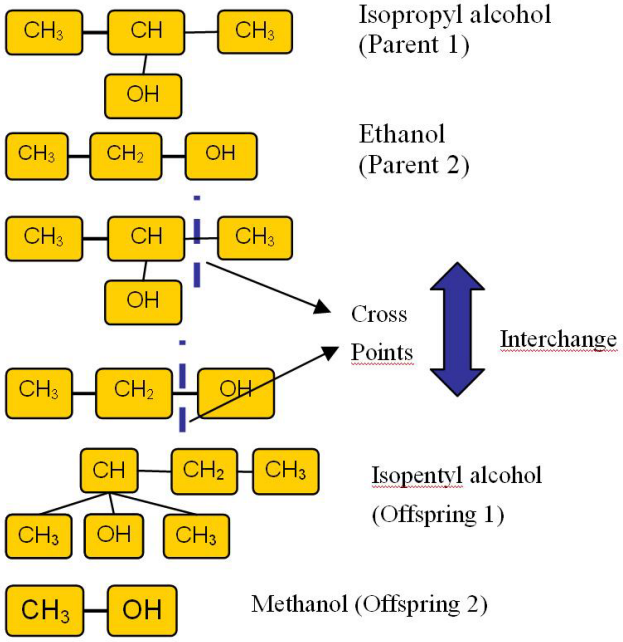

Figure 3. Molecules cross example.

2.3.5. *Selection*

The individual's selection was performed by tournament.

3. RESULTS AND ANALYSIS

The results depicted in figures 5 to 8 are the average of 50 runs of the solvent design problem. As model problem, the design of an extractant of acetic acid from an aqueous solution was done.

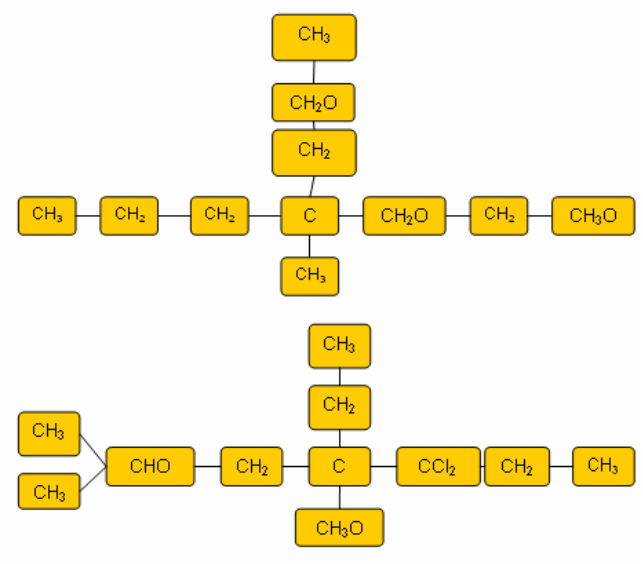

Figure 4. Best two molecules designed

The designed molecules with the best fitness contain esters, ethers or chloride groups. These results are in agreement with the results of Wang and Achenie (Wang & Achenie, 2002), Kim Diwekar and Tomazi (KIM, *et al.*, 2004) (Kim & Diwekar, 2002), and are oxygenated groups such as the groups suggested by Harper (Harper, *et al.*, 1999).

Due to our CAMD software does not include restrictions on the design; our results are large molecules, compared to methyl dimethyl ester and diisobutyl ketone reported previously. However, our molecules have the right groups.

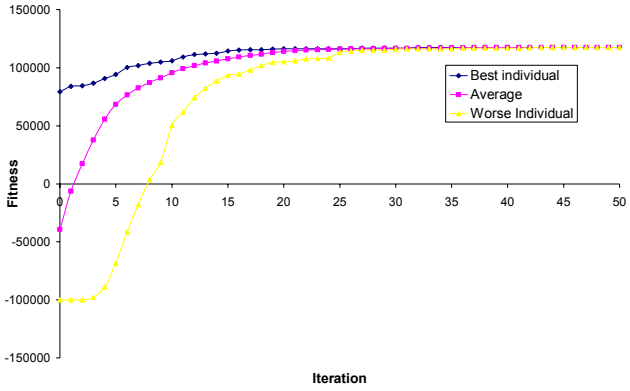

Figure 5. Fitness Evolution

Figure 5 shows fast convergence of the proposed genetic algorithm HAEA. It needs less than 30 iterations to find a local

optimum. For the best individual the algorithm get an improvement close to 50% in the fitness function.

Good molecule fitness implies a good distribution coefficient and good selectivity. As it was mentioned above, these properties are inverse related. A good extraction coefficient allows a good extraction capacity, it is very favorable for the unit operation due to less extraction agent consumption. A high selectivity is beneficial for extraction operation since it makes easier the next separation operations in the process, saving money.

The found fitness value of 120000 is very high, compared to values such as 12,43 for the selectivity and 0,66 for the distribution coefficient for methyl dimethyl ester, and 20,3 for the selectivity and 0,28 for the distribution coefficient for diisobutyl ketone. However, the lack of restrictions could explain this fact.

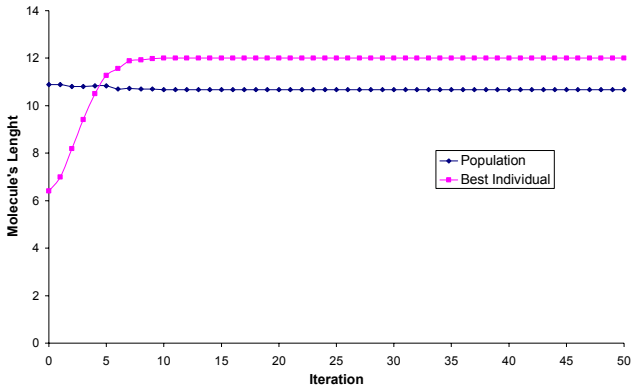

Figure 6. Individual Length Evolution

The optimum length of the designed molecule is close to 12 groups. Figure 6 shows how algorithm makes bigger the best molecules and stabilizes the population length in a value close to 10 groups.

For a group contribution method, a higher number of groups imply a higher value of the property on evaluation. A long molecule is difficult to synthesize, thus we will introduce in future works, restrictions on the molecule design limiting its size.

A molecule with 12 groups is relatively large compared to other results founded in the literature. We hope that the restrictions introduction could be an effective way to design more realistic molecules.

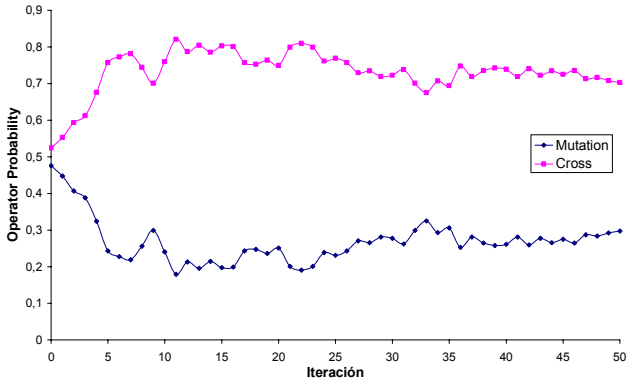

Figure 7. Population Operators Probabilities Evolution

The algorithm HAEA is an adaptive genetic algorithm, thus operators probabilities are adjusted by it. Figures 7 and 8 show similar trends, at the run's beginning the algorithm cross individuals as the preferential operator and the probability mutation is low, allowing good exploration of the search space. Around the tenth iteration cross probability diminish progressively, and mutation probability is a little higher for every iteration, allowing exploiting on the found optimum.

We are introducing more operators in the genetic algorithm such as group addition and deletion. We hope that their introduction could make more efficient the search process.

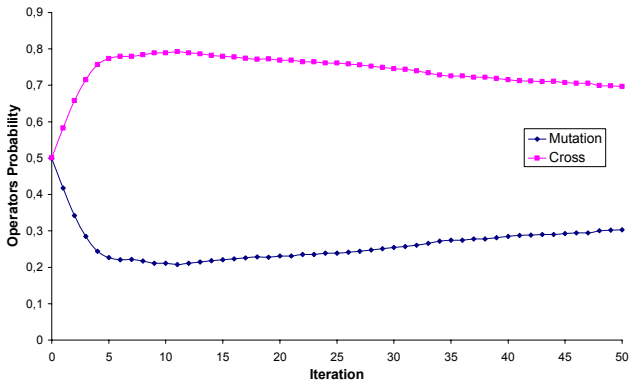

Figure 8. Operators Probabilities Evolution for the Best Molecule

4. CONCLUSIONS

Due to the high level of complexity of molecular design, genetic algorithms are a valuable tool for the CAMD.

The designed molecules for the extraction of acetic acid are in agreement with literature results.

The adaptive genetic algorithm HAEA is a promising tool for the computer aided molecular design.

We need to incorporate some restrictions in the molecular design to get more realistic results.

5. ACKNOWLEDGMENTS

The authors are very grateful for the economic support from the "División de investigación sede Bogotá (DIB)" from National University of Colombia and with Nicolás Moreno for his help with programming.

6. REFERENCES

[1] Achenie LEK & Sinha M (2003) Interval global optimization in solvent design. *Reliable Computing* **9**: 317-338.

[2] Achenie LEK & Sinha M (2004) The design of blanket wash solvents with environmental considerations. *Advances in Environmental Research* **8**: 213-227.

[3] Buxton A, Livingston AG & Pistikopoulos EN (1999) Optimal Design of Solvent Blends for Environmental Impact Minimization. *AIChE Journal* **45**: 817-843.

[4] Duvedi A & Achenie LEK (1997) On the design of environmentally benign refrigerant mixtures: a mathematical programming approach. *Computers chem. Engng* **21**: 915-923.

[5] GANI R & BRIGNOLE EA (1983) MOLECULAR DESIGN OF SOLVENTS FOR LIQUID EXTRACTION BASED ON UNIFAC. *Fluid Phase Equilibria* **13**: 331-340.

[6] Gani R, Jiménez-González C & Constable DJC (2005) Method for selection of solvents for promotion of organic reactions. *Computers and Chemical Engineering* **29**: 1661–1676.

[7] Gmehling J, Li J & Schiller M (1993) Modified UNIFAC model. 2. Present parameter matrix and results for different thermodynamic properties. *Industrial and Engineering Chemistry Research* **32**: 178-193.

[8] Gmehling J, Wittig R, Lohmann J & Joh R (2002) A modified UNIFAC (Dortmund) model. 4. Revision and extension. *Industrial and Engineering Chemistry Research* **41**: 1678-1688.

[9] Harper PM, Gani R, Kolar P & Ishikawa T (1999) Computer-aided molecular design with combined molecular modeling and group contribution. *Fluid Phase Equilibria* **158-160**: 337–347.

[10] Hostrup M, Harper PM & Gani R (1999) Design of environmentally benign processes: Integration of solvent design and separation process synthesis. *Computers and Chemical Engineering* **23**: 1395-1414.

[11] Ihmels EC & Gmehling J (2003) Extension and Revision of the Group Contribution Method GCVOL for the Prediction of Pure Compound Liquid Densities. *Ind. Eng. Chem. Res.* **42**: 408-412.

[12] Karunanithi AT, Achenie LEK & Gani R (2005) A New Decomposition-Based Computer-Aided Molecular/Mixture Design Methodology for the Design of Optimal Solvents and Solvent Mixtures. *Ind. Eng. Chem. Res.* **44**: 4785-4797.

[13] Karunanithi AT, Achenie LEK & Gani R (2006) A computer-aided molecular design framework for crystallization solvent design. *Chemical Engineering Science* **61**: 1243 – 1256.

[14] Kim K-J & Diwekar UM (2002) Efficient Combinatorial Optimization under Uncertainty. 2. Application to Stochastic Solvent Selection. *Ind. Eng. Chem. Res.* **41**: 1285-1296.

[15] KIM K-J, DIWEKAR UM & TOMAZI KG (2004) ENTRAINER SELECTION AND SOLVENT RECYCLING IN COMPLEX BATCH DISTILLATION. *Chem. Eng. Comm.* **191**: 1606-1633.

[16] Klamt A, Eckert F & Hornig M (2001) COSMO-RS: A novel view to physiological solvation and partition questions. *Journal of Computer-Aided Molecular Design* **15**: 355–365.

[17] Lehmann A & Maranas CD (2004) Molecular design using quantum chemical calculations for property estimation. *Industrial and Engineering Chemistry Research* **43**: 3419-3432.

[18] Marcoulaki EC & Kokossis AC (2000) On the development of novel chemicals using a systematic optimisation approach. Part II. Solvent design. *Chemical Engineering Science* **55**: 2547-2561.

[19] Marcoulaki EC, Kokossis AC & Batzias FA (2000) Novel chemicals for clean and efficient processes using stochastic optimization. *Computers and Chemical Engineering* **24**: 705-710.

[20] Meniai AH & Newsham DMT (1999) Molecular solvent design for liquid-liquid extraction using the UNIQUAC model. *Fluid Phase Equilibria* **158-160**: 327-335.

[21] Sheldon TJ, Folic M & Adjiman CS (2006) Solvent Design Using a Quantum Mechanical Continuum Solvation Model. *Ind. Eng. Chem. Res* **45**: 1128-1140.

[22] Stefanis SK, Buxton A, Livingston AG & Pistikopoulos EN (1996) A methodology for environmental impact minimization: Solvent design and reaction path synthesis issues. *Computers and Chemical Engineering* **20**: S1419-S1424.

[23] Van Dyk B & Nieuwoudt I (2000) Design of solvents for extractive distillation. *Industrial and Engineering Chemistry Research* **39**: 1423-1429.

[24] van Dyk B & Nieuwoudt I (2002) DESIGNER SOLVENTS FOR LIQUID-LIQUID EXTRACTION. ed.^eds.), p.^pp. 1246-1252. Cape Town (South Africa).

[25] Venkatasabrumanian V, Chan K & Caruthers JM (1994) Computer aided molecular design using genetic algorithms. *Computers and Chemical Engineering* **18**: 833-844.

[26] Wang Y & Achenie LEK (2002) Computer aided solvent design for extractive fermentation. *Fluid Phase Equilibria* **201**: 1-18.

[27] Weidlich U & Gmehling J (1987) MODIFIED UNIFAC MODEL. 1. PREDICTION OF VLE, h**E, AND gamma ** infinity. *Industrial and Engineering Chemistry Research* **26**: 1372-1381.

[28] Xu W & Diwekar UM (2005) Environmentally Friendly Heterogeneous Azeotropic Distillation System Design: Integration of EBS Selection and IPS Recycling. *Ind. Eng. Chem. Res.* **44**: 4061-4067.

[29] Xu W & Diwekar UM (2005) Improved Genetic Algorithms for Deterministic Optimization and Optimization under Uncertainty. Part II. Solvent Selection under Uncertainty. *Ind. Eng. Chem. Res.* **44**: 7138-7146.