Development of Corresponding States Model for Estimation of the Surface Tension of Chemical Compounds

Farhad Gharagheizi
Dept. of Chemical Engineering, Buinzahra Branch, Islamic Azad University, Buinzahra, Iran

Ali Eslamimanesh
MINES ParisTech, CEP/TEP - Centre Énergétique et Procédés, 35 Rue Saint Honoré, 77305 Fontainebleau, France

Mehdi Sattari
Saman Energy Giti Co., Tehran, 3331619636, Iran

Amir H. Mohammadi
MINES ParisTech, CEP/TEP - Centre Énergétique et Procédés, 35 Rue Saint Honoré, 77305 Fontainebleau, France

Thermodynamics Research Unit, School of Chemical Engineering, University of KwaZulu-Natal, Howard College Campus, King George V Avenue, Durban 4041, South Africa

Dominique Richon
Technical University of Denmark, Center for Energy Resources Engineering (CERE), Dept. of Chemical and Biochemical Engineering, DK-2800 Kgs. Lyngby, Denmark

Thermodynamics Research Unit, School of Chemical Engineering, University of KwaZulu-Natal, Howard College Campus, King George V Avenue, Durban 4041, South Africa

DOI 10.1002/aic.13824
Published online May 29, 2012 in Wiley Online Library (wileyonlinelibrary.com).

The gene expression programming (GEP) strategy is applied for presenting two corresponding states models to represent/predict the surface tension of about 1,700 compounds (mostly organic) from 75 chemical families at various temperatures collected from the DIPPR 801 database. The models parameters include critical temperature or temperature/critical volume/acentric factor/critical pressure/reduced temperature/reduced normal boiling point temperature/molecular weight of the compounds. Around 1,300 surface tension data of 118 random compounds are used for developing the first model (a four-parameter model) and about 20,000 data related to around 1,600 compounds are applied for checking its prediction capability. For the second one (a five-parameter model), about 10,000 random data are applied for its development, and 11,000 data are used for testing its prediction ability. The statistical parameters including average absolute relative deviations of the results form dataset values (25 and 18% for the first and second models, respectively) demonstrate the accuracy of the presented models. © 2012 American Institute of Chemical Engineers *AIChE J,* 59: 613–621, 2013
Keywords: physicochemical properties, gene expression programming, corresponding states, accurate model, optimization, surface tension

Introduction

It is currently well-accepted that the efficient processes in chemical, petroleum, pharmaceutical, and polymer industries are generally designed in the case of availability of the values of the thermophysical, physicochemical, and/or thermodynamic properties of the involved components or their relevant mixtures.[1–4] This can be expected by the corresponding experimental measurements of such properties (beside the integrity of the processes' design), which are obtained through accurate experimental procedures/techniques and apparatuses in modern laboratories. However, the broad experimental efforts, since the past two centuries, indicate that generating these kinds of data are intrinsically time-consuming, costly, and with probable non-negligible uncertainties mainly due to very high (or very low) temperature/pressure conditions, extremely low compositions of a particular species in the mixtures, inappropriate design of the apparatuses, unreliable experimental techniques, human mistakes during the measurements, carelessness in calibration of the instruments, and so forth.[5,6]

Therefore, there have been considerable numbers of theoretical studies to correlate the existing experimental data for particular conditions and for specific chemical compounds or mixtures. These correlations may be later applied to predict the corresponding properties for other systems at specific conditions within the domain of their applicability.[1]

Additional Supporting Information may be found in the online version of this article.

Correspondence concerning this article should be addressed to A. H. Mohammadi at amir-hossein.mohammadi@mines-paristech.fr.

Nevertheless, the mentioned properties may not be calculated/estimated merely using the basic thermodynamics.[1] Different molecular theories are generally capable of such evaluations.[1] Although, the thermodynamic concepts lead to reduction in the complexity of the molecular theories by relating one physical property to another one and, consequently, contribute to generating two main categories of correlations: Empirical correlations and semiempirical ones.[1]

Acceptable results can be obtained, normally, through the empirical correlations within the range of the conditions and the compounds and/or mixtures, which have been applied for their development; however, any extrapolation may not be recommended.[1] Semiempirical correlations use some theoretical basis introducing particular parameters to improve the empirical correlations prediction capability. As a matter of fact, incorporating empirical expressions with the theoretical relations generally provides powerful methods for developing reliable and predictive equations.[1]

In any case, certain parameters of the aforementioned correlations should be regressed over selected experimental data. Numerous mathematical methods including linear and nonlinear regression methods and various kinds of optimization techniques have been so far proposed for this purpose. There is generally no doubt for the scientific community that the applied methods for development of the correlations (at least the genetic-type ones) are supposed to satisfy the following criteria[5,7–14]:

1. More probability for convergence to the global optimum;
2. High probability of finding a mathematically-correct solution;
3. No requirement for determination of the network topology in advance; which can be automatically determined as the training process ends;
4. Low probability to face overfitting/underfitting problems;
5. Acceptable generalization performance;
6. Fewer adjustable parameters;
7. Relying on the population-based initialization;
8. No requirement to define the final correlation form for the algorithm;
9. Use of the basis of stochastic evolutionary principles;
10. Ability to handle nondifferentiable, nonlinear, and multimodal cost functions;
11. Few, robust, and easy to choose control variables to steer the minimization of the objective function;
12. No sensitivity to starting points that is, starting decision variables or objective function values; and
13. Consistent and consecutive modification of the solutions in each generation.

The preceding criteria would eventually contribute to obtaining reliable predictive correlations that yield acceptable results in short computational time.

The genetic algorithm (GA), first introduced by Holland[15] is considered as a heuristic optimization technique (among the evolutionary algorithms) that pursues the process of natural evolution. In biology, an organism contains a set of rules, explaining the way that the organism is constructed from the very small building blocks of life. These rules are generally encoded in the genes of an organism that are connected together in the shape of strings called chromosomes. The GA algorithm[15] generally generates chromosomes (population of strings), which encode solutions to optimization problems through specific operators like selection, mutation, and crossover.[13]

The final solutions of a GA optimization process[15] are encoded in fixed-length binary (0 and 1) strings. The modifications of this algorithm mainly focus on manipulation of the mentioned operators. The genetic programming (GP)[16,17] is an effective improvement of the GA, in which the solutions are presented as nonlinear structures of parse trees (treated as functions) instead of fixed-length binary solutions. This modification results in searching among variety of possible functions for finding the final solution.[16,17] Considering the drawbacks of the GP[16,17] (which will be discussed later), Ferreira[18] introduced a very fruitful modification to the original GP algorithm.[16,17] In the new strategy, called "gene expression programming (GEP)",[18] ramified structures of different sizes and shapes (parse trees) are completely encoded in the linear solutions of fixed length that finally lead to more probability of obtaining the global optimum of the model parameters.[13,18] The details of the GEP[18] algorithm are described in the next section.

As the GEP[18] strategy has been, up to now, applied for several electrical, mechanical, and bioengineering purposes such as constructing a quantitative structure-activity relationship (QSAR) for investigation of the human dopamine sulfotransferases,[19] development of stage-discharge curves of rivers,[20] splitting tensile strength of concrete,[21] transformer fault diagnosis,[22] prediction of lateral outflow over triangular labyrinth side weirs,[23] designing electronic circuits,[24] etc., it is of great interest to employ the same strategy (GEP[18]) for determination of the physicochemical properties of the chemical compounds (mostly organic ones). This article deals with representation/prediction of the surface tension values of about 1,700 chemical compounds (mostly organic) at various temperatures. To achieve this goal, a corresponding states model is developed applying the GEP[18] algorithm.

Theory

The surface effects on the natural phenomena and industrial applications such as reactions over the surface of a catalyst, boiling heat transfer, condensation, and microscale channel flows processes such as lubrication, corrosion, adherence, detergency, and reactions in electrochemical cells have been the subject of many studies especially in the past decade.[25–28] Liquid-vapor interface behavior is important to investigate the performance of detergents, in chemical engineering separations like absorption and distillation, estimation of the capillary pressure to investigate the effects of surface forces on fluid distribution within a reservoir, determination of condensate recovery in the case of retrograde condensation in gas condensate reservoirs, and in the performance of biological membranes.[25–30]

There are unequal asymmetric forces acting upon a molecule, which are zero at equilibrium.[1,27–30] At low-gas densities, the molecules at the surface experience a sidewise attraction toward the bulk liquid, meanwhile they are attracted a little in the direction of the bulk gas.[1,27–30] These attractive forces tend to pull the surface toward the bulk-liquid phase. Therefore, the surface layer is in tension and, at equilibrium, it tends to minimize its area compatible with the mass of material, container restraints, and external forces.[1,27–30] This molecular tension at the surface is quantitatively expressed as surface (interfacial) tension, which refers to the force exerted at the interface per unit length.[1,27–30] As a matter of fact, it is a macroscopic thermophysical property that affects behavior of fluids in a variety of processes, as already mentioned.

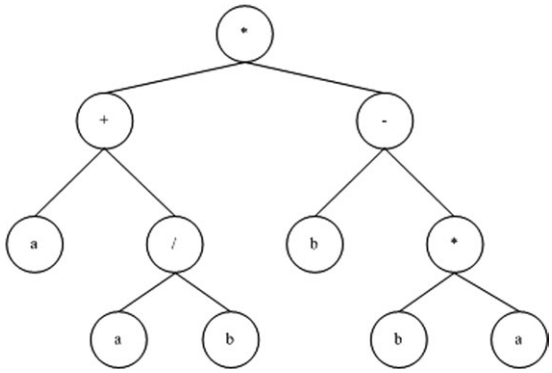

Figure 1. A typical computer LISP program in the GP[16,17] algorithm represented as a parse tree (expression tree), which stands for the algebraic expression [a + (a/b)]×[b − (a × b)] by a two-gene chromosome.

Presentation of the methods for evaluation of this property was, perhaps, begun in 1923, when Macleod[31] proposed an empirical equation to correlate the experimental values of surface tension based on the difference between the liquid and vapor densities (at equilibrium) of a chemical compound at given temperature and a constant characteristic of the liquid phase as follows[27,28,31]

$$\sigma^{1/4} = (P)(\rho_L - \rho_V)/Mw \tag{1}$$

where, σ is the surface tension in (dyn.cm^{-1}), P denotes the parachor in (kg$^{1/4}$ m^3 s$^{-1/2}$ kmol^{-1}), ρ is the density in (g.cm^{-3}), M is the molecular weight, and subscripts L and V refer to the liquid and vapor phases, respectively.

Since that time, lots of efforts have been undertaken for the determination of surface tension of pure compounds. These new works have been mainly based on parachor parameter,[32–40] modification of the original correlation of Macleod,[31,41–46] corresponding state principles,[47–53] molecular dynamics (MD) simulations,[26,54–64] finite-size scaling approach,[26] transferable potentials for phase equilibria (TraPPE), and optimized potentials for liquid simulations (OPLS) force fields,[65] equations of state,[66–68] molecular layer structure theory (MLST),[69] correlation between surface tension and liquid compressibility,[70] and statistical-mechanic accompanied by corresponding state principle.[30] A rigorous review on these methods has been already well-presented.[35] Recently, our group has proposed two accurate models based on group contribution (GC) method,[28] and quantitative structure property relationship (QSPR) strategy[27] for representation/prediction of surface tensions of 752 and 1,604 chemical compounds at various temperatures, respectively. The obtained results show absolute average relative deviations of the calculated/estimated properties from the applied data: about 2 and 4% and squared correlation coefficients: 0.997 and 0.985 through applying the GC[34] and QSPR[35] models, respectively. However, presenting more easy-to-use correlations for determination of surface tension values of chemical compounds (mostly organic ones) may be of more interest of some researchers and engineers.

Mathematical Strategy

Genetic programming

As mentioned earlier, the GP[16,17] is an extension of the genetic algorithms. The defined problem (the forms of the functions, number of parameters, etc.) does not affect the main organization of the GP searches manner.[13,16,17] The main distinction between the GP[16,17] and the GA[15] is that in the former, the chromosomes consist of nonlinear structures similar to parse trees although they are similar to the GA[15] linear structures, which are naked replicators working as genotype and phenotype.[13] These parse trees, adopted like the protein molecules, include diverse forms of functionality. Therefore, the final solution of a specific problem can be found among more various types of functions. It is worth pointing out that the genetic operators (e.g., recombination, crossover, and mutation) also operate during the computational steps of the GP[16,17] similar to those of the original GA,[15] but they resemble to pruning and grafting of trees. As indicated by Ferreira,[13] the main disadvantage of the GP is that the complex replicators (parse trees structures) can only be modified in limited ranges because their reproduction should be done only on the parse trees. These improvements include modifying or exchanging definite branches of the corresponding parse trees,[13] that may lead to invalid (unacceptable) trees structures. A typical computer LISP program based on the GP[16,17] algorithm is shown in Figure 1. It should be noted that the GP uses these kinds of computer programs for data representation.[21]

Gene Expression Programming

The GEP,[18] resulting from the modification and extension of the GP,[16,17] is employing computer programs in order to solve a problem. In the latter technique, the populations individuals are symbolic expression trees unlike those of GEP[18] in which the individuals are encoded as linear chromosomes, which are later translated into the expression parse trees that is, the genotype and phenotype are eventually separated. As a result, the GEP[18] algorithm possesses many of the advantages of the evolutionary systems.[13] Another element of the GEP[18] is that the chromosomes are such designed that can permit the creation of multiple genes. Therefore, the novel structures of the genes in the GEP[18] algorithm, allows encoding of any program for efficient evolution of the solutions.[13] The organized structure of the genes also permits powerful and efficient genetic operators looking for the solutions in the entire feasible region of the problem.[13] These operators are recombined directly on the linear encoding (before it is translated into a tree). Recombination, as a

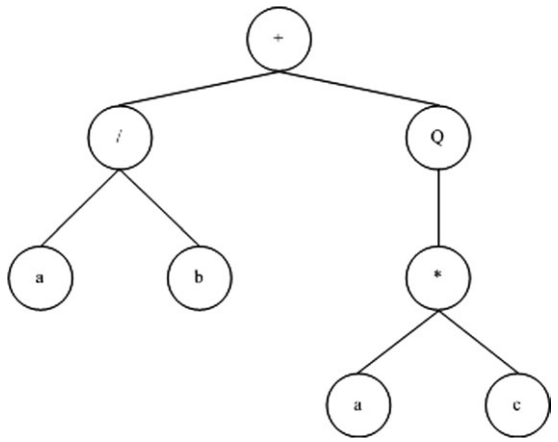

Figure 2. A typical Karva language program in the GEP[16,17] strategy, which represents the algebraic expression [(a/b)]+[($\sqrt{a \times c}$)] by a two-gene chromosome.

matter of fact, is sharing the information from the genes of the parents to the gene of the offspring. Therefore, the modified parts of the resulting expression trees generally experience little resemblance to their previous ancestors.[71] On the other hand, it can be stated that the GEP[18] special codes use Karva language to make it possible to infer exactly the phenotype given the sequence of a gene and vice versa.[21] For instance, the algebraic expression (a/b) + ($\sqrt{a \times c}$) can be easily shown as a diagram or expression tree (ET) like Figure 2, with the Karva language representation of */Q * a b a c (Q denotes the squared root function). Each character places in one position from 0 to 7, and can be shown as 0 1 2 3 4 5 6 7. To recapitulate, the operators in this algorithm result in generation of the valid parse trees that can be a complex mathematical structure or even artificial neural networks.[13]

The GEP General Computational Steps. Ferreira[13] expressed the general computational algorithm of the GEP[18] strategy as follows[13]:

1. Initialization of the population comprising the random generated chromosomes of a certain number of individuals;
2. Fitness of the individuals based on fitness functions (cases);
3. Selecting the individuals according to their fitness to reproduce with modification;
4. The new individuals are treated using the same procedure including expression of the genomes, confrontation of the selection environment, selection, and reproduction with modification;
5. Repeating the aforementioned steps for a certain number of generations or until a good solution has been found (convergence of the algorithm according to the defined criteria).

The same algorithm has been followed in this work for developing the corresponding states models.

Database

As the quality and generality of the applied database has a direct influence on the developed models for prediction of physicochemical properties, the DIPPR 801 database,[72] has been herein used (including experimental and pseudoexperimental values). The reported surface tension values of about 1,700 compounds (mostly organic) at various temperatures have been employed for developing and validating two corresponding states models.

Developing the models

As previously explained, one of the significant features of the GEP[18] mathematical strategy is that there is no need to assume specific functional forms to find the best representation/prediction of the applied data.[13,18,73] Thus, an accurate functional (correlation) form involving selected parameters (perhaps the most efficient ones) are obtained through the evolutionary algorithm itself. Our previous literature survey[27,28] shows that the data of surface tension of the pure chemical compounds can be normally correlated by the corresponding state principle parameters such as critical pressure (P_c), critical/reduced temperature (T_c/T_r), critical volume (V_c), acentric factor (ω), normal boiling point temperature (T_b), and/or reduced normal boiling point temperature (T_{br}) along with the molecular weight (Mw).[1,47–53] In this work, it is preliminary assumed that the surface tension value can be formulated as the functions of the aforementioned properties (plus temperature) as follows

$$\sigma = f(T, T_c, T_r, P_c, V_c, \omega, T_b, T_{br}, Mw) \tag{2}$$

Having defined the parameters of the correlation, the following computational algorithm has been applied:

1. Initialization of the population or generating, randomly, the chromosomal structures of the individuals by setting many correlations presented as pars trees using the operators (−, +,*, /, ^), and terminals as functions of the input data and the output desired results (surface tension values)
2. Calculation of the fitness value for every individual of the generated population by the following objective function (OF)

$$OF(i) = \frac{100}{\mathrm{N}} \sum_{i}^{\mathrm{N}} \frac{\left|\sigma(i)^{\mathrm{rep/pred}} - \sigma(i)^{\mathrm{exp}}\right|}{\sigma(i)^{\mathrm{exp}}} \tag{3}$$

where N is the number of the data points used in the GEP[18] algorithms, and superscripts rep/pred and exp denote the determined surface tension values by the final developed correlation and applied data, respectively.

3. Selection of the individuals to stand for appropriate parents for replacement, which were evaluated from the fitness values. In this work, the tournament technique[73,74] has been applied to provide an acceptable diversity of the population in each generation.
4. Use of the genetic operators including replication, mutation, and inversion for gene reproduction with modification of computational steps:
   (a) Replication operator: It copies exactly the chromosomes of the individuals chosen in the selection step (step 3).[13]
   (b) Mutation operator: It contributes to efficient adaption of populations of individuals.[13] In this study, the point mutation has been used, in which a random node (in the structures of the chromosomes) is selected, and the stored information is replaced with a different random primitive of the same arity taken from the initial (old) set.[75] Having defined the mutation rate (p_m), the mutation can occure everywhere in the structural organization of chromosomes although with preservation of the original structure.[13] Generally, the mutation can be performed through changing the heads of genes symbols and terminals of the tails.[13,18]
   (c) Inversion operator: This operator is applied to create new individuals through modification of the heads of randomly selected genes. It has already been proven that all the new individuals created by inversion are correct programs.[13] The performance of this operator can be set choosing a value for the inversion rate (p_i).[13]
5. Transposition and insertion sequence elements: The transposable elements of gene expression programming are portion of the genome that can be activated and jumped to another place in the chromosome, which includes three types as implemented by Ferreira[13]: "*Short fragments with either a function or terminal in the first position transpose to the head of genes, short*

Table 1. The Parameters of the GEP[18] Algorithm Applied in the Computational Route

GEP[18] algorithm parameters	Value
Number of chromosomes	30
Head size	8
Number of genes	7
Linking function	+
Generations without change	2000
Fitness function	AARD%[a]
Mutation	0.044
Inversion	0.1
IS transposition	0.1
RIS transposition	0.1
One-point recombination	0.3
Two-point recombination	0.3
Gene recombination	0.1
Gene transposition	0.1
Constant per gene	2
Operators used:	+
	−
	*
	/
	$\sqrt{}$
	exp
	$\log_e$
	power

[a] $\%AARD = 100/\mathrm{N}\sum_{i}^{N}\frac{|\mathrm{rep.(i)/pred.(i)-exp.(i)}|}{\mathrm{exp.(i)}}$, where N is the number of data.

fragments with a function in the first position that transpose to the root of genes (root IS elements or RIS elements), and entire genes that transpose to the beginning of chromosomes."

6. Recombination: This step, which is conducted in three manners including one- and two-point recombination, and also gene recombination,[13] randomly chooses two chromosomes to exchange specific material with each other, leading to the appearance of two new chromosomes.[13] Consequently, new A generation is created. The preceding procedure is repeated until the defined stopping criteria (can be user-defined convergence criteria or maximum number of generations) are satisfied. The details of this procedure along with comprehensive examples are provided by Ferreira.[13]

Results and Discussion

The aforementioned calculation procedure has been pursued to obtain accurate and simple models. During the calculation steps (for developing and testing the first correlation), the main dataset has been randomly divided into the ldquo;-trainingrdquo; set (1,164 data points, about 5% of the whole dataset), the ldquo;validation (optimization)rdquo; set (145 data points, about 1% of the whole dataset), and the ldquo;test (prediction)rdquo; set (20,135 data points, about 94% of the whole dataset). The process of division of database into three subdatasets is performed randomly. As a matter of fact, the GEP[18] algorithm computational steps define the required parameters, which yield an accurate model from the introduced parameters (T, T_c, T_r, P_c, V_c, ω, T_b, T_{br}, Mw). Therefore, one can consider several independent parameters for a particular problem and obtain the ones, which have the most effects on the desired output results. The first corresponding states model can be represented as follows

$$\sigma(N \cdot m^{-1}) = 8.948226 * 10^{-4}\left[\frac{A^2}{Mw}\sqrt{\frac{A\omega}{Mw}}\right]^{0.5} \quad (4)$$

where

$$A = (T_c(K) - T(K) - \omega) \quad (5)$$

The number of the digits of the coefficient in Eq. 4 has been determined by performing sensitivity analysis of the calculated/estimated results to the corresponding value. The statistical parameters of the obtained results indicate that the absolute average relative deviation of the represented/predicted values from the used surface tension data[72] is about 25%. This issue shows an overall acceptable accuracy of the model for determination of the surface tension of many organic compounds at various temperatures. Comparison between the represented/predicted surface tension values applying Eq. 4 and the applied data[72] are presented in supporting information (please see online for Supporting Information).

The significant parameters of the GEP[18] calculation steps are reported in Table 1. As can be seen, the numbers of treated chromosomes, genes, the mutation and inversion coefficients, and the applied operators are among the effective parameters of the algorithm required for obtaining an acceptable correlation of interest. The capability of the proposed model for evaluation of the surface tension of the investigated chemicals have been compared to the calculated/estimated results by several most widely-used correlations available in the literature. Table 2 reports the corresponding results. It reveals that the obtained model in this work (first model) leads to reasonable deviations of the determined surface tension values from the applied data[72] compared to the studied models. The surface tension values of the treated chemicals calculated through the studied models are reported in Supporting Information.

As evident in the provided tables (in this article and supporting information), this model does not bring about accurate representations for some of the investigated data. It is probably due to development of this model on the basis of the surface tension data of limited number compounds from particular chemical families (only seven chemical families). On the other hand, we would rather point out that the time of the calculations using the GEP[18] algorithm is of drastic importance. Pursuing the calculation steps in each generation needs parallel computing, and, consequently, high amounts of time. For development of the first correlation, we have defined a stopping criterion for the algorithm, which is the number of the determined parameters of the correlation by the GEP[18] algorithm. It has been done because our primary objective has been development of a correlation with the

Table 2. Comparison Between the Results of the Obtained Models and Well-Known Ones[47–50]

Correlation	Number of parameters	AARD %
Brock and Bird[47]	4	28
Curl and Pitzer[48] and Pitzer[49]	4	36
Zuo and Stenby[50]	4	28
This work (Eq. 4)	4	25
This work (Eq. 6)	5	18

(The results of our previous works[27,28] have not been included in this table because they are not generally considered as simple models)

same number of parameters than the investigated expressions from the literature using a small portion of the data for training and optimization processes. Thus, it has been possible to obtain a more accurate correlation through continuation of the calculation steps following more generations from the subsequent populations along with a different stopping criterion and application of more surface tension data from wide ranges of chemical families for the training step. This task has been also undertaken in this work. The results contribute to obtaining a five-parameter correlation (the AARD % of the corresponding results is about 18%) as follows

$$\sigma(N \cdot m^{-1}) = 10^{-4}(P_c/bar)^{\frac{2}{3}}(T_c/K)^{\frac{1}{3}}(1-T_r)^{\frac{11}{9}} \times \left[7.728729T_{br} + 2.476318(T_{br}^3 + (V_c/m^3 \cdot kmol^{-1}))\right] \quad (6)$$

To develop and test the second correlation (Eq. 6), the main dataset has been randomly divided into the "training" set (ca. 8,000 data points equaling to around 40% of the whole dataset), the "validation (optimization)" set (2,000 data points equaling to 9% of the whole dataset), and the "test (prediction)" set (11,000 data points equaling to 51% of the whole dataset). Figures 3 and 4 clearly indicate the corresponding results applying Eq. 6. In addition, Table 3 reports the deviations of the results of the aforementioned correlations for the chemical families categorized in 75 chemical families. These tables can easily recommend the most accurate model among the studied ones for calculation/estimation of the surface tension values of the chemicals from each chemical family.

One significant point should be emphasized in our discussion. As mentioned earlier, we have previously developed two models[27,28] for the same purpose using different strategies. The numbers of the data points treated in previous works[27,28] for developing the models (i.e., the data used in training sets and optimization sets) were much higher than in this work. In previously presented models,[27,28] around 90% of the corresponding data were applied for their development. Use of the fewer number of data applied for training and optimization of the current models shows the high capability of the GEP[18] algorithm to develop correlations with acceptable predictions. In addition, the proposed models are easy to apply and do not require any computer programs for computations. The obtained results interpret that the applied algorithm may be promising for evaluation of other thermophysical properties. Another factor to consider is the limitation of the applied mathematical algorithm for handling such a large dataset. At the time of preparing this article, we were not able to use more data for the training and optimization sets of the applied GEP[18] algorithm.

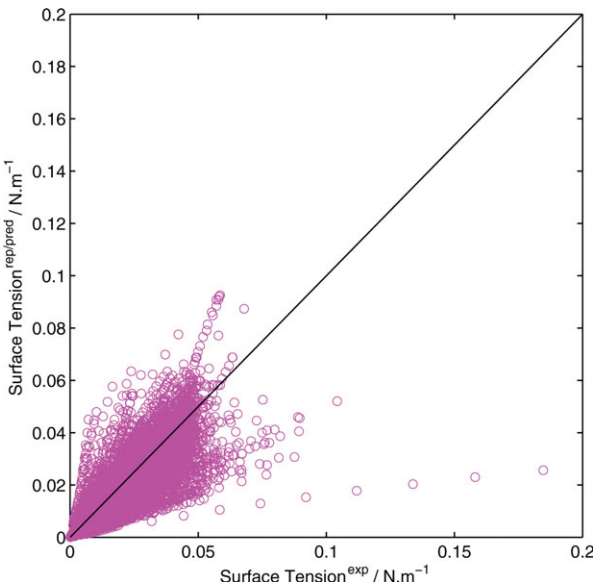

Figure 3. Comparison between the whole represented/predicted results of the second developed model (Eq. 6), and the reported values in the DIPPR[72] of surface tension of investigated compounds at various temperatures.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

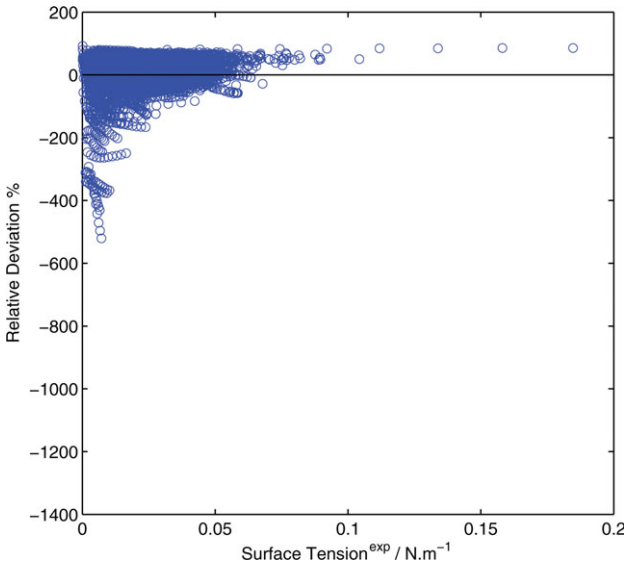

Figure 4. Relative deviations of the whole represented/predicted surface tension values of the investigated chemical compounds by the second developed model (Eq. 6) from the applied data.[72]

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Careful scrutiny of the results show that the proposed models lead to high deviations of some of the evaluated surface tension values from the reported ones in the DIPPR 801.[72] These results may not be corresponded to particular compounds or chemical families. However, it can be concluded that the deviations of the results using the first correlation for the compounds containing halogens, some compounds in polyfunctional organics, polyfunctional nitriles, and polyols are generally higher than the corresponding values for the compounds from other family groups. Similarly, the deviations of the obtained surface tension values using the second correlation for several compounds in polyfunctional organics and polyols are generally more considerable compared to the other families.

Apart from that, the applied surface tension values[72] contain definite uncertainties, which generally lie between 1 to 5% (see Supporting Information). These uncertainties affect, indeed, the prediction capability of the obtained correlation. It may be possible to eliminate the probable outliers (the data, for which the evaluated results show very high

Table 3. The Absolute Average Relative Deviations of the Investigated Models Results from the Reported Values in the DIPPR 801[72] for Each Chemical Family of the Studied Compounds

No.	Chemical family	AARD%/ Brock and Bird[47]	AARD%/ Pitzer Curl[48] and Pitzer[49]	AARD%/ Zuo and Stenby[50]	AARD%/ This work (Eq.4)	AARD%/ This work (Eq.6)	Number of compounds in each of family
1	1-ALKENES	2.9	6.6	3.0	4	2	283
2	2,3,4-ALKENES	3.3	4.6	3.3	4.3	3.5	294
3	ACETATES	5.3	9.1	5.3	12	7	317
4	ALDEHYDES	11	14	11	10	10	455
5	ALIPHATIC ETHERS	6.4	9.6	6.5	6.6	5.3	455
6	ALKYLCYCLOHEXANES	5.1	4.1	4.6	4.8	2.7	230
7	ALKYLCYCLOPENTANES	2.4	5.5	2.6	2.8	3.0	151
8	ALKYNES	5.1	6.9	5.1	3.9	4.8	204
9	ANHYDRIDES	27	33	27	15	14	124
10	AROMATIC ALCOHOLS	30	36	30	15	17	457
11	AROMATIC AMINES	20	24	20	15	14	470
12	AROMATIC CARBOXYLIC ACIDS	54	82	59	16	20	122
13	AROMATIC CHLORIDES	8.4	7.9	8.3	28	10	182
14	AROMATIC ESTERS	13	18	13	18	11	328
15	C, H, BR COMPOUNDS	13	17	14	41	11	230
16	C, H, F COMPOUNDS	11	12	11	48	13	508
17	C, H, I COMPOUNDS	9.7	14	11	44	7	117
18	C, H, MULTIHALOGEN COMPOUNDS	12	11	11	54	12	540
19	C, H, NO2 COMPOUNDS	13	17	13	17	14	351
20	C1/C2 ALIPHATIC CHLORIDES	6.7	5.8	6.3	38	8	271
21	C3 & HIGHER ALIPHATIC CHLORIDES	7.6	6.8	7.6	24	9	349
22	CYCLOALIPHATIC ALCOHOLS	30	36	30	15	14	151
23	CYCLOALKANES	3.0	3.1	2.3	8.3	3.7	70
24	CYCLOALKENES	3.4	4.8	3.1	8.2	3.6	151
25	DIALKENES	5.5	7.0	5.3	6.0	5.6	342
26	DICARBOXYLIC ACIDS	32	47	34	16	18	153
27	DIMETHYLALKANES	1.3	4.8	1.3	2.3	4.0	272
28	DIPHENYL/POLYAROMATICS	18	18	18	22	18	251
29	EPOXIDES	5.4	7.4	5.4	13	8	193
30	ETHYL & HIGHER ALKENES	5.5	6.0	5.4	4.7	6.1	146
31	FORMATES	6.7	11	6.4	7.6	5.3	169
32	ISOCYANATES/DIISOCYANATES	17	17	16	24	20	116
33	KETONES	7.9	11	7.7	6.8	5.5	596
34	MERCAPTANS	5.4	9.1	5.5	7.0	5.1	273
35	METHYLALKANES	1.3	3.9	1.2	2.7	1.8	186
36	METHYLALKENES	5.4	6.4	5.4	4.8	6.3	291
37	MULTIRING CYCLOALKANES	3.1	4.2	3.2	5.0	4.8	67
38	N-ALCOHOLS	31	39	31	27	11	369
39	N-ALIPHATIC ACIDS	25	37	25	17	10	312
40	N-ALIPHATIC PRIMARY AMINES	6.9	11	6.1	6.0	3.7	183
41	N-ALKANES	3.4	8.8	3.8	4.9	2.7	495
42	N-ALKYLBENZENES	2.2	6.9	2.7	3.5	2.5	279
43	NAPHTHALENES	6.3	9.0	6.2	5.3	5.8	241
44	NITRILES	17	20	18	33	20	395
45	NITROAMINES	20	30	21	13	14	75
46	ORGANIC SALTS	57	62	56	15	40	153
47	OTHER ALIPHATIC ACIDS	34	40	34	30	20	322
48	OTHER ALIPHATIC ALCOHOLS	29	39	30	16	11	602
49	OTHER ALIPHATIC AMINES	7.1	12	7.1	5.3	4.6	248
50	OTHER ALKANES	1.9	4.0	2.0	2.4	4.5	276
51	OTHER ALKYLBENZENES	5.5	5.6	5.5	7.2	5.6	659
52	OTHER AMINES, IMINES	21	26	21	17	17	450
53	OTHER CONDENSED RINGS	25	25	25	24	26	114
54	OTHER ETHERS/DIETHERS	8.5	12	8.6	11	8	370
55	OTHER HYDROCARBON RINGS	9.9	10	9.9	11	10	174
56	OTHER MONOAROMATICS	6.1	5.4	6.1	7.6	6.6	222
57	OTHER ORGANIC COMPOUNDS	7.1	8.4	7.4	27	11	81
58	OTHER POLYFUNCTIONAL C, H, O	33	49	33	16	15	754
59	OTHER POLYFUNCTIONAL ORGANICS	167	174	165	112	137	102
60	OTHER SATURATED ALIPHATIC ESTERS	15	21	15	16	11	256
61	PEROXIDES	43	57	46	15	17	133
62	POLYFUNCTIONAL ACIDS	54	86	59	25	23	239
63	POLYFUNCTIONAL AMIDES/AMINES	41	52	42	32	21	377
64	POLYFUNCTIONAL C, H, N, HALIDE, (O)	36	40	35	33	31	146
65	POLYFUNCTIONAL C, H, O, HALIDE	24	27	24	32	18	525
66	POLYFUNCTIONAL C, H, O, N	43	54	44	29	24	339
67	POLYFUNCTIONAL C, H, O, S	36	42	36	23	19	143
68	POLYFUNCTIONAL ESTERS	34	56	36	19	17	314
69	POLYFUNCTIONAL NITRILES	47	55	46	48	16	59
70	POLYOLS	101	141	105	40	43	451
71	PROPIONATES AND BUTYRATES	3.3	7.3	3.1	7.8	4.0	205
72	SILANES/SILOXANES	14	14	14	27	15	595
73	SULFIDES/THIOPHENES	6.4	9.0	6.5	13	6	538
74	TERPENES	5.3	5.6	5.3	5.3	3.7	104
75	UNSATURATED ALIPHATIC ESTERS	6.1	12	6.5	10	5	279

deviations from the used data) from the proposed models results and develop more accurate ones; however, our aim, herein, has been to investigate the ability of all of the investigated correlations for representation/prediction of the whole surface tension values from one of the most comprehensive datasets in the literature.[72]

In the final analysis, it can be pointed out that one of the significant achievements of this study is that the GEP[18] algorithm has eventually generated an empirical correlation (Eq. 6) that is very similar to the semiempirical models of Curl and Pitzer,[48] Pitzer,[49] and Brock and Bird,[47] which have been developed based on theory (see Ref. [1] for details of these models).

Conclusion

In this communication, the gene expression programming[18] mathematical algorithm was applied to develop two corresponding states models including a four-parameter and a five-parameter one for representation/prediction of the surface tension values of about 1,700 chemical compounds at various temperatures collected from the DIPPR 801 database.[72] The parameters of the models include the critical temperature, or temperature/critical volume/acentric factor/critical pressure/reduced temperature/reduced normal boiling point temperature/molecular weight of the compounds. Around 1,300 surface tension data for the chemical compounds from methylalkanes, dimethylalkanes, other alkanes, 1-alkenes, methylalkenes, ethyl, and higher alkenes chemical families were applied for developing the first model (ca. 6% of the whole dataset), and about 20,000 data (ca. 94% of the whole dataset) were used for its testing. Regarding the second expression, around 40%, 9%, and 51% of the reported data[72] were used for training, optimization, and testing processes, respectively.

The statistical parameters of the obtained correlations show 25 and 18% absolute average relative deviations of the results from the reported values[72] for the first correlation and the second one, respectively. Comparison of the calculated/estimated results of the proposed models and those of through application of the well-known models[47–50] available in open literature demonstrate their high capability for determination of the surface tension of the many chemicals from 75 various chemical families at various temperatures. Indeed, using more and/or more accurate experimental surface tension values (in the case of availability), and considering more calculation time for the GEP[18] algorithm to converge to more probable global optimum of the objective function of the problem shall contribute to developing more predictive tools for the same purpose.

Acknowledgment

Ali Eslamimanesh thanks MINES ParisTech for providing him a PhD scholarship.

Literature Cited

1. Poling BE, Prausnitz JM, O'Connell JP. *Properties of Gases and Liquids*. 5th ed. New York: McGraw-Hill; 2001.
2. Shacham M, Brauner N. A dynamic library for physical and thermodynamic properties correlations. *Ind Eng Chem Res*. 2000;39:1649–1657.
3. Shacham M, Brauner N. High precision correlations of thermo-physical properties. *Comput Aided Chem Eng*. 2004;18:1129–1134.
4. Shore H, Brauner N, Shacham M. Modeling physical and thermodynamic properties via inverse normalizing transformations. *Ind Eng Chem Res*. 2002;41:651–656.
5. Eslamimanesh A, Gharagheizi F, Mohammadi AH, Richon D. Phase equilibrium modeling of structure H clathrate hydrates of methane + water "insoluble" hydrocarbon promoter using QSPR molecular approach. *J Chem Eng Data*. 2011;56:3775–3793.
6. Eslamimanesh A, Mohammadi AH, Richon D. Thermodynamic consistency test for experimental data of water content of methane. *AIChE J*. 2011;57:2566–2573.
7. Price K, Storn R. Differential evolution. *Dr. Dobb's J*. 1997;22:18–24.
8. Chiou JP, Wang FS. Hybrid method of evolutionary algorithms for static and dynamic optimization problems with applications to a fed-batch fermentation process. *Comput Chem Eng*. 1999;23:1277–1291.
9. Schwefel HP. *Numerical Optimization of Computer Models*. New York: John Wiley & Sons, Inc; 1981.
10. Goldberg DE. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison-Wesley; 1989.
11. Davis L. *Handbook of Genetic Algorithms*. New York: Van Nostrand Reinhold; 1991.
12. Storn R. Differential Evolution - A simple and efficient heuristic for global optimization over continuous spaces. *J Global Optim*. 1997;11:341–359.
13. Ferreira C. *Gene Expression Programming*. 2nd ed. The Netherlands: Springer-Verlag; 2006.
14. Deb K. *Multi-Objective Optimization using Evolutionary Algorithms*. West Sussex, UK: Wiley; 2002.
15. Holland JH. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence, 1975. University of Michigan Press*. 2nd ed. MIT Press; 1992.
16. Cramer NL. *A Representation for the Adaptive Generation of Simple Sequential Programs*. In: Grefenstette JJ, ed. *Proceedings of the First International Conference on Genetic Algorithms and their Applications, Erlbaum*; 1985:183–187.
17. Koza JR. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA: MIT Press; 1992.
18. Ferreira C. Gene expression programming: a new adaptive algorithm for solving problems. *Complex Syst*. 2001;13:87–129.
19. Si H, Zhao J, Cui L, Lian N, Feng H, Duan YB, Hu Z. Study of human dopamine sulfotransferases based on gene expression programming. *Chem Biol Drug Des*. 2011;78:370–377.
20. Azamathulla HM, Ghani AA, Leow, CS, Chang, CK, Zakaria NA. Gene-expression programming for the development of a stage-discharge curve of the pahang river. *Water Resour Manage*. 2011;25:2901–2916.
21. Özcan F. Gene expression programming based formulations for splitting tensile strength of concrete. *Constr Build Mater*. 2012;26:404–410.
22. Dong Z, Zhu Y. Transformer fault diagnosis based on gene expression programming classifier. *Adv Mater Res*. 2012;354–355:1022–1026.
23. Kisi O; Emin Emiroglu M, Bilhan O, Guven A. Prediction of lateral outflow over triangular labyrinth side weirs under subcritical conditions using soft computing approaches. *Expert Syst Appl*. 2012;39:3454–3460
24. Yan X, Wei W, Liang Q, Hu C, Yao Y. *Designing electronic circuits by means of gene expression programming II*. In: *ICES'07 Proceedings of the 7th international conference on Evolvable systems: from biology to hardware*; Wuhan, China; 2007:319–330.
25. Ramírez-Verduzcoa LF, Romero-Martínez A, Trejo A. Prediction of the surface tension, surface concentration, and the relative Gibbs adsorption isotherm of binary liquid systems. *Fluid Phase Equilib*. 2006;246:119–130.
26. Wemhoff AP, Carey VP. Surface tension prediction using characteristics of the density profile through the interfacial region. *Int J Thermophys*. 2006;27:413–436.
27. Gharagheizi F, Eslamimanesh A, Mohammadi AH, Richon D. Handling a very large data set for determination of surface tension of chemical compounds using quantitative structure-property relationship strategy. *Chem Eng Sci*. 2011;66:4991–5023.
28. Gharagheizi F, Eslamimanesh A, Mohammadi AH, Richon D. Use of artificial neural network-group contribution method to determine surface tension of pure compounds. *J Chem Eng Data*. 2011;56:2587–2601.
29. Danesh A. *PVT and Phase Behaviour of Petroleum Reservoir Fluids*. The Netherlands: Elsevier; 1998.
30. Escobedo J, Mansoori A. Surface tension prediction for pure fluids. *AIChE J*. 1996;42:1425–1433.
31. Macleod DB. Relation between surface tension and density. *Trans Faraday Soc*. 1923;19:38–42.
32. Sugden SA. Relation between surface tension, density, and chemical composition. *J Am Chem Soc*. 1924;125:1177–1189.

33. Sugden S. *The Parachor and Valency*. London: Routledge Sons, Ltd; 1930.
34. Gharagheizi F, Eslamimanesh A, Mohammadi AH, Richon D. Determination of parachor of various compounds using an artificial neural network–group contribution method. *Ind Eng Chem Res.* 2011;50:5815–5823.
35. Gharagheizi F, Eslamimanesh A, Mohammadi AH, Richon D. QSPR approach for determination of parachor of chemical compounds. *Chem Eng Sci*. 2011;66:2959–2967.
36. Bayliss NS. Atomic radii from parachor data and from electron diffraction data. *J Am Chem Soc*. 1937;59:444–447.
37. Schechter DS, Guo B. Parachors based on modern physics and their uses in IFT prediction of reservoir fluids. In: SPE 30785 Proceedings of the Annual Conference, Dallas, TX; 1995.
38. Baker O, Swerdloff W. Calculations of surface tension-3: Calculations of surface tension parachor values. *Oil Gas J*. 1955;43:141.
39. Fanchi JR. Calculation of parachors for compositional simulation, an update. *SPE Res Eng Aug*. 1990;5:433–436.
40. Quayle R. The parachor of chemical compounds. *Chem Rev*. 1953;53439–53589.
41. Weinaug CF, Kalz DL. Surface tensions of methane-propane mixtures. *Ind Eng Chem*. 1943;35:239–246.
42. Lee ST, Chien MCH. A new multicomponent surface tension correlation based on scaling theory. SPE/DOE Symposium on EOR, SPE/DOE 12643, Tulsa, OK; April 15–18, 1984.
43. Hugill JA, van Welsenes AJ, Surface tension: A simple correlation for natural gas + condensate systems. *Fluid Phase Equilib*. 1986;29:383–390.
44. Gasem KAM, Dulcamara PB, Dickson BK, Robinson RL Jr. Test of prediction methods for interfacial tensions of CO, and ethane in hydrocarbon solvents. *Fluid Phase Equilib*. 1989;53;39–50.
45. Fanchi, JR. Calculation of parachors for compositional simulation. *JPT*. 1985;85:2049–2050.
46. Ali JK. Prediction of parachors of petroleum cuts and pseudo-components. *Fluid Phase Equilib*. 1994;95:383–398.
47. Brock JR, Bird RB. Surface tension and the principle of corresponding states. *AIChE J*. 1955;1:174–177.
48. Curl RF, Pitzer K. Volumetric and thermodynamic properties of fluids-enthalpy, free energy, and entropy. *Ind Eng Chem*. 1958;50: 265–274.
49. Pitzer KS. *Thermodynamics*. 3d ed. New York: McGraw-Hill; 1995.
50. Zuo YX, Stenby EH. Corresponding-states and parachor models for the calculation of interfacial tensions. *Can J Chem Eng*. 1997;75:1130–1137.
51. Rice P, Teja AS. A generalized corresponding-states method for the prediction of surface tension of pure liquids and liquid mixtures. *J colloid Interf Sci*. 1982;86:158–163.
52. Sastri SRS, Rao KK. A simple method to predict surface tension of chemical liquids *Chem Eng J*. 1995;59:181–186.
53. Riedel L. Eine neue universelle Dampfdruckformel. Untersuchungen über eine Erweiterung des Theorems der übereinstimmenden Zustände. Teil I. *Chem Ing Tech*. 1954;26:83–89.
54. Weng G, Park S, Lukes JR, Tien CL. Molecular dynamics investigation of thickness effect on liquid films. *J Chem Phys*. 2000;113:5917–5923.
55. Enders S, Kahl H, Mecke M, Winkelmann J. Molecular dynamics simulation of the liquid-vapor interface: I. The orientational profile of 2-center Lennard-Jones and of Stockmayer fluid molecules. *J Mol Liquids*. 2004;115:29–39.
56. Harris JG. Liquid-vapor interfaces of alkane oligomers: structure and thermodynamics from molecular dynamics simulations of chemically realistic models. *J Phys Chem*. 1992;96:5077–5086.
57. Alejandre J, Tildesley DJ, Chapela GA. Molecular dynamics simulation of the orthobaric densities and surface tension of water. *J Chem Phys*. 1995;102:4574–4583.
58. Nijmeijer MJP, Bakker AF, Bruin C, Sikkenk JH. A molecular dynamics simulation of the Lennard-Jones liquid-vapor interface. *J Chem Phys*. 1988;89:3789–3792.
59. Dunikov DO, Malyshenko SP, Zhakhovskii VV. Corresponding states law and molecular dynamics simulations of the Lennard-Jones fluid. *J Chem Phys*. 2001;115:6623–6631.
60. Barker JA. Surface tension and atomic interactions in simple liquids. *Mol Phys*. 1993;80:815–820.
61. Sinha BS, Dhir VK, Freund J, Darve E. Surface tension evaluation in Lennard-Jones fluid system with untruncated potentials. In: Proceedings of ASME Summer Heat Transfer Conference; July 21–28, 2003, Las Vegas, NV.
62. Holcomb CD, Clancy P, Zollweg JA. Global corresponding states representation of the interfacial tension and capillary constant for the binary mixtures argon + krypton, methane + krypton, and krypton + ethane. *Mol Phys*. 1993;78:437–459.
63. Mecke M, Winkelmann J, Fischer J. Molecular dynamics simulation of the liquid-vapor interface: The Lennard-Jones fluid. *J Chem Phys*. 1997;107:9264–9270.
64. Binder K, Muller M. Computer simulation of profiles of interfaces between coexisting phases: Do we understand their finite size effects? *Int J Mod Phys C*. 2000;11:1093–1108.
65. Benet J, MacDowell LG, Menduin C. Liquid vapor phase equilibria and surface tension of ethane as predicted by the TraPPE and OPLS models. *J Chem Eng Data*. 2010;55:5465–5470.
66. Tolman RC. The superficial density of matter at a liquid-vapor boundary. *J Phys Chem*. 1949;17:118–127.
67. Sher I, Haber S, Hetsroni G. A new state model of liquid-vapor interfaces to yield analytical expression for surface tension. *Chem Eng Sci*. 2005;60:711–716.
68. Abbas S, Ahlstro P, Nordhölm S. Estimation of the surface tension of polar fluids long-range contributions. *Langmuir*. 1998;14:396–406.
69. Do DD, Ustinov E, Do HD. Phase equilibria and surface tension of pure fluids using a molecular layer structure theory (MLST) model. *Fluid Phase Equilib*. 2003;204:309–326.
70. Davis HT, Scriven LE. A simple theory of surface tension at low vapor pressure. *J Phys Chem*. 1976;80:2805–2806.
71. Gene Expression Programming, http://en.wikipedia.org/wiki/Gene_expression_programming. Accessed August 2011.
72. Project 801. Evaluated Process Design Data, Public Release Documentation, Design Institute for Physical Properties (DIPPR). American Institute of Chemical Engineers (AIChE); 2006.
73. Cai W, Pacheco-Vega A, Sen M, Yang KT. Heat transfer correlations by symbolic regression. *Int J Heat Mass Transfer*. 2006;49:4352–4359.
74. Goldberg DE, Deb K. *A Comparison of Selection Schemes used in Genetic Algorithms*. In: Rawlins GJE, ed. *Foundations of Genetic Algorithms*. Waltham, MS: Morgan Kaufman Publishers; 1991:69–93.
75. Poli R, Langdon WB, McPhee NF. *A Field Guide to Genetic Programming*. R Poli R, Langdon WB, Nicholas F, eds. San Francisco CA: McPhee Publisher; 2008.

Manuscript received Jan. 20, 2012, and revision received Apr. 4, 2012.