Fluid Phase Equilibria 329 (2012) 71–77

Contents lists available at SciVerse ScienceDirect

Fluid Phase Equilibria

journal homepage: www.elsevier.com/locate/fluid

Gene expression programming strategy for estimation of flash point temperature of non-electrolyte organic compounds

Farhad Gharagheizi[a,*], Poorandokht Ilani-Kashkouli[a], Nasrin Farahani[b], Amir H. Mohammadi[c,d,**]

[a] *Department of Chemical Engineering, Buinzahra Branch, Islamic Azad University, Buinzahra, Iran*
[b] *Department of Chemistry, Buinzahra Branch, Islamic Azad University, Buinzahra, Iran*
[c] *MINES ParisTech, CEP/TEP – Centre Énergétique et Procédés, 35 Rue Saint Honoré, 77305 Fontainebleau, France*
[d] *Thermodynamics Research Unit, School of Chemical Engineering, University of KwaZulu-Natal, Howard College Campus, King George V Avenue, Durban 4041, South Africa*

ARTICLE INFO

Article history:
Received 13 January 2012
Received in revised form 30 April 2012
Accepted 21 May 2012
Available online 28 May 2012

Keywords:
Gene expression programming
Flammability characteristics
Flash point
Genetic programming
Genetic algorithm

ABSTRACT

The accuracy and predictability of correlations and models to determine the flammability characteristics of chemical compounds are of drastic significance in various chemical industries. In the present study, the main focus is on introducing and applying the gene expression programming (GEP) mathematical strategy to develop a comprehensive empirical method for this purpose. This work deals with presenting an empirical correlation to predict the flash point temperature of 1471 (non-electrolyte) organic compounds from 77 different chemical families. The parameters of the correlation include the molecular weight, critical temperature, critical pressure, acentric factor, and normal boiling point of the compounds. The obtained statistical parameters including root mean square of error of the results from DIPPR 801 data (8.8, 8.9, 8.9 K for training, optimization and prediction sets, respectively) demonstrate improved accuracy of the results of the presented correlation with respect to previously-proposed methods available in open literature.

1. Introduction

The term flash point (FP) refers to the lowest temperature at which a liquid gives off sufficient vapor to form an ignitable mixture with air near the surface of the liquid or within the vessel used [1]. FP values are essential information for the safe transportation, storage, and use of combustible liquids [2–5].

Experimental measurement of FP is expensive and may contain high practical uncertainties. Therefore, calculation of FP of various compounds has been the subject of many theoretical studies in order to develop accurate models. These investigations can be classified into three main categories: “empirical correlations”, “quantitative structure-property relationship (QSPR)” models, and the well-known “group contribution” methods. It should be noted that the latter is a special form of QSPRs; however, they are considered as a different class due to their easy to use nature and wide range of applications. Good reviews are available in the literature for various methods proposed for FP [6–8].

The first category contains those correlations that need at least one of the other physical properties such as normal boiling point, density, vapor pressure, critical properties, and enthalpy of vaporization [8]. To refer some of those models that lie in this class, we can refer to those empirical correlations proposed by Prugh (200 compounds, average absolute error 11 K, and maximum deviation 500%) [9], Fuji and Herman (168 compounds, for 89% of compounds within ±10 K) [10], Patil (950 compounds, several models with AARD% of 10%) [11], Suzuki et al. (400 compounds, average absolute error 13.52 K) [12], Satyrayana and Kakati (250 compounds, AARD% of 8.3%, maximum deviation 32.72%) [13], Satyrayana and Rao (1221 compounds, several correlations) [14], Metcalfe and Metcalfe (201 compounds, average absolute error 8.6 K, maximum error 26.2 K) [15], Hshieh (207 compounds, average absolute error 11.06 K) [16], Catoire and Naudet (evaluated using 1471 compounds: [4] AARD of 2.44% and average absolute error of 8.28 K) [7] and Gharagheizi et al. (the former model: AARD of 2.4% and average absolute error of 8.06 K; the latter model: AARD of 2.14%) [4,5].

The second category is the QSPR models in which FP is correlated using some chemical structure-based parameters called “molecular descriptors”. These correlations just relate the FP to the chemical structure and do not need any other physical properties. We can refer to the QSPR models presented by Tetteh et al. (400 compounds, average absolute error of approximately 11 K) [17], Katritzky et al. (the former: 271 and compounds, root mean square error of 23.03 K; the latter: 758 compounds, AARD of 3.49% and average absolute error of 10.65 K) [18,19], Gharagheizi and Alamdari (1378 compounds, AARD of 10.2%) [2]. There are numerous

* Corresponding author. Fax: +98 21 77926580.
** Corresponding author. Tel.: +33 1 64694970; fax: +33 1 64694968.
E-mail addresses: fghara@gmail.com (F. Gharagheizi), amir-hossein.mohammadi@mines-paristech.fr (A.H. Mohammadi).

0378-3812/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.fluid.2012.05.015

studies in this category; however, the general models are regarded in this study. It is obvious that developing this type of correlations is much more difficult than the ones for particular chemical families such as hydrocarbons. The most important drawback of the QSPR models is the complex procedure of calculation of the molecular descriptors from chemical structure. As a result, these correlations are not simple to use.

The third category includes the group contribution models (GC). In this kind of methods, FP is correlated with the number of occurrences of some chemical substructures. It seems the only general model from this class is the one proposed by the first author and his co-workers (1030 compounds, AARD of 10.2%) [3]. Of course, there are some other GC methods that proposed just for some particular chemical families. They are not considered in this study. Perhaps, the only important weak point of the GC models is the large number of parameters that they require. In addition, in the recently proposed version of GC, namely, artificial neural network-group contribution (ANN-GC), the complexity of the model is another issue.

A comprehensive comparison between these three categories is pretty difficult because there are several factors to be taken into account, for instance, simplicity of the model, accuracy of the model, simplicity of the parameters, and the comprehensiveness of the method for covering the wider applicability domain. The latter includes both the number of compounds, and the diversity of chemical compounds employed while developing and validating the model.

According to the reported statistical parameters of the models, the first category seems to be more convincing than others due to the simplicity basis, accuracy and comprehensiveness.

As a result of statistical parameters of the models, it can be concluded that despite significant progress in the estimation of FP using the QSPR and GC methods, the empirical correlations give more comprehensive and more accurate results. The latter can normally give acceptable results within the range of the conditions and the compounds, implemented for their development. Semi-empirical correlations use some theoretical basis in the form of parameters to improve the prediction capabilities.

In any case, certain parameters of the aforementioned correlations should be regressed over the experimental data. Many mathematical methods, including linear/nonlinear regression methods and various kinds of optimization techniques have been so far proposed for this purpose.

The genetic algorithm (GA), firstly introduced by Holland [20], is considered as a heuristic optimization technique (among the evolutionary algorithms) that follows the process of natural evolution. It generally generates solutions (chromosomes) to optimization problems through specific operators like selection, mutation, and crossover [21]. The final solutions 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) [22] 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 [22]. Considering the drawbacks of the GP (which will be discussed later), Ferreira [21] introduced a very fruitful modification to the original GP algorithm [22]. In the new strategy, called "gene expression programming (GEP)" [23], 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 [21,23]. The description of the GEP [21] algorithm is given in the next section.

The GEP [21] strategy has been, up to now, implemented for several electrical, mechanical purposes such as development of stage-discharge curves of rivers [24] and splitting tensile strength of concrete [25]; it is of great interest to employ the same algorithm for determination of the flammability characteristics of chemical compounds such as FP. It should be noted that our group has very recently applied the method for development of some corresponding states models for thermal conductivity of gases [26], viscosity of gases [27], and solubility parameters [28].

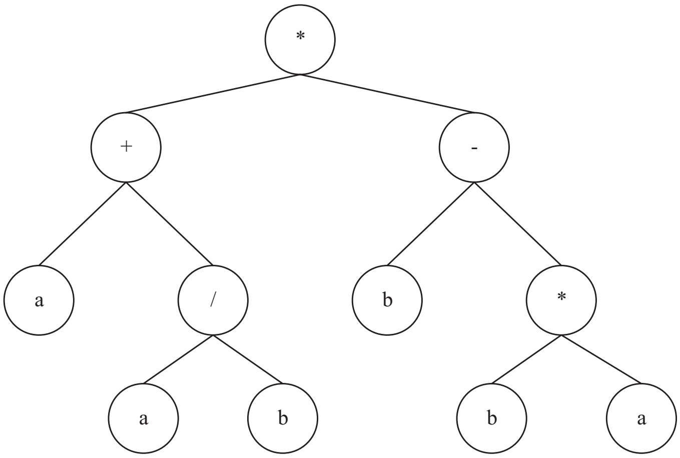

Fig. 1. A typical computer LISP program in the GP algorithm represented as a parse tree (expression tree), which stands for the algebraic expression $[a+(a/b)]\times[b-(a\times b)]$ by a two-gene chromosome.

2. Mathematical strategy

2.1. Genetic programming

As mentioned earlier, the GP [22] 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 [22,23]. The main distinction between the GP [22] and the GA [20] is that in the former, the chromosomes consist of nonlinear structures similar to parse trees though they are similar to the GA [22] linear structures, which are naked replicators working as genotype and phenotype [21]. 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 (such as recombination, crossover, and mutation) also operate during the computational steps of the GP similar to the original GA [20] but they resemble to pruning and grafting of trees [21]. As indicated by Ferreira [21], the main disadvantages of the GP is that the complex replicators (parse trees structures) can be only modified in limited ranges because their reproduction should be done only on the parse trees. These modifications include modifying or exchanging definite branches of the corresponding parse trees [21], that may be occasionally lead to invalid (unacceptable) trees structures. A typical computer LISP program based on the GP [22] algorithm is shown in Fig. 1. It should be noted that the GP utilizes these kinds of computer programs for data representation.

2.2. Gene expression programming

The GEP [23], resulting from the modification and extension of the GP [22], is applying computer programs in order to solve a problem. In the latter technique, the population individuals are symbolic expression trees unlike those of GEP [21], in which the individuals are encoded as linear chromosomes, which are later translated into the expression parse trees, i.e. the genotype and phenotype are eventually separated one another. As a consequence, the GEP