Scientia Iranica C (2013) 20 (3), 543–548

Sharif University of Technology

Scientia Iranica

Transactions C: Chemistry and Chemical Engineering

www.sciencedirect.com

Modeling intermolecular potential of He–F_2 dimer from symmetry-adapted perturbation theory using multi-gene genetic programming

M. Amiri [a,*], M. Eftekhari [b], M. Dehestani [a], A. Tajaddini [c]

[a] *Department of Chemistry, Shahid Bahonar University of Kerman, Kerman, P.O. Box: 7616913, Iran*

[b] *Department of Computer Engineering, Shahid Bahonar University of Kerman, Kerman, P. Code: 7616913111, Iran*

[c] *Department of Mathmatics, Shahid Bahonar University of Kerman, Kerman, P.O. Box: 76169-14111, Iran*

Received 16 April 2012; revised 18 August 2012; accepted 31 December 2012

KEYWORDS

Potential energy;
SAPT;
MGGP;
Lennard-Jones potential.

Abstract Any molecular dynamical calculation requires a precise knowledge of interaction potential as an input. In an appropriate form, such that the potential, with respect to the coordinates, can be evaluated easily and accurately at arbitrary geometries (in our study parameters for geometry are R and θ), a good potential energy expression can offer the exact intermolecular behavior of systems. There are many methods to create mathematical expressions for the potential energy. In this study for the first time, we utilized the Multi-gene Genetic Programming (MGGP) method to generate a potential energy model for the He–F_2 system. The MGGP method is one of the most powerful methods used for non-linear regression problems. A dataset of size 714 created by the SAPT 2008 program is used to generate models of MGGP. The results obtained show the power of MGGP for producing an efficient nonlinear regression model, in terms of accuracy and complexity.

1. Introduction

Potential Energy (PE) is the energy stored in a molecule. This energy also is the portion of the energy of a system which is associated with its position in a force field [1]. Usually, in chemistry, when we talk about PE, it is related to the energy that was created because of the configuration of the molecule. When we have two parts, like our system (He–F_2), PE will change the change of geometry parameters (R, θ), or, in other words, with a change in configuration. The most stable configuration has the smallest potential energy. Our knowledge about the most stable configuration is very essential. For example, to propose a good mechanism for chemistry reactions we must have correct information about changes of potential energy in order to change configuration.

An accurate model of PE is necessary in scattering calculation, and calculation of the second virial coefficients, etc. For example, to calculate second virial coefficients $(B_{12}(T))$, we must evaluate the following integral:

$$B_{12}(T) = \pi N_A \int_0^{\infty} \int_0^{\pi} \{1 - \exp[-V(R,\theta)/kT]\} \times R^2 \sin\theta dR d\theta, \quad (1)$$

where N_A is Avogadro's number and k is Boltzmann's constant. $V(R, \theta)$ is potential energy expression.

In the present study, we obtained several models for the potential energy of the He–F_2 system by the MGGP method. MGGP is a promising variant of genetic programming (GP). Multi-gene Genetic Programming (MGGP) effectively combines the model structure selection ability of the standard GP with the parameter estimation power of classical regression to capture nonlinear interactions [2]. In order to achieve this purpose, 714 data, collected by the SAPT 2008 program, are used. SAPT2008

* Corresponding author. Tel.: +98 341 3202 106; fax: +98 341 3222 033.
E-mail addresses: mohammadamiri1985@yahoo.com (M. Amiri), m.eftekhari@uk.ac.ir (M. Eftekhari), dehestani2002@yahoo.com (M. Dehestani), atajadini@uk.ac.ir (A. Tajaddini).

Peer review under responsibility of Sharif University of Technology.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.scient.2012.12.040

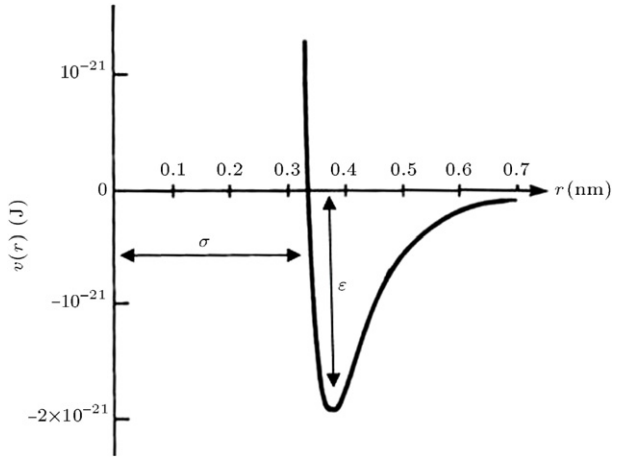


Figure 1: The potential energy for Ar–Ar system.

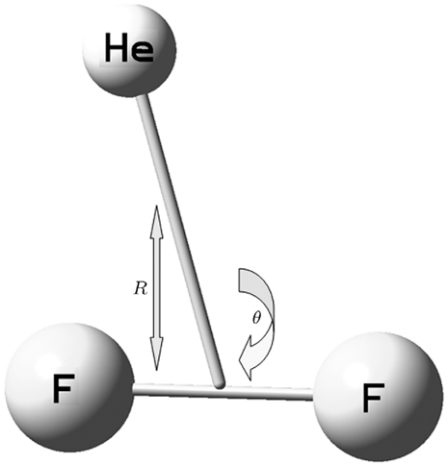


Figure 2: The coordinate system for He–F_2.

is a computer program implementing the many-body version of the symmetry-adapted perturbation theory (SAPT). SAPT is designed to calculate the interaction energy of a dimer, or a system consisting of two arbitrary monomers. Each monomer can be an atom or a molecule [3]. Obtained equations are suitable models in terms of both accuracy and simplicity.

2. Potential energy and computational details

2.1. Potential energy with one variable

Figure 1 shows the typical behavior of PE as a function of r (the intermolecular distance for a diatom like Ar–Ar) nearby $r = r_e$ (r_e or $r_{\text{equilibrium}}$ is the intermolecular distance, whose potential energy, at this distance, has the smallest energy). The curve is similar to a parabola, and, therefore, it is represented by the following equation:

$$V = 1/2kx^2, \tag{2}$$

where $x = r - r_e$ and $k = (d^2V/dr^2)_{r=r_e}$. Eq. (2) is the familiar harmonic oscillator potential. The harmonic oscillator potential cannot be used for modeling the interactions between internal atoms of a molecule and an external atom, because, as indicated in Figure 1, with an increase in r, the behavior of the potential energy does not show the behavior of the harmonic oscillator. For this reason, we have to add cubic, quarter and other higher order terms to the Eq. (2). The resulted equation is an anharmonic oscillator potential.

The general function is as follows [4]:

$$V(x) = d_0x^2\left(1 + \sum_{i=1}^{\infty} d_ix^i\right), \tag{3}$$

where $x = r - r_e$.

2.1.1. Lennard-Jones potential

A simple but practical function for estimating the interaction energy between two neutral atoms or molecules is the Lennard-Jones potential, as [1]:

$$V(R) = 4\varepsilon[(\sigma/R)^m - (\sigma/R)^n], \tag{4}$$

where constants ε and σ are the depth of potential energy curve and the distance at which the potential energy vanishes, respectively, as shown in Figure 1. There is a small flexibility in the functional form, except when the powers m and n are 12 and 6, respectively. This model has a minimum number of parameters to show the PE curve. Because there are not enough parameters to adequately reproduce an exact potential, the Lennard-Jones potential is not a precise model [4]. One action undertaken in this study is to adjust the Lennard-Jones potential for creating new and exact models for our collected data.

2.2. Potential energy with two variables

Introducing the PE expression for a system with two variables is much more complicated than a system with one variable. Our system, He–F_2, has two variables. An atom (He) + diatomic (F–F) is like a linear rigid rotor. Figure 2 shows the coordinate system for the He–F_2.

The PE model, $V(R, \theta)$, is expanded in terms of Legendre polynomials (P_λ):

$$V(R, \theta) = \sum_{\lambda} V_\lambda(R)P_\lambda(\cos\theta). \tag{5}$$

Coefficient V_λ is dependent only on R [4].

3. Computational details

The coordinate system for the He–F_2 compound is shown in Figure 2, where R is the intermolecular distance between the center mass of F_2 and He, r is the F–F bond length and θ is the angle between R and r.

We used 714 data points in the following order to make models for the potential energy surface:

$R = 1.6(0.25)2, 2(0.1)3, 3.05, 3.1, 3.15, 3.2, 3.38, 3.4, 3.6, 3.8, 4, 4.5, 5, 6$ Å,

$\theta = 0^\circ(5)90^\circ, 57.5^\circ, 87.5.$

For the first time, we calculated potential energy ($V(R, \theta)$) by the SAPT2008 program [3] and cc-pVQZ-F12 is used as a basis set [5]. In a previous study, potential energy was calculated by the super molecular approach [6], which has less accuracy in comparison with the symmetry-adapted perturbation theory (SAPT) approach.

The resulted energies in this research (calculated by SAPT2008) and in two previous research works, which used a super molecular approach, are compared in Table 1.

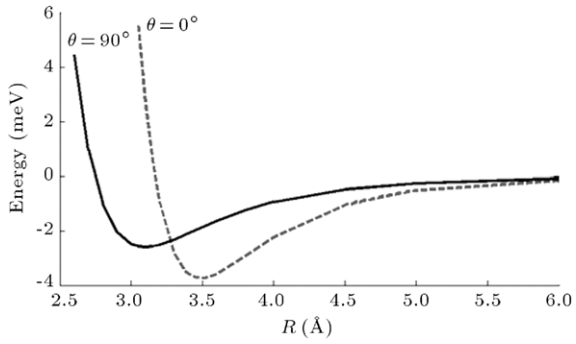


Figure 3: Dependence of the potential energy surface of the He–F_2 on the inter monomer distance R at $r = 1.412$ Å, $\theta = 0°$ and $\theta = 90°$.

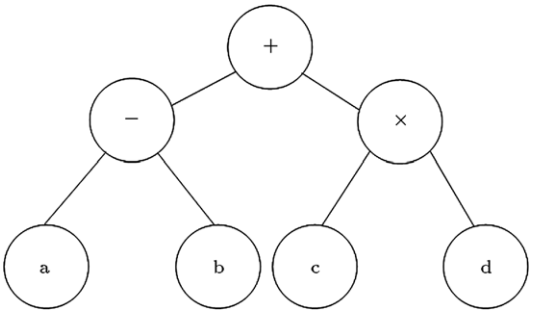


Figure 4: Tree structure of the expression $(a - b) + (c \times d)$.

Table 1: Potential energies He–F_2 in comparison with two previous works at $r = 1.412$ Å. All energies are in micro hartrees.

	$\theta = 0°$		$\theta = 90°$	
R (Å)	Our work	Chan et al. [6]	Our work	Chan et al. [6]
1.60	301688.28	–	31008.62	–
2.00	72115.92	–	5693.13	–
2.40	12443.73	–	737.20	–
2.80	1505.57	–	−37.03	–
3.00	347.90	314.80	−91.15	−145.70
3.25	−75.47	−108.00	−89.83	−124.50
3.50	−137.38	−162.30	−67.65	−91.70
3.75	−114.23	−131.40	−47.75	−64.20
4.00	−82.04	−92.80	−33.20	−44.50
4.50	−38.53	−42.80	−17.09	−21.90
5.00	−18.79	−20.60	−9.03	−11.50
6.00	−5.48	−6.00	−2.98	−3.70

Figure 3 shows the potential energy curve in linear ($\theta = 0$) and T-shaped configurations. It can be seen that linear configuration has less energy, therefore, is more stable than T-shaped configuration.

4. Genetic programming

Genetic Programming (GP), which was first offered by Koza [7], is a biologically inspired machine learning method. GP and Genetic Algorithm (GA) are very similar to each other. In GP, at first, a population of computer programs which is represented by a tree structure (Figure 4) is generated, then mutation and crossover are done on the best selected trees to create a new population, this process is repeated until a maximum number of generations is obtained. The GP method is usually called symbolic regression [8]. Figure 5 shows this process as a flowchart.

Figure 4 shows the structure of a tree which exhibits the mathematical expression $(a - b) + (c \times d)$. Trees are flexible structures through which logical expressions or mathematical relations could be appropriately shown. Leaves of the tree usually specify variables or constants and are chosen from a pre-defined terminal set, while other nodes specify operators or functions and are chosen from a pre-defined function set. In the tree of Figure 4, a, b, c, d and $+$, $-$, $\times$ are used as variables and operators, respectively. The division operator and variable e are not used in the tree. Pre-defined sets are as follows:

$$\text{function set} = \{+, -, \times, /\}, \tag{6}$$

$$\text{terminal set} = \{a, b, c, d, e\}. \tag{7}$$

4.1. Multi-gene GP

In this work, Multi-Gene GP (MGGP), one newly developed version of GP, is utilized for performing efficient modeling via symbolic regression. In Multi-gene GP, each chromosome consists of some "genes" and, thereby, each model is a weighted linear combination of these genes. Each of the genes is a standard GP tree.

In Figure 6 there are two genes or GP trees, x_1 and x_2 are input variables. y is a linear combination of two genes, and d_0, d_1 and d_2 are the weights that are obtained by least squares. In the present study, we used GPTIPS (a toolbox written for Matlab) for obtaining one efficient regression model.

5. Simulation results

5.1. Parameters setup

In the present study, we used GPTIPS to create a mathematical expression for the PE of the He–F_2 system. Parameters of algorithms are as follows. The number of generations, $G = 300$, size of population, $Popsize = 200$, x-over probability, $Pc = 0.7$, and mutation probability, $Pm = 0.01$. Also, function set $= \{+, -, /, \hat{}, \sin, \sinh, \tan, \tanh, \cos\}$, and terminal set $= \{R, \theta\}$.

5.2. Results and discussion

The produced expression by Multi-gene GP is given in the following:

$$\begin{aligned} V(R,\theta) = & -((24.13\tan(\tanh(R_2))) + (24.13(R + \theta^{1/4})) \\ & \times ((0.4739\theta) - \cos(R) + 12.57))/\sinh(R^2) \\ & - (1091|\theta - 0.2231|^{1/2})\tan(\sinh(\sin(\theta^{1/2}))) \\ & \times (\sin(\theta^{1/2}) + 0.5119)/(5\sinh(R^2)) + 0.6577 \\ & + ((188000\tanh((0.2813\theta) + 1.6062) \\ & \times (\tanh((0.0325\theta) + \sinh(R)) \\ & - 0.6309)))/\sinh(R^2). \end{aligned} \tag{8}$$

The correlation coefficient (R^2) for the obtained model is 0.99992. Figure 7 depicts the real data (resulted from SAPT program) versus the predicted data.

5.3. Adjusting Lennard-Jones parameters

In this study, also, we adjusted and optimized Lennard-Jones potential's parameters for our system.

At first, we tried to obtain relations for m and n in Lennard-Jones potential equation. Parameters of algorithms were set as follows. The number of generations, $G = 320$, size of population, $Popsize = 220$, x-over probability, $Pc = 0.7$, and mutation probability, $Pm = 0.01$. Also, function set $= \{+, \times, \sin,$

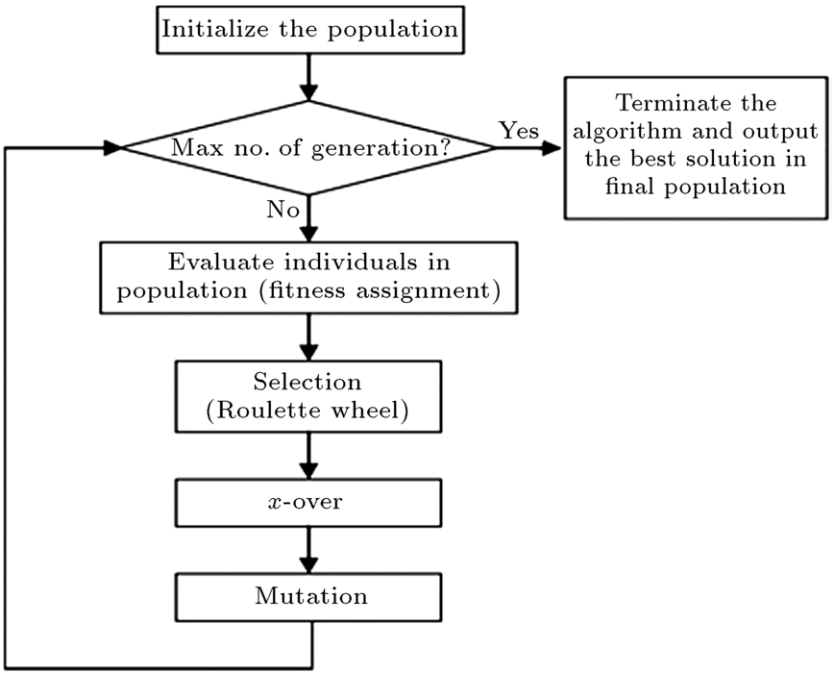
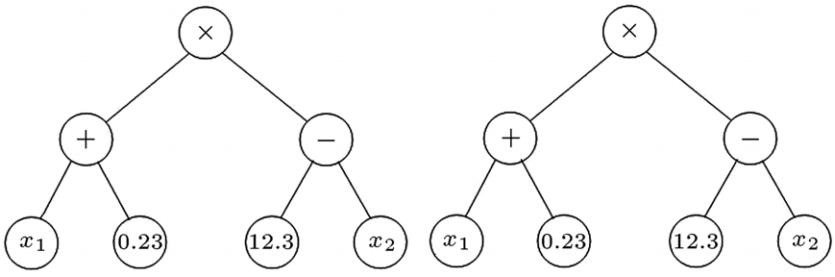


Figure 5: Flowchart of GA and GP algorithms.



Multiple gene model:

$$y = d_0 + d_1((x_1 + 0.23)(12.3 - x_2)) + d_2((0.71 - x_1)\sin(x_2))$$

Figure 6: A typical multiple gene model.

sinh, tan, tanh, cos}, and terminal set $= \{R, \theta\}$. In the following, the derived relations for m and n are given, respectively:

$$m = 6.48, \tag{9}$$

$$n = 3.2R + 5.164, \tag{10}$$

and the modified Lennard-Jones equation is represented as :

$$V(R, \theta) = 4\varepsilon(\theta)[(\sigma(\theta)/R)^{6.48} - (\sigma(\theta)/R)^{3.2R+5.164}]. \tag{11}$$

Values of $\varepsilon(\theta)$ and $\sigma(\theta)$ for some angle are listed in Table 2.

Figure 8 shows the accuracy of predictions of Eq. (11), with regard to real values (E_{SAPT}).

As the third idea, we tried to achieve an accurate equation by adjusting four parameters of the Lennard-Jones potential equation, namely; m, n, ε and σ. The results are given in the following:

$$m = \cos(4.997\theta) + 5.452, \tag{12}$$

$$n = 2.0\theta + \cos(9.994R), \tag{13}$$

ε and σ are constant values, 4.997 and 6.431, respectively. Accordingly, the final equation is represented as:

$$V(R, \theta) = 4(4.997)[(6.431/R)^{\cos(4.997\theta)+5.452} - (6.431/R)^{2.0\theta+\cos(9.994R)}]. \tag{14}$$

Table 2: Values of two parameters, $\varepsilon(\theta)$ and $\sigma(\theta)$.

θ (°)	ε (meV)	σ (Å)
0	−3.59	3.17
35	−2.01	3.28
45	−1.75	3.26
50	−1.71	3.25
70	−1.88	2.98
85	−2.50	2.77
90	−2.58	2.72

Parameters of the algorithm are as follows. The number of generations, $G = 320$, size of population, $Popsize = 220$, x-over probability, $Pc = 0.7$, and mutation probability, $Pm = 0.01$. Also, function set $= \{+, \times, \cos\}$, and terminal set $= \{R, \theta\}$.

The accuracy of Eq. (14), with regard to real values (E_{SAPT}), is shown in Figure 9.

6. Second virial coefficients

The second virial coefficients were calculated from the calculated potential energy Eq. (8), and Gauss and Ramberg

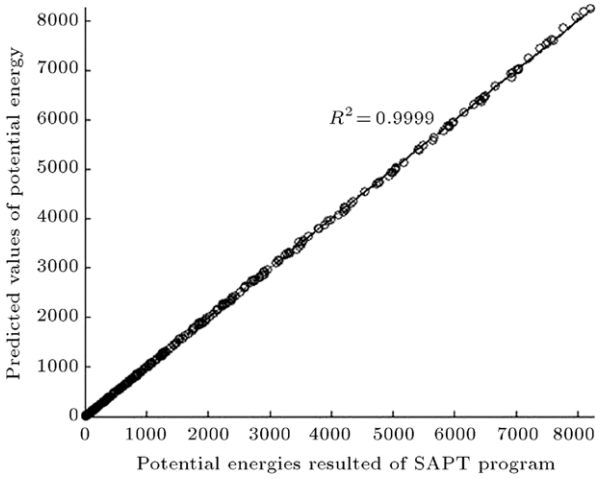


Figure 7: Predicted potential energies using Eq. (8) versus values computed by SAPT program.

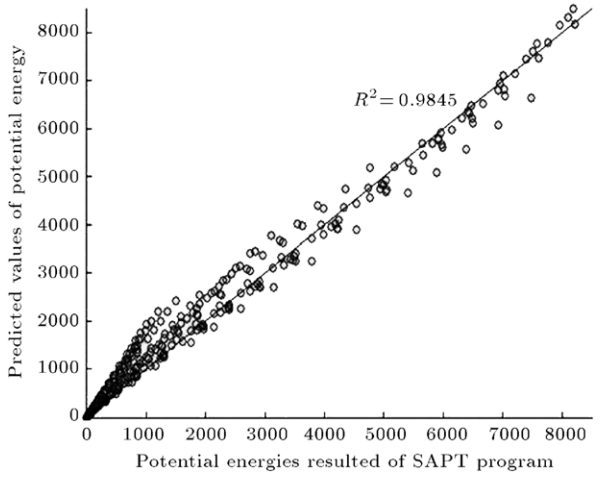


Figure 8: Predicted potential energies using Eq. (11) versus values computed by SAPT program.

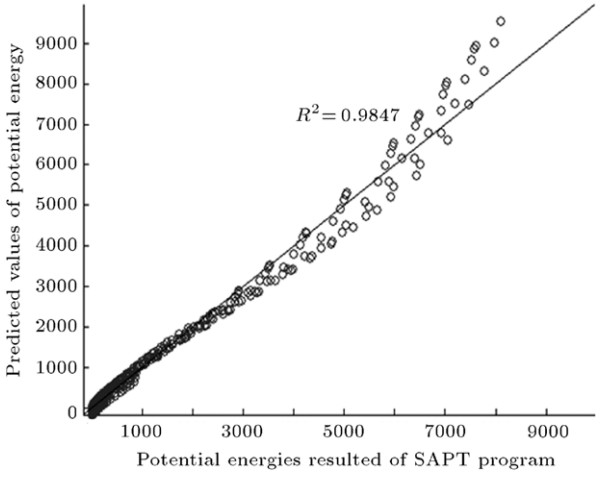


Figure 9: Predicted potential energies using Eq. (14) versus values computed by SAPT program.

methods [9] are used for evaluating the integral (Eq. (1)). The results are shown in Figure 10. The figure for Eq. (13) is similar

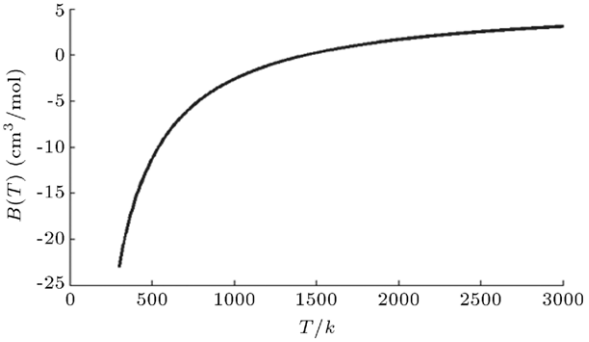


Figure 10: Second virial coefficients resulted of Eq. (8).

to Eq. (8). Unfortunately, there are no second virial coefficients in previous work to compare with our results.

7. Conclusion

This study can be divided into two parts, generally. First, chemistry computation, in which we applied the SAPT approach using the SAPT2008 program to achieve potential energies, He–F_2, for the first time. The SAPT method has good accuracy in comparison with the super molecular method used by previous researchers [6]. In the second step, to create models for achieved energies in the previous part, the MGGP method was utilized. In order to create simple and precise models, the MGGP was employed, both for modeling PE Eq. (8) and for adjusting the Lennard-Jones potential (Eqs. (11) and (14)). It is reasonable that a simple equation is more desirable than a complex one because of its application; for example, calculation of the second virial coefficients.

References

[1] Lide, D.R., *Handbook of Chemistry and Physics*, 84th Edn., pp. 2–55, CRC, Boca Raton (2004).
[2] Gandomi, A.H. and Alavi, A.H. "A new multi-gene genetic programming approach to nonlinear system modeling. Part I: materials and structural engineering problems", *Neural Computing & Applications*, 21, pp. 171–187 (2012).
[3] Bukowski, R., Cencek, W., Jankowski, P., Jeziorski, B., Jeziorska, M., Kucharski, S.A., Lotrich, V.F., Misquitta, A.J., Moszynski, R., Patkowski, K., Podeszwa, R., Rybak, S., Szalewicz, K., Williams, H.L., Wheatley, R.J., Wormer, P.E.S. and Zuchowski, P.S. "SAPT 2008: An Ab Initio Program for Many – Body Symmetry – Adapted Perturbation Theory Calculations of Intermolecular Interaction energies" (2008).
[4] Sathyamurthy, N. "Computational fiting of ab initio potential energy surfaces", *Computer Physics Reports*, 3, pp. 1–70 (1985).
[5] Bischoff, A.F., Wolfsegger, S., Tew, D.P. and Klopper, W. "Assessment of basis sets for F12 explicitly-correlated molecular electronic-structure methods", *Molecular Physics*, 107, pp. 963–975 (2009).
[6] Chan, K.W., Power, T.D., Jai-nhuknan, J. and Cybulski, S. "An ab initio study of He–F_2, Ne–F_2, and Ar–F_2 van der Waals complexes", *Journal of Chemical Physics*, 110, pp. 860–869 (1999).
[7] Koza, J.R., *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, 2nd Edn., MIT, Massachusetts, USA (1992).
[8] Searson, D.P., Leahy, D.E. and Willis, M.J. "GPTIPS: An open source genetic programming toolbox for multigene symbolic regression", 1st Int. Conf. on Multiconference of Engineers and Computer Scientists, 1, Hong Kong, China, pp. 77–80 (2010).
[9] Burden, R.L. and Faires, J.D. "Numerical analysis", In *Introduction to Theory and Application of Modern Numerical Approximation Techniques*, 7th Edn., pp. 207–211, Brooks Cole, California, USA (2001).

Mohammad Amiri was born in Kerman, Iran, in 1985. He received his B.S. degree in Pure Chemistry and M.S. degree in Physical Chemistry from Shahid Bahonar University, Kerman, Iran, in 2008 and 2011, respectively. His research interests include: computational chemistry, especially modeling and optimization, fuzzy logic and it's applications, and new methods for optimization, such as neural networks, genetic algorithms and GPTIPS.

Mahdi Eftekhari was born in Kerman, Iran, in 1978. He received a B.S. degree in Computer Engineering in 2001 and his M.S. and Ph.D. degrees in Artificial Intelligence from Shiraz University, Iran, in 2004 and 2008, respectively. He has been faculty member of the Computer Engineering Department at Shahid Bahonar University of Kerman, Iran, since 2008. His research interests include: fuzzy systems and modeling, evolutionary algorithms, data mining, machine learning and application of intelligent methods in bioinformatics. He is author and co-author of about 30 papers in cited journals and conferences, and a member of the Iranian Society of Fuzzy Systems.

Maryam Dehestani was born in Kerman, Iran, in 1967. She received a B.S. degree in Chemistry from Shahid Bahonar University, Kerman, Iran, in 1988, and M.S. and Ph.D. degrees in Physical Chemistry from the University for Teacher Training, Iran, in 1991 and 2001, respectively. She has been a faculty member of the Chemistry Department at Shahid Bahonar University of Kerman, Iran, since 2002. His research interests include: quantum chemistry, molecular spectroscopy, computational chemistry and nanostructures calculations. He is the author and co-author of about 90 papers in cited journals and conferences.

Azita Tajaddini was born in Kerman, Iran, in 1974. She received a B.S. degree in Applied Mathematics from Vali-e-Asr University, Rafsanjan, Iran, in 1995, and M.S. and Ph.D. degrees in Applied Mathematics from Shahid Bahonar University, Kerman, Iran, in 1997 and 2008, respectively, where she is currently faculty member in the Mathematics Department. Her research interests include: inverse eigen value problem and numerical linear algebra. She is author and co-author of 6 papers in cited journals.