International Journal of Advanced Science and Technology
Vol.63, (2014), pp.23-34
http://dx.doi.org/10.14257/ijast.2014.63.03

On Symbolic Regression for Optimizing Thermostable Lipase Production

Hossam Faris[1], Alaa Sheta[2] and Rania Hiary[3]

[1]*King Abdulla II School for Information Technology,*
The University of Jordan, Amman, Jordan
[2]*Computers and Systems Department,*
Electronics Research Institute (ERI), Cairo, Egypt
[3]*Information and Communication Technology in Education Program,*
Al-Albait University, Mafrag, Jordan

hossam.faris@ju.edu.jo, asheta66@gmail.com, rania.i.hiary@gmail.com

Abstract

Theromostable lipases have wide range of biotechnological applications in the industry. Therefore, there is always high interest in investigating their features and operating conditions. However, Lipase production is a challenging and complex process due to its nature which is highly dependent on the conditions of the process such as temperature, initial pH, incubation period, time, inoculum size and agitation rate. Efficient optimization of the process is a common goal in order to improve the productivity and reduce the costs. In this paper, we apply a Symbolic Regression Genetic Programming (GP) approach in order to develop a mathematical model which can predict the lipase activities in submerged fermentation (SmF) system. The developed GP model is compared with a neural network model proposed in the literature. The reported evaluation results show superiority of GP in modeling and optimizing the process.

Keywords: *genetic programming, symbolic regression, lipase production*

1. Introduction

In the last few years, there has been a growing commercial interest in investigating the features of theromostable lipases due to their wide range of biotechnological applications in industry and business. Theromostable lipases have some advantageous features such as:

- Extra thermostability,
- Extra resistance to many environments changes than their mesophilic homologues [1–4].
- Theromostable lipases can be produced at low cost.

Different types of thermostable enzymes have been successfully used in industrial applications, mainly as replacements for thermolabile enzymes [5–8]. On the industrial production scale, microbial extracellular enzymes show remarkable advantages especially in biotechnological applications such as diary-based products, detergents, drugs, cosmetics and leather processes [9–11].

However, optimizing lipase production and predicting its activity is a complex and challenging problem [10]. This because lipase production is highly dependent on its operating

ISSN: 2005-4238 IJAST

conditions that affect its growth. Such factors include nutritional and physico-chemical factors such as temperature, initial pH, incubation period, time, inoculum size and agitation rate [4, 11].

Therefore, adopting an optimization method and choosing a modeling technique are vital in lipase manufacturing and important for producing reliable lipase products with high standards [12, 13]. Efficient optimization and modeling is a common goal in order to improve the productivity of the process and reduce the costs [4].

In literature, Artificial Neural Networks (ANNs) are one of the most machine learning techniques applied for lipase production optimization and prediction. ANN is a mathematical model consists of a number of information processing units called "neurons". Neurons are interconnected in different ways to form the structure of the network. Usually, neurons are organized in layers. Each neuron has a number of inputs and computes a single output. Each input has a weight that expresses the importance or contribution of its corresponding input to the output. ANN performs the training process by adjusting the weights of the neuron until a minimum value of error is reached.

In [4], authors compared between response surface methodology (RSM) and ANN for prediction of extracellular thermostable lipase production based on the six independent factors mentioned earlier. In their work, they choose a multilayer full feedforward incremental backpropagation network with Gaussian transfer function. Although both approaches had good quality predictions, authors showed that ANN has a better modeling power and prediction results. However, ANNs in general have some drawbacks; ANNs relatively require large amounts of data for the training stage and they are commonly referred to as a black input/output box, it is commonly hard to interpret their results.

In another work, authors in [17] optimized the culture parameters of lipase production in submerged (SmF) and solid-state (SSF) systems using artificial neural network. They implemented two ANN based approaches; multilayer normal and full feed forward backpropagation. Authors claimed that ANN models in SmF and SSF were highly predictive and enhanced the lipase production although the obtained conditions were close together.

There were other approaches based on using Genetic Programming (GP) and Genetic Algorithms (GA) were applied for modeling lipase production. GP and GA are inspired by the evolution concepts and theories. In [3, 20], authors applied GP as evolutionary computation methodology for developing an efficient model for the fermentation process. Authors compared their results with other results obtained from traditional experimental design approaches. In [18], a modified genetic algorithm is proposed for a parameter identification of an E. coli fedbatch fermentation model. Authors made some adjustments of the genetic parameters regarding the fermentation processes to improve the conventional genetic algorithm. Authors claim that the modified GA for a parameter identification of the problem can be efficient and effective

In this paper, we investigate the use of Symbolic Regression Genetic Programming to generate a mathematical model which can predict the Lipase activities in submerged fermentation (SmF) system. The developed GP model should be able to estimate the lipase activities with high performance rates. Furthermore, a comparison between the developed GP model and neural network model proposed in [17] will be conducted. The reported evaluation results of the GP model are promising.

The rest if the paper is structured as follows; in Section 2, the GP approach applied in this research is described and discussed. Section 3 specifies evaluation criteria used in order evaluate the performance of the developed GP model. Materials and methods are listed in Section 4 and data collected is presented in Section 5. Finally, experiments and results are discussed in Section 6.

2. Symbolic Regression via Genetic Programming

The term "Symbolic regression" was introduced by J. Koza [15]. The goal of this approach is to generate a mathematical expression of any functional form which estimates the values of a specified target variable based on values of a set of input variables while minimizing some error criteria. Unlike traditional linear and nonlinear regression methods which fit parameters to an equation of a predetermined form, symbolic regression searches both the space of models along with the space of all possible parameters (coefficients) simultaneously such that it can find the best model which minimizes the error criterion. Symbolic regression has an advantage when the underlying function is quite complex [26].

Symbolic regression was used in GP to evolve a population of trees [16]. For a system with m input variables of dimension $R^{n\times m}$ to produce a model output $\hat{y}$ with dimension $R^{n\times 1}$, where n is the number of observations taken and u is an input variable, we could produce a tree structure which introduces the mathematical model:

$$\hat{y} = f(u_1, \dots, u_i) \qquad (1)$$

A. *Initial Population and Representation*

The evolutionary cycle of GP starts by generating a number of candidate models called individuals. Each individual consists of a randomly generated tree. Simple individual tree is shown in Figure 1.

GP individuals can be defined using a Terminal set *T* and a Function set *F*. The set *T* contains operands such as constants and variables, while the set *F* contains simple arithmetic operators such as *addition*, *subtraction*, *multiplication* and *division*, also it could contains other non-linear operators and functions like; *sqrt*, *exp*, *sin*, *cos*. Both sets *F* and *T* are used to develop and form tree structures which represent a model for the problem.

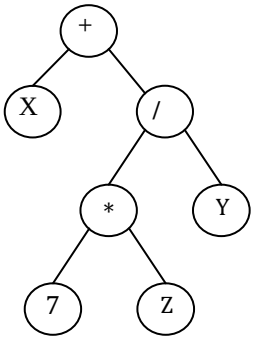

Figure 1. Example of a symbolic tree (model)

B. *Fitness evaluation*

Fitness evaluation: each individual is evaluated according to specific evaluation criteria. In our case Pearson R^2 evaluator (coefficient of determination) is applied:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y} - y)^2}{\sum_{i=1}^{n}(\bar{y} - y)^2} \qquad (2)$$

Where $\hat{y}$ is the model prediction value and y is the experimental value. The goal is to maximize R^2.

C. *Tournament selection*

In genetic programming, some individuals are selected from the population for reproduction. Tournament selection is the one of the most famous methods for selecting those individuals. In this method, *k* individuals are chosen randomly then the best individual among *k* is the chosen for reproduction. *k* is typically referred to as tournament size [23].

D. *Subtree swapping crossover*

Crossover operation selects two individuals (parents) and cuts their trees at some randomly chosen position. The subtrees are swapped to generate two new individuals (children). The following example in Figure 2 shows the operation.

E. *Mutation*

Mutation is an operator applied on one individual tree. A randomly point in the tree is chosen then the subtree under this point is replaced by a new randomely generated subtree. An example of the mutation operator is shown in Figure 3. Typically, mutation operator is performed with a probability much less than crossover [24].

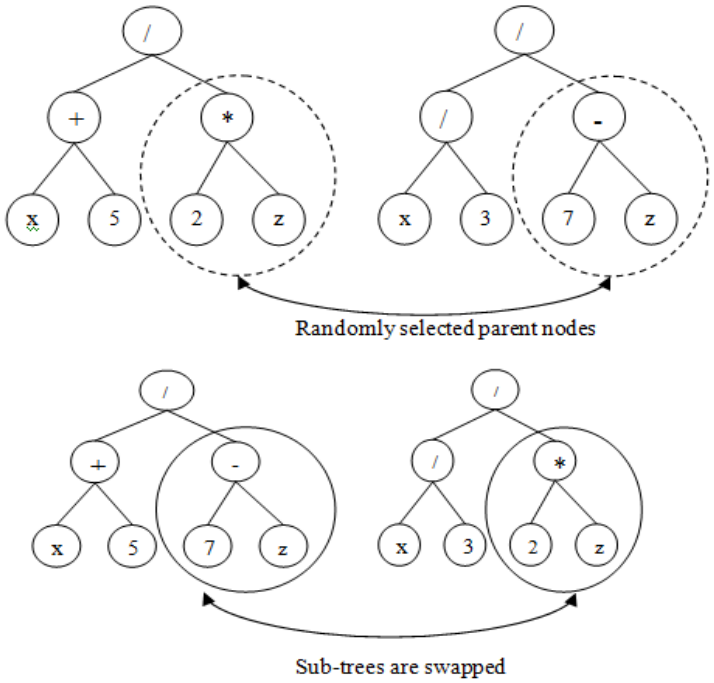

Figure 2. Example of GP crossover operator

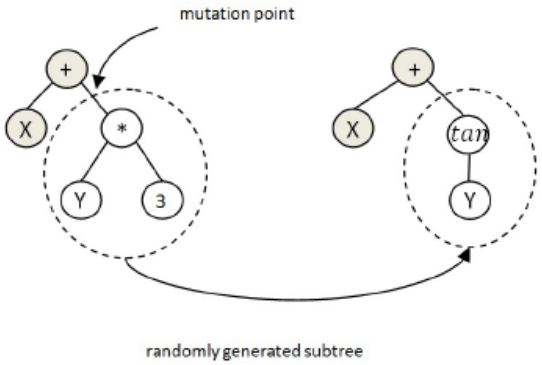

Figure 3. Example of GP mutation operator

F. *Elitism*

In this operation GP copies one or more individuals to the next evolutionary cycle without any change in their structure. Usually, individuals selection in this operation is based on their fitness value [25].

G. *Termination condition*

Finally, GP stops its evolutionary cycle when it finds an individual with a required fitness level or when the maximum number of generations is reached. Both termination conditions can be set before starting GP.

3. Model Evaluation

In order to assess the equality of the developed GP model, the following validation criteria are applied:

1) Variance-Accounted-For (VAF):

$$VAF = \left[1 - \frac{\mathrm{var}(\mathrm{y}-\hat{\mathrm{y}})}{var(\mathrm{y})}\right] \times 100\% \qquad (3)$$

2) Mean squared error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{y}_i - \hat{y}_i)^2 \qquad (4)$$

3) Root-mean-square error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{y}_i - \hat{y}_i)^2} \qquad (5)$$

Where y and $\hat{y}$ are the actual lipase activates and the estimated, respectively. *n* is the number of observations used in the experiments.

4. Materials and Methods

A. *Bacterial Strain*

For the purpose of applying GP, we use the same dataset developed in [17]. Author in [17] described the methodology of producing the bacterial strain in details. The bacterial strain was isolated from oily food waste in Serdang, Selangor, Malaysia and identified as

Acinetobacter sp. by the German Collection of Microorganisms and Cell Cultures (DSMZ), Braunschweig, Germany.

B. *Lipase Production in SmF*

In [17] author stated that:

"*The selected SmF lipase production medium was composed of (% w/v): peptone (5), yeast extract (1), NaCl (0.05), CaCl2 (0.05), lactose (1); and coconut oil (1% v/v). The medium was sterilized for 20min at 121C. The SmF cultures were performed in 250mL blue cap bottles in a rotary incubator shaker (0250rpm). The agitation, inoculum size, initial pH, temperature, and time were adjusted according to the central composite rotatable design (CCRD). After lipase production, the cell-free supernatant was obtained by centrifugation at 12,000 g, 4C for 10min prior to lipase assay."*

5. Data collection

Experimental data was collected in [17] for five variables in for the lipase production in SmF system. The variable selected levels were incubation temperature (27- 45 0C); initial pH (6-9); moisture content (60-100%); olive oil (0-20%) and incubation period (72-168h). The experimental produced lipase activity in SmF is presented in Table 1. The data set consists of 22 records. The first 19 record were used for training and the last 3 records for testing (in bold). In [17], authors used ANN to develop a predictive model to predict the lipase production in SmF.

Table 1. Actual and estimated lipase activities in SmF based for ANN, and GP models

Temp. C^0	pH	Inoculum (%)	Time (h)	Agitation (rpm)	Lipase Act. Measured	NN Model [17]	GP Model
40.9	8.3	1.9	79.8	45.1	8	7.9	7.9
40.9	6.7	4.1	79.8	45.1	3.6	3.7	3.6
31.1	8.3	4.1	40.2	154.9	10.6	7	10.0
40.9	8.3	4.1	40.2	45.1	5.5	7	5.2
40.9	8.3	1.9	40.2	154.9	4.8	4.1	5.3
40.9	6.7	1.9	79.8	154.9	2.4	1.1	3.2
31.1	6.7	4.1	79.8	154.9	7.2	7.5	7.1
31.1	8.3	1.9	79.8	154.9	8.5	6.2	8.0
40.9	6.7	4.1	40.2	154.9	5.6	6.3	5.2
31.1	8.3	4.1	79.8	45.1	5.2	5.6	5.1
31.1	6.7	1.9	40.2	45.1	8.3	8.6	8.6
27	7.5	3	60	100	11.2	10.2	11.6
45	7.5	3	60	100	9.4	10.4	9.2
36	6	3	60	100	9.8	9.9	9.1
36	9	3	60	100	15	13.4	15.4
36	7.5	1	60	100	9.2	9.5	9.4
36	7.5	5	60	100	7.6	3.4	8.1

36	7.5	3	24	100	9.2	11.9	8.8
36	7.5	3	96	100	8.4	8.3	8.8
36	**7.5**	**3**	**60**	**0**	**4.2**	**5.8**	**4.5**
36	**7.5**	**3**	**60**	**200**	**10.6**	**9.8**	**7.8**
36	**7.5**	**3**	**60**	**100**	**7.2**	**10**	**8.8**

6. Experiments and Results

A. *HueristicLab framework*

HueristicLab framework was used to develop a GP model for the experiments designed in this work. HeuristicLab is a flexible and extensible graphical user interface software for heuristic optimization based on Microsoft .Net and C#[1].

GP evolves a number of trees (models) automatically in cycles. HueristicLab is used to define the number of trees to be combined. As the number of tress increased the model complexity increased but a possible solution could be found. Training data are used to build the model while the testing data are used, after the run, to evaluate the developed GP models.

B. *GP Setup*

Before running the GP algorithm, there are number of parameters should be set by the user such as the population size, probability of crossover and mutation and the type of the selection and crossover mechanisms. There are also other parameters that control the complexity of the generated models such as the maximum tree length and the maximum tree depth. Setting small values for the latter two parameters can lead to producing simpler and compact models but less efficient ones in term of accuracy. Experience in the domain can help in setting all these parameters taking in consideration the simplicity versus efficiency issue. In our experiments, the values for all GP parameters are shown in Table 2.

C. *GP Model evaluation*

The prediction results of the best obtained GP model for the training and testing experiments are shown in Table 3. In Figure 4, we show the convergence of the GP evolutionary cycle. It can be noticed that the evolutionary cycle converged to the best model after around 330 cycles. Figure 5 shows the actual and estimated lipase activities based the developed GP model. In order to compare the results of the GP model with the ANN model proposed in [17], the VAF, MSE and RMSE were computed for both models, GP and ANN. The computed values are given in Table 3.

As shown in Table 3, for training data set, the VAF is 0.98% and the MSE is 0.17, whereas for the testing data set, VAF is 0.51% and MSE is 3.41. According to the evaluation results obtained, it can be noticed that the GP model outperformed the ANN model for training and testing sets.

[1]HeuristicLab is a framework for heuristic and evolutionary algorithms that is developed by members of the Heuristic and Evolutionary Algorithms Laboratory (HEAL).http://dev.heuristiclab.com

Table 2. GP TUNING PARAMETERS

Parameter	Value
Mutation probability	15%
Population size	1000
Maximum generations	1000
Maximum Tree Depth	12
Maximum Tree Length	20
Selection mechanism	Tournament selector
Elites	1
Operators	{+ ,- , *, /, sin, cos, tan}

Table 3. Evaluation results for the developed GP model compared to ANN

	GP model		ANN model [17]	
	training	testing	training	testing
VAF	0.98	0.51	0.71	0.40
MSE	0.17	3.41	2.80	3.68
RMSE	0.42	1.85	1.67	1.92

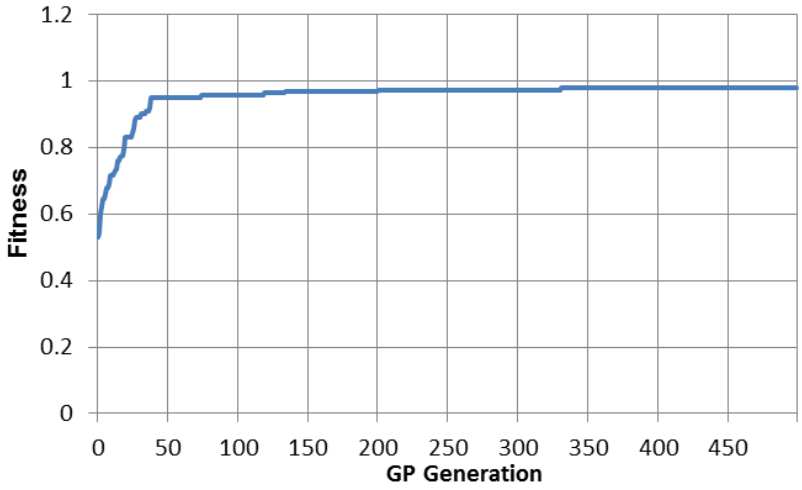

Figure 4. Convergence of GP ecolutionary cycle

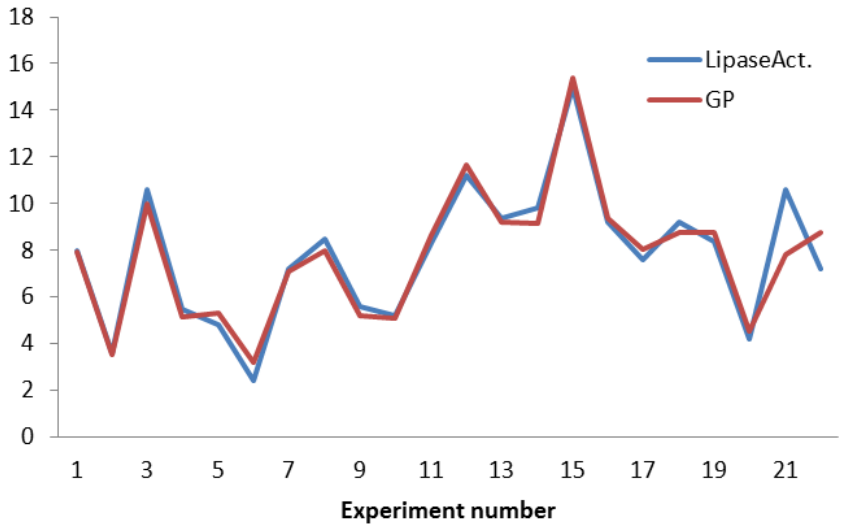

Figure 5. Actual and GP Model estimated Lipase Activities

7. Conclusion

In this paper, a genetic programming approach was used to develop a symbolic regression model to predict the lipase activity. GP model in this research was based on five independent variables showing observed values of lipase activity in SmF system. They are the Temperature, pH, Inoculum, Time and Agitation. Performance of the developed model was evaluated and compared based on VAF, MSE and RMSE criteria. GP showed promising results compared to the artificial neural network based model proposed previously in the literature.

References

[1] T. Ghose and V. Bisaria, "Development of biotechnology in india", in History of Modern Biotechnology I (A. Fiechter, ed.), Advances in Biochemical Engineering/Biotechnology, vol. 69, **(2000)**, pp. 87–124.
[2] Y. Abdel-Fattah, "Optimization of thermostable lipase production from a thermophilicgeobacillus sp. using box-behnken experimental design", Biotechnology Letters, vol. 24, **(2002)**, pp. 1217–1222.
[3] W. Sheta, W. M. Aly, Y. R. Abdel-Fattah and A. F. Sheta, "Designing a biological experiment using genetic programming: An evolutionary methodology to improve the production of thermostable lipase enzyme", WSEAS Transactions on Systems, vol. 4, no. 2, **(2003)**, pp. 1221– 1232.
[4] A. Ebrahimpour, R. Rahman, D. EanCh´ang, M. Basri and A. Salleh, "A modeling study by response surface methodology and artificial neural network on culture parameters optimization for thermostable lipase production from a newly isolated thermophilic geobacillus sp. strain ARM", BMC Biotechnology, vol. 8, no. 1, **(2008)**, pp. 96.
[5] B. Zamost, H. Nielsen and R. Starnes, "Thermostable enzymes for industrial applications", Journal of Industrial Microbiology, vol. 8, **(1991)**, pp. 71–81.
[6] O. Olsen, R. Borriss, O. Simon and K. Thomsen, "Hybrid bacillus (1- 3,1-4)-?-glucanases: engineering thermostable enzymes by construction of hybrid genes", Molecular and General Genetics MGG, vol. 225, **(1991)**, pp. 177–185.
[7] P. Turner, G. Mamo and E. Karlsson, "Potential and utilization of thermophiles and thermostable enzymes in biorefining", Microbial Cell Factories, vol. 6, **(2007)**, pp. 1–9.
[8] D. -H. Chung, J. Huddleston, J. Farkas and J. Westpheling, "Identification and characterization of CbeI: a novel thermostable restriction enzyme from caldicellulosiruptorbescii DSM 6725 and a member of a new subfamily of HaeIII-like enzymes", Journal of Industrial Microbiology and Biotechnology, vol. 38, **(2011)**, pp. 1867–1877.

[9] C. Schmidt-Dannert, M. L. Rua, H. Atomi and R. D. Schmid, "Thermoalkalophilic lipase of bacillus thermocatenulatus. i. molecular cloning, nucleotide sequence, purification and some properties", Biochimica et BiophysicaActa, vol. 1301, no. 1-2, **(1996)**, pp. 105114.
[10] C. X. Cceres, D. M. Freire, R. E. Cceres and R. F. Segovia, "Mathematical modeling of lipase production by penicilliumrestrictum in batch fermentation", in Proc. of 4th Mercosur Congress on Process Systems Engineering, **(2005)**.
[11] A. N. Emmanuel, A. David, Y. K. Benjamin and E. T. Yannick, "A hybrid neural network approach for batch fermentation simulation", Australian Journal of basic and applied sciences, vol. 3, no. 4, **(2009)**, pp. 3930–3936.
[12] R. Hiary, A. Sheta and H. Faris, "Fermentation process modeling using takagi-sugeno fuzzy model", WSEAS Trans. on Systems, vol. 11, no. 8, **(2012)**, pp. 375–384.
[13] A. Sheta and R. Hiary, "Modeling lipase production process using artificial neural networks", in Proceedings of the 3rd IEEE International Conference on Multimedia Computing and Systems, (Tangier, Morocco), **(2012)** May 10-12, pp. 1158–1163.
[14] D. P. Searson, D. E. Leahy and M. J. Willis, "GPTIPS: An open source genetic programming toolbox for multigene symbolic regression", in Proceedings of the International Multi-conference of Engineers and Computer Scientists 2010 (IMECS 2010), vol. 1, (Hong Kong), **(2010)** March 17-19, pp. 77–80.
[15] J. Koza, "Evolving a computer program to generate random numbers using the genetic programming paradigm", in Proceedings of the Fourth International Conference on Genetic Algorithms, Morgan Kaufmann, La Jolla, CA, **(1991)**.
[16] J. R. Koza, "Genetic Programming: On the Programming of Computers by Means of Natural Selection", The MIT Press, **(1992)**.
[17] A. Khoramnia, A. Ebrahimpour, B. K. Beh and O. M. Lai, "Production of a solvent, detergent, and thermotolerant lipase by a newly isolated acinetobacter sp. in submerged and solid-state fermentations", Journal of Biomedicine and Biotechnology, vol. 2011, no. 702179, **(2011)**.
[18] O. Roeva, "A modified genetic algorithm for a parameter identification of fermentation processes", Biotechnology and Biotechnological Equipment, vol. 20, no. 1, **(2006)**, pp. 202–209.
[19] H. Iba, T. Hitoshi, G. Hung and S. Taisuke, "System identification using structured genetic algorithms", in Proceedings of the Fifth International Conference on Genetic Algorithms, Morgan Kaufmann, **(1993)**, pp. 279–286.
[20] A. Sheta, R. Hiary, H. Faris and N. Ghatasheh, "Optimizing Thermostable Enzymes Production Using Multigene Symbolic Regression Genetic Programming", World Applied Sciences Journal, vol. 22, no. 4, **(2013)**, pp. 485–493.
[21] S. Chen, S. A. Billings and P. M. Grant, "Non-linear system iden- tification using neural networks", International J. Control, vol. 51, **(1990)**, pp. 1191–1214.
[22] M. Hinchliffe, H. Hiden, B. McKay, M. Willis, M. Tham and G. Barton, "Modelling chemical process systems using a multi-gene genetic programming algorithm", in Late Breaking Papers at the Genetic Programming 1996 Conference Stanford University July 28- 31, 1996 (J. R. Koza, ed.), (Stanford University, CA, USA), Stanford Bookstore, **(1996)** July 28-31, pp. 56–65.
[23] B. L. Miller, B. L. Miller, D. E. Goldberg and D. E. Goldberg, "Genetic algorithms, tournament selection, and the effects of noise", Complex Systems, vol. 9, **(1995)**, pp. 193–212.
[24] R. E. Keller and W. Banzhaf, "Genetic programming using mutation, reproduction and genotype-phenotype mapping from linear binary genomes into linear lalr(1) phenotypes", in in Proceedings of Genetic Programming 1996 Conference, MIT Press, **(1996)**, pp. 116–122.
[25] S. Yang, "Genetic algorithms with memory and elitism based immigrants in dynamic environments", Evolutionary Computation, **(2008)**, pp. 385– 416.
[26] J. Duffy and J. Engle-Warnick, "Using Symbolic Regression to Infer Strategies from Experimental Data", Evolutionary Computation in Economics and Finance, volume 100 of Studies in Fuzziness and Soft Computing, chapter 4, Physica Verlag, **(1999)**.

Authors

Hossam Faris is an Assistant professor at Business Information Technology department/King Abdulla II, School for Information Technology/ University of Jordan. H. Faris received his BA, M.Sc. degrees (with excellent rates) in Computer Science from Yarmouk University and Al-Balqa` Applied University in 2004 and 2008 respectively in Jordan. Since then, he has been awarded a full scholarship to peruse his PhD degrees in e-

Business at University of Salento, Italy, where he obtained his PhD degree in 2011. His research interests include Knowledge Management Systems, Ontologies, e-Business, Search and retrieval algorithms and Genetic Algorithms

Alaa Sheta received his B.E., M.Sc. degrees in Electronics and Communication Engineering from Faculty of Engineering, Cairo University in 1988 and 1994, respectively. A. Sheta received his Ph.D. degree from the Computer Science Department, School of Information Technology, George Mason University, Fairfax, VA, USA in 1997. Prof. Sheta served as the Vice Dean of Prince Abullah Bin Ghazi Faculty of Science and Information Technology, Al-Balqa Applied University, Salt, Jordan on the academic year 2008/2009. He has been the Dean Assistant for Planning and Development during the years 2006-2008. Currently, Prof. Sheta is chairman of the Computers and Systems Department, Electronics Research Institute (ERI), Cairo, Egypt. His research interests include Modeling and Simulation of Dynamical Nonlinear Systems, Robotics, Evolutionary Computation, Automatic Control, Fuzzy Logic, and Neural Networks.

Rania Al-Hiary received her B.Sc in Biology from the Faculty of Science, Jordanian University and her M.Sc. in Information and Communication Technology in Education, Prince Hussein bin Abdullah faculty of Information Technology, Albayt University, Al-Mafrag, Jordan. In her M.Sc. thesis project, she explored the effect of using robotics laboratories for helping Tenth grade students to gain creative thinking skills. She published number of research articles in international conferences and journals. She has been a teacher at number of high schools in Jordan. She was also part of the Red Cross/Red Crescent society in Amman, Jordan. Rania's scientific research interest includes bioinformatics, lipase activity modeling; computer simulation; innovation in education and education leadership.

