Proceedings of the 2013 Federated Conference on
Computer Science and Information Systems pp. 371–376

Influence of the Population Size on the Genetic Algorithm Performance in Case of Cultivation Process Modelling

Olympia Roeva
Institute of Biophysics and
Biomedical Engineering,
Bulgarian Academy of Science,
Acad. G. Bonchev Str., bl. 105,
1113 Sofia, Bulgaria
E-mail: olympia@biomed.bas.bg

Stefka Fidanova
Institute of Information and
Communication Technology,
Bulgarian Academy of Science,
Acad. G. Bonchev Str., bl. 25A,
1113 Sofia, Bulgaria
E-mail: stefka@parallel.bas.bg

Marcin Paprzycki
Systems Research Institute,
Polish Academy of Sciences, Warsaw
and
Management Academy,
Warsaw, Poland
E-mail:marcin.paprzycki@ibspan.waw.pl

***Abstract*—In this paper, an investigation of the influence of the population size on the genetic algorithm (GA) performance for a model parameter identification problem, is considered. The mathematical model of an *E. coli* fed-batch cultivation process is studied. The three model parameters – maximum specific growth rate (μ_{max}), saturation constant (k_S) and yield coefficient ($Y_{S/X}$) are estimated using different population sizes. Population sizes between 5 and 200 chromosomes in the population are tested with constant number of generations. In order to obtain meaningful information about the influence of the population size a considerable number of independent runs of the GA are performed. The observed results show that the optimal population size is 100 chromosomes for 200 generations. In this case accurate model parameters values are obtained in reasonable computational time. Further increase of the population size, above 100 chromosomes, does not improve the solution accuracy. Moreover, the computational time is increased significantly.**

I. Introduction

METAHEURISTICS, such as genetic algorithms (GA), are widely used to solve various optimization problems. The GA are highly relevant for industrial applications, because they are capable of handling problems with non-linear constraints, multiple objectives, and dynamic components – properties that frequently appear in the real-world problems [15]. Since their introduction and subsequent popularization [16], the GA have been frequently used as an alternative optimization tool to the conventional methods and have been successfully applied in a variety of areas, and still find increasing acceptance [1], [3], [7], [11], [23], [28], [29].

The metaheuristic algorithms require of setting the values of several algorithm components and parameters. These parameters values have great impact on performance and efficacy of the algorithm [13], [22], [30], [14]. Therefore, it is important to investigate the algorithm parameters influence on the performance of the developed metaheuristic algorithms. The aim is to find the optimal parameters values for the considered optimization problem. The optimal values for the parameters depend mainly on i) the problem; ii) the instance of the problem to deal with and iii) the computational time that will be spend in solving the problem. Usually in the algorithm parameters tuning a compromise between solution quality and search time should be done.

For the parameter setting of metaheuristics, several automated approaches exist. These methods use i) a single step of parameter tuning (prior to the practical use of the algorithm), or parameter control (self adaptation to the problem being optimized) [19]. Parameter control is well suited when one wants good average performances across diverse problems, but the needed computation overhead leads to less efficiency on specific problems, compared to parameter tuning [9]. Best known parameter tuning techniques are racing [8], sequential parameter optimization [5] and meta-parameter setting (sometimes referred as meta-algorithm [5]).

Population sizing has been one of the important topics to consider in evolutionary computation [2], [12], [31]. Various results about the appropriate population size can be found in the literature[25], [27]. Researchers usually argue that a "small" population size could guide the algorithm to poor solutions [17], [24], [31] and that a "large" population size could make the algorithm expend more computation time in finding a solution [17], [20], [21]. Due to significant influence of population size to the solution quality and search time [27] a more thorough research should be done for this GA parameter.

The main goal of this research is to carry out investigation of the influence of one of the key GA parameters – population size (number of chromosomes) – on the algorithm performance for identification of a cultivation process model. Parameter identification of non-linear cultivation process models is a hard combinatorial optimization problem for which exact algorithms or traditional numerical methods do not work efficiently. A non-linear mathematical model of fed-batch cultivation process of the most important host organism for recombinant protein production — bacteria *Escherichia coli* – is considered [27].

The paper is organized as follows. The problem formulation is given in Section 2. The numerical results and a discussion

978-1-4673-4471-5/$25.00

are presented in Section 3. Conclusion remarks are done in Section 4.

II. Problem Formulation

A. E. coli Fed-batch Cultivation Model

Application of the general state space dynamical model [6] to the *E. coli* cultivation fed-batch process leads to the following nonlinear differential equation system [27]:

$$\frac{dX}{dt} = \mu_{max}\frac{S}{k_S + S}X - \frac{F_{in}}{V}X \tag{1}$$

$$\frac{dS}{dt} = -\frac{1}{Y_{S/X}}\mu_{max}\frac{S}{k_S + S}X + \frac{F_{in}}{V}(S_{in} - S) \tag{2}$$

$$\frac{dV}{dt} = F_{in} \tag{3}$$

where X is the biomass concentration, [g/l]; S is the substrate concentration, [g/l]; F_{in} is the feeding rate, [l/h]; V is the bioreactor volume, [l]; S_{in} is the substrate concentration in the feeding solution, [g/l]; μ_{max} is the maximum value of the specific growth rate, [h^{-1}]; k_S is the saturation constant, [g/l]; $Y_{S/X}$ is the yield coefficient, [-].

The initial process conditions are [4]:

- t_0 = 6.68 h,
- $X(t_0)$ = 1.25 g/l and $S(t_0)$ = 0.8 g/l,
- S_{in} = 100 g/l.

For the considered non-linear mathematical model of *E. coli* fed-batch cultivation process the parameters that should be identified are:

- maximum specific growth rate (μ_{max}),
- saturation constant (k_S),
- yield coefficient ($Y_{S/X}$).

B. Genetic Algorithm

GA was developed to model adaptation processes mainly operating on binary strings and using a recombination operator with mutation as a background operator. The GA maintains a population of chromosomes, $P(t) = x_1^t$, ..., x_n^t for generation t. Each chromosome represents a potential solution to the problem and is implemented as some data structure S. Each solution is evaluated to give some measure of its "fitness". Fitness of a chromosome is assigned proportionally to the value of the objective function of the chromosomes. Then, a new population (generation $t+1$) is formed by selecting more fit chromosomes (selection step). Some members of the new population undergo transformations by means of "genetic" operators to form new solution. There are unary transformations m_i (mutation type), which create new chromosomes by a small change in a single chromosome ($m_i : S \to S$), and higher order transformations c_j (crossover type), which create new chromosomes by combining parts from several chromosomes ($c_j : S \times \ldots \times S \to S$). After some number of generations the algorithm converges – it is expected that the best chromosome represents a near-optimum (reasonable) solution. The combined effect of selection, crossover and mutation gives so-called reproductive scheme growth equation [15]:

$$\xi(S, t+1) \geq$$

$$\xi(S,t) \cdot eval(S,t) / \bar{F}(t) \left[1 - p_c \cdot \frac{\delta(S)}{m-1} - o(S) \cdot p_m\right]$$

The structure of the herewith used GA is shown by the pseudocode below (Figure 1).

begin
  $i = 0$
  Initial population $P(0)$
  Evaluate $P(0)$
  while (not done) **do**
  (test for termination criterion)
  begin
    $i = i + 1$
    Select $P(i)$ from $P(i-1)$
    Recombine $P(i)$
    Mutate $P(i)$
    Evaluate $P(i)$
  end
end

Fig. 1. Pseudocode for GA

Three model parameters are represented in the chromosome – μ_{max}, k_S and $Y_{S/X}$. The following upper and lower bounds of the model parameters are considered [29]:

$$0 < \mu_{max} < 0.7,$$

$$0 < k_S < 1,$$

$$0 < Y_{S/X} < 30.$$

Roulette wheel, developed by Holland [16] is the herewith used selection method. The probability, P_i, for each chromosome is defined by:

$$P[\text{Individual } i \text{ is chosen}] = \frac{F_i}{\sum_{j=1}^{PopSize} F_j}, \tag{4}$$

where F_i equals the fitness of chromosome i and $PopSize$ is the population size.

To reproduce the chromosomes simple crossover and binary mutation according to [29] are applied. In proposed genetic algorithm fitness-based reinsertion (selection of offspring) is used.

For the considered here model parameter identification, the type of the basic operators in GA are as follows [29]:

- encoding – binary,
- fitness function – linear ranking,
- selection function – roulette wheel selection,
- crossover function – simple crossover,
- mutation function – binary mutation,
- reinsertion – fitness-based.

The values of GA parameters are [29]:

- generation gap, ggap = 0.97,
- crossover probability, xovr = 0.75,
- mutation probability, mutr = 0.01,
- maximum number of generations, maxgen = 200.

C. Optimization Criterion

In practical view, modelling studies are performed to identify simple and easy-to-use models that are suitable to support the engineering tasks of process optimization and, especially of control. The most appropriate model must satisfy the following conditions:

(i) the model structure should be able to represent the measured data in a proper manner;

(ii) the model structure should be as simple as possible compatible with the first requirement.

The optimization criterion is a certain factor, whose value defines the quality of an estimated set of parameters. To evaluate the mishmash between experimental and model predicted data the Least Square Regression is used.

The objective consists of adjusting the parameters (μ_{max}, k_S and $Y_{S/X}$) of the non-linear mathematical model function (Eq. (1) - Eq. (3)) to best fit a data set. A simple data set consists of n points (data pairs) (x_i, y_i), $i = 1, 2, \ldots, n$, where x_i is an independent variable and y_i is a dependent variable whose value is found by observation. The model function has the form $f(x, \beta)$, where the m adjustable parameters are held in the vector β, $\beta = [\mu_{max}\ k_S\ Y_{S/X}]$. The goal is to find the parameter values for the model which "best" fits the data. The least squares method finds its optimum when the sum S of squared residuals:

$$S = \sum_{i=1}^{n} r_i^2$$

is a minimum. A residual is defined as the difference between the actual value of the dependent variable and the value predicted by the model. A data point may consist of more than one independent variable. For an example, when fitting a plane to a set of height measurements, the plane is a function of two independent variables, x and z, say. In the most general case there may be one or more independent variables and one or more dependent variables at each data point.

$$r_i = y_i - f(x_i, \beta).$$

III. Numerical Results and Discussion

All computations are performed using a PC/Intel Core i5-2320 CPU @ 3.00GHz, 8 GB Memory (RAM), Windows 7 (64 bit) operating system and Matlab 7.5 environment.

A series of numerical experiments are performed to evaluate the influence of the population size in GAs on the accuracy of the obtained solution. Using mathematical model of the *E. coli* cultivation process (Eq. (1) - Eq. (3)) the model parameters – maximum specific growth rate (μ_{max}), saturation constant (k_S) and yield coefficient ($Y_{S/X}$) – are estimated. For the identification procedures consistently different population sizes (from 5 to 200 chromosomes in the population) are used. The number of generations is fixed to 200. Because of the stochastic characteristics of the applied GA series of 30 runs for each population size are performed.

TABLE I
ALGORITHM PERFORMANCE FOR VARIOUS POPULATION SIZES - OBJECTIVE FUNCTION VALUES

Population size	Objective function S		
	Average	Best	Worst
5	6.1200	4.8325	9.2958
10	5.8000	4.8548	9.6175
20	4.7660	4.4753	5.3634
30	4.6519	4.4816	5.0094
40	4.6359	4.4437	4.9669
50	4.6070	4.4488	4.8636
60	4.5886	4.4625	4.8013
70	4.5648	4.4384	4.7357
80	4.5782	4.4474	4.7463
90	4.5711	4.4496	4.7211
100	4.5406	4.4252	4.7017
110	4.5455	4.4332	4.7319
150	4.5511	4.4575	4.6717
200	4.5453	4.4359	4.7206

TABLE II
ALGORITHM PERFORMANCE FOR VARIOUS POPULATION SIZES - COMPUTATIONAL TIME

Population size	Computational time, s		
	Average	Best	Worst
5	4.9457	4.5552	5.6004
10	6.0039	5.6316	6.3648
20	7.6482	7.3008	7.9561
30	11.1115	10.8265	11.5129
40	12.9824	12.4957	13.3537
50	14.9087	14.3989	15.5377
60	17.2766	16.6141	20.3113
70	19.7601	19.1725	20.0617
80	22.1880	21.7153	22.6669
90	24.3414	23.9150	24.8198
100	26.8644	26.4890	27.8306
110	29.7057	29.1878	30.2642
150	39.7273	39.1407	40.3887
200	52.4782	51.3087	55.8952

In the Table I, obtained average, best and worst objective function values for considered population sizes, are presented. The results observed for computational time are listed in Table II.

The numerical experiments show that increasing the size of the population of 5 to 100 chromosomes significantly improves the resulting value of the objective function (average results) – from 6.1200 to 4.5406 (see Table I). The further increase in the size of population (more than 100 chromosomes) does not lead to more accurate results. The subsequent increase in the population size leads only to an increase in computational time without improving the value of the objective function (average results) – from 26.8644 s (100 chromosomes) to 52.4782 s (200 chromosomes) vs. $S = 4.5406$ to $S = 4.5453$ (see Table II).

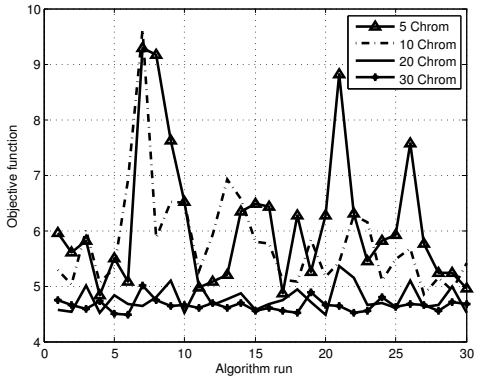

Fig. 2. Objective function values obtained during the 30 algorithm runs for 5, 10, 20 and 30 chromosomes in the population

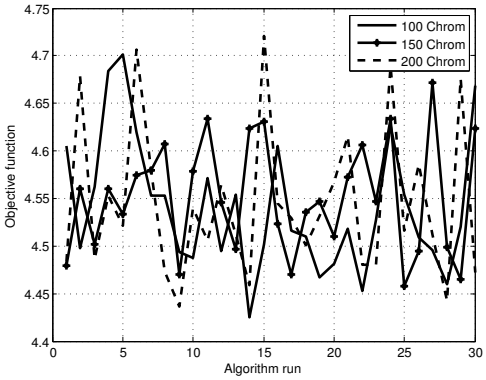

Fig. 3. Objective function values obtained during the 30 algorithm runs for 100, 150 and 200 chromosomes in the population

For better interpretation the obtained numerical results are graphically visualized in the next figures. On Figure 2 the objective function values, obtained during the 30 GA runs for 5, 10, 20 and 30 chromosomes in the population, are shown. The graphical results show that the GA could not find accurate solution using small population size – 5 or 10 chromosomes. It is need at least 20 chromosomes in population for achieving a better solution. On Figure 3 the objective function values, obtained during the 30 algorithm runs for 100, 110, 150 and 200 chromosomes in the population, are shown. Here, it could be seen that using large population size (110, 150 or 200 chromosomes) did not result in an improvement of the objective function values. The ANOVA test is applied and the values of the objective function for population size equal and more than 100 are statistically equal. Moreover, as can be seen from Figure 5 increasing the population size result in an acceleration of computational time. When the population size increases it leads to increase of the needed computational resources like time and memory which can be a problem for large-scale tests. Therefore we can conclude that populations with 100 individuals is optimal with respect to the value of the objective function and the needed computational resources.

All numerical experiments for the influence of the population size on the objective function value and on the computational time are summarized in Figure 4 and Figure 5. It can be concluded that for the considered here non-linear cultivation model parameter identification problem the optimal population size is 100 chromosomes in the population (for 200 generations).

In the Table III the best parameter values (μ_{max}, k_S and $Y_{S/X}$), obtained using GA with 100 chromosomes in the population, are presented. According to [10], [18], [32] the values of the estimated model parameters are in admissible boundaries.

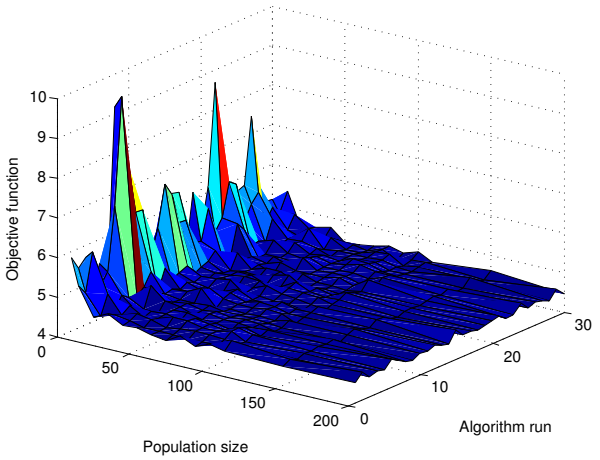

Fig. 4. Influence of the population size on the objective function value

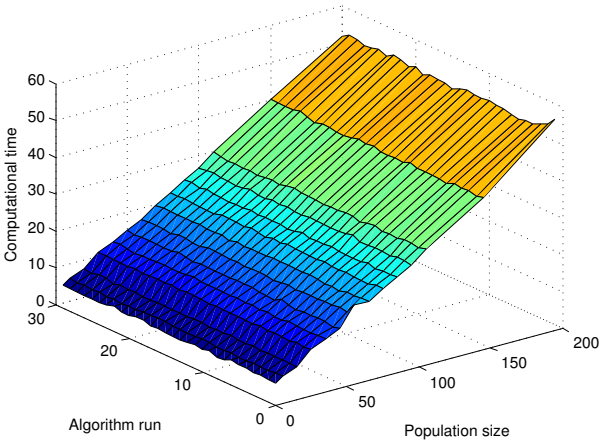

Fig. 5. Influence of the population size on the computational time

IV. Conclusion

A good selection of the GA parameters improve both computation time and solution accuracy. Finding good parameter values is not a trivial task and requires human expertise as

TABLE III
BEST PARAMETER VALUES OF THE MODEL (100 CHROMOSOMES)

Parameter	Value
μ_{max}, [1/h]	0.4881
k_S, [g/l]	0.0120
$Y_{S/X}$, [-]	2.0193

well as time. In this paper, the influence of the one of key GA parameters (population size) on the GA performance, is studied. As a test problem, the *E. coli* fed-batch cultivation model parameter identification, is considered. The three model parameters (maximum specific growth rate (μ_{max}), saturation constant (k_S) and yield coefficient ($Y_{S/X}$)) are identified. For a fixed number of the generations (200) different population sizes of the GA are explored. The numerical experiments are started with 5 chromosomes in the population and consistently increased to 200 chromosomes. The obtained results show that the optimal population size, for the considered here case study, is 100 chromosomes. Thus, accurate model parameters values are obtained with reasonable computational efforts. The use of smaller populations result in lower accuracy of the solution, obtained for a smaller computational time. The further increase of the population size increases the accuracy of solution. This effect is observed to a population size of 100 chromosomes. The use of larger populations does not improve the solution accuracy and only increase the needed computational resources.

ACKNOWLEDGMENT

This work has been partially supported by the Bulgarian National Scientific Fund under the Grants DID 02/29 "Modeling Processes with Fixed Development Rules (ModProFix)" and DMU 02/4 "High quality control of biotechnological processes with application of modified conventional and metaheuristics methods". Work presented here is a part of the Poland-Bulgarian collaborative Grant "Parallel and distributed computing practices" and by European Commission project ACOMIN.

REFERENCES

[1] S. Akpinar and G. M. Bayhan, "A Hybrid Genetic Aalgorithm for Mixed Model Assembly Line Balancing Problem with Parallel Workstations and Zoning Constraints", Engineering Applications of Artificial Intelligence, Vol. 24, No. 3, 2011, pp. 449-457.

[2] J. T. Alander, "On optimal population size of genetic algorithms", In Proceedings of the IEEE Computer Systems and Software Engineering, 1992, pp. 65–69.

[3] H. N. Al-Duwaish, "A Genetic Approach to the Identification of Linear Dynamical Systems with Static Nonlinearities", International Journal of Systems Science, Vol. 31, No. 3, 2000, pp. 307-313.

[4] M. Arndt and B. Hitzmann, "Feed Forward/feedback Control of Glucose Concentration during Cultivation of *Escherichia coli*", 8th IFAC Int. Conf. on Comp. Appl. in Biotechn, Canada, 2001, pp. 425-429.

[5] T. Bartz-Beielstein, "Experimental Research in Evolutionary Computation: The New Experimentalism", Natural Computing Series, Springer, 2006.

[6] G. Bastin and D. Dochain, "On-line Estimation and Adaptive Control of Bioreactors", Els. Sc. Publ, 1991.

[7] K. K. Benjamin, A. N. Ammanuel, A. David and Y. K. Benjamin, "Genetic Algorithm using for a Batch Fermentation Process Identification", J of Applied Sciences, Vol. 8, No. 12, 2008, pp. 2272-2278.

[8] M. Birattari, T. Stützle, L. Paquete and K. Varrentrapp, "A racing algorithm for configuring metaheuristics", In GECCO 02: Proceedings of the Genetic and Evolutionary Computation Conference, 2002, pp. 11–18, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc.

[9] J. Clune, S. Goings, B. Punch and E. Goodman, "Investigations in metagas: panaceas or pipe dreams?", In GECCO 05: Proceedings of the 2005 workshops on Genetic and evolutionary computation, 2005, pp. 235–241, New York, NY, USA, ACM.

[10] J. Contiero, C. Beatty, S. Kumari, C. L. DeSanti, W. L. Strohl, A. Wolfe, "Effects of mutations in acetate metabolism on high-cell-density growth of *Escherichia coli*", Journal of Industrial Microbiology and Biotechnology, Vol. 24, 2000, pp. 421-430.

[11] M. F. J. da Silva, J. M. S. Perez, J. A. G. Pulido and M. A. V. Rodriguez, "AlineaGA - A Genetic Algorithm with Local Search Optimization for Multiple Sequence Alignment", Appl Intell, Vol. 32, 2010, pp. 164-172.

[12] P. A. Diaz-Gomez and D. F. Hougen, "Initial Population for Genetic Algorithms: A Metric Approacs", Proceedings of the 2007 International Conference on Genetic and Evolutionary Methods, GEM 2007, June 25-28, 2007, Las Vegas, Nevada, USA, Hamid R. Arabnia and Jack Y. Yang and Mary Qu Yang (Eds), pp. 43-49, CSREA Press.

[13] Á. E. Eiben, R. Hinterding and Z. Michalewicz, "Parameter Control in Evolutionary Algorithms", IEEE Transactions on Evolutionary Computation, Vol. 3, No. 2, 1999.

[14] Fidanova S., "Simulated Annealing: A Monte Carlo Method for GPS Surveying", Computational Science - 2006, Lecture Notes in Computer Science No 3991, 2006, pp. 1009-1012.

[15] D. E. Goldberg, "Genetic Algorithms in Search, Optimization and Machine Learning", Addison Wesley Longman, London, 2006.

[16] J. H. Holland, "Adaptation in Natural and Artificial Systems", 2nd Edn. Cambridge, MIT Press, 1992.

[17] V. K. Koumousis and C. P. Katsaras, A sawtooth genetic algorithm combining the effects of variable population size and reinitialization to enhance performance, IEEE Transactions on Evolutionary Computation, Vol. 10, No. 1, 2006, pp. 19–28.

[18] D. Levisauskas, V. Galvanauskas, S. Henrich, K. Wilhelm, N. Volk, A. Lubbert, "Model-based optimization of viral capsid protein production in fed-batch culture of recombinant *Escherichia coli*", Bioprocess and Biosystems Engineering, Vol. 25, 2003, pp. 255-262.

[19] F. G. Lobo, C. F. Lima and Z. Michalewicz, (Ed.), "Parameter Setting in Evolutionary Algorithms", Studies in Computational Intelligence, Vol. 54, 2007.

[20] F. G. Lobo and D. E. Goldberg, "The parameterless genetic algorithm in practice, Information SciencesInformatics and Computer Science, Vol. 167, No. 1-4, 2004, pp. 217–232.

[21] F. G. Lobo and C. F. Lima, A review of adaptive population sizing schemes in genetic algorithms, In Proceedings of the Genetic and Evolutionary Computation Conference, 2005, pp. 228–234.

[22] R. Nowotniak and J. Kucharski, "GPU-based Tuning of Quantum-Inspired Genetic Algorithm for a Combinatorial Optimization Problem", In Proceedings of the XIV International Conference System Modeling and Control, 2011, ISSN 978-83-927875-1-8.

[23] J. P. Paplinski, "The Genetic Algorithm with Simplex Crossover for Identification of Time Delays", *Intelligent Information Systems*, 2010, pp. 337-346.

[24] M. Pelikan, D. E. Goldberg, and E. Cantu-Paz, "Bayesian optimization algorithm, population sizing, and time to convergence", Illinois Genetic Algorithms Laboratory, University of Illinois, Tech. Rep., 2000.

[25] C. R. Reeves, "Using Genetic Algorithms With Small Populations", In Proceedings of the Fifth International Conference on Genetic Algorithms, 1993, pp. 92-99.

[26] E. Ridge, "Design of Experiments for the Tuning of Optimisation Algorithms", PhD Thesis, The University of York, Department of Computer Science, 2007.

[27] O. Roeva, "Improvement of Genetic Algorithm Performance for Identification of Cultivation Process Models", Advanced Topics on Evolutionary Computing, Book Series: Artificial Intelligence Series-WSEAS, 2008, pp. 34-39.

[28] O. Roeva and Ts. Slavov, "Fed-batch Cultivation Control based on Genetic Algorithm PID Controller Tuning", Lecture Notes on Computer Science, Springer-Verlag Berlin Heidelberg, Vol. 6046, 2011, pp. 289-296.

[29] O. Roeva and S. Fidanova, "Chapter 13. A Comparison of Genetic Algorithms and Ant Colony Optimization for Modeling of *E. coli* Cultivation Process", In book Real-World Application of Genetic Algorithms, In Tech, 2012, pp. 261-282.

[30] A. Saremi, T. Y. ElMekkawy and G. G. Wang, "Tuning the Parameters of a Memetic Algorithm to Solve Vehicle Routing Problem with Backhauls Using Design of Experiments", International Journal of Operations Research, Vol. 4, No. 4, 2007, pp. 206-219.

[31] A. Piszcz and T. Soule, "Genetic programming: Optimal population sizes for varying complexity problems", In Proceedings of the Genetic and Evolutionary Computation Conference, 2006, pp. 953–954.

[32] B. Zelic, D. Vasic-Racki, C. Wandrey, R. Takors, "Modeling of the pyruvate production with *Escherichia coli* in a fed-batch bioreactor", Bioprocess and Biosystems Engineering, Vol. 26, 2004, pp. 249-258.