A Hybrid Genetic Algorithm for Parameter Identification of Bioprocess Models

Olympia Roeva

Institute of Biophysics and Biomedical Engineering, BAS
105 Acad. G. Bonchev Str., 1113 Sofia, Bulgaria,
olympia@clbme.bas.bg

Abstract. In this paper a hybrid scheme using GA and SQP method is introduced. In the hybrid GA-SQP the role of the GA is to explore the search place in order to either isolate the most promising region of the search space. The role of the SQP is to exploit the information gathered by the GA. To demonstrate the usefulness of the presented approach, two cases for parameter identification of different complexity are considered. The hybrid scheme is applied for modeling of *E. coli MC4110* fed-batch cultivation process. The results show that the GA-SQP takes the advantages of both GA's global search ability and SQP's local search ability, hence enhances the overall search ability and computational efficiency.

1 Introduction

Robust and efficient methods for parameter identification are of key importance in system biology and related areas. Nowadays the most common direct methods used for global optimization are evolutionary algorithms such as genetic algorithms (GA). The principal advantages of GA are domain independence, nonlinearity and robustness. The GA effectiveness has been already demonstrated for identification of fed-batch cultivation processes [2, 11]. The same qualities that make the GA so robust also can make it more computationally intensive and slower than other methods [8]. On the contrary, local search methods have faster convergence due to the use of local information for determination of the most promising search direction by creating logical movements. One of the leading methods for solving constrained non-linear optimization problems is sequential quadratic programming (SQP) [3, 5]. Algorithms in this class guarantee global convergence and typically require few iterations to locate a solution point. However, local search methods can easily be entrapped in local minima. An approach that overcomes the above disadvantages is to combine GA with local search methods, to design more efficient methods with relatively faster convergence than the pure GA. Hybrid GA have received significant interest in recent years and are being increasingly used to solve real-world problems [1]. Different local search methods have got attention in such combinations [7, 9, 13].

In this paper a parameter identification of an *E. coli MC4110* fed-batch fermentation process using hybrid GA is proposed. To improve the performance of the conventional GA, a combine scheme using the GA and SQP method is introduced. Thus, optimizers work jointly to locate efficiently quality design points better than either could alone.

This paper is organized as follows. Outline of the introduced hybrid algorithm is described in Section 2. In Section 3 a discussion of the obtained numerical results of *E. coli* cultivation process model parameter identification is presented. Conclusion remarks are done in Section 4.

2 Outline of the hybrid GA-SQP

GA is very effective at finding optimal solutions to a variety of complex optimization problems because it does not impose many of the limitations of the traditional techniques. The same characteristics that make the GA so robust can make it more computationally intensive and hence slower than other methods. To improve the performance of the conventional GA, a hybrid scheme using GA and SQP method is proposed.

Background of the GA. 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 individuals, $P(t) = x_1^t$, ..., x_n^t for generation t. Each individual 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 an individual is assigned proportionally to the value of the objective function of the individuals. Then, a new population (generation $t+1$) is formed by selecting more fit individuals (selected 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 individuals by a small change in a single individual ($m_i : S \to S$), and higher order transformations c_j (crossover type), which create new individuals by combining parts from several individuals ($c_j : S \times \ldots \times S \to S$). After some number of generations the algorithm converges - it is expected that the best individual represents a near-optimum (reasonable) solution. The combined effect of selection, crossover and mutation gives so-called reproductive scheme growth equation [4]:

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

A pseudo code of a GA is presented as:

```
**1** Set generation number to zero ($t = 0$)
**2** Initialise usually random population of individuals ($P(0)$)
**3** Evaluate fitness of all initial individuals of population
**4** Begin major generation loop in $k$:
    **4.1** Test for termination criterion
    **4.2** Increase the generation number
    **4.3** Select a sub-population (select $P(i)$ from $P(i-1)$)
    **4.4** Recombine the genes of selected parents (recombine $P(i)$)
    **4.5** Perturb the mated population stochastically (mutate $P(i)$)
    **4.6** Evaluate the new fitness (evaluate $P(i)$)
**5** End major generation loop
```

Background of the SQP algorithm. SQP is one of the most popular and robust algorithms for nonlinear continuous optimization. The general optimization problem to minimize an objective function f under nonlinear equality and inequality constraints is [5]:

$$\min_{x\in\mathcal{R}^n} \; f\left(x\right), \, c\left(x\right) = 0, \, b\left(x\right) \geq 0, \, x_l \leq x \leq x_u,$$

where x is an n-dimensional parameter vector. It is assumed that all problem functions $f(x)$, $c(x)$ and $b(x)$ are continuously differentiable on the whole $\mathcal{R}^n$.

At an iteration x_k (for the equality constrains), a basic SQP algorithm defines an appropriate search direction d_k as a solution to the QP subproblem

$$\begin{aligned} &\min_{d\in\mathcal{R}^n} \; f\left(x_k\right) + g\left(x_k\right)^{\mathrm{T}} d + \tfrac{1}{2} d^{\mathrm{T}} \nabla^2_{xx} f\left(x\right) + \sum_{i=1}^{t} \lambda^i \nabla^2_{xx} c^i\left(x\right) \\ &\text{s.t. } c\left(x_k\right) + A\left(x_k\right) d = 0 \end{aligned}$$

is equal to, or is a symmetric approximation for, the Hessian of the Lagrangian.

A pseudo code of SQP algorithm could be presented as:

```
**1** Set the initial point $x = x_0$
**2** Set the Hessian matrix ($H_0 = I$)
**3** Evaluate $f_0, g_0, c_0$ and $A_0$
**4** Solve the QP subproblem to find search direction $d_k$
**5** Update $x_{k+1} = x_k + \alpha_k d_k$
**6** Evaluate $f_{k+1}, g_{k+1}, c_{k+1}$ and $A_{k+1}$
**7** Convergence check
    If yes, go to exit
    If no, obtain $H_{k+1}$ by updating $H_k$ and go back to Step 4
```

3 Numerical results and discussion

E. coli MC4110 fed-batch cultivation model. The mathematical model of the considered process can be represented by [11]:

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

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

$$\frac{dA}{dt} = \frac{1}{Y_{A/X}} \mu_{max} \frac{A}{k_A + A} X - \frac{F_{in}}{V} A \tag{3}$$

$$\frac{dpO_2}{dt} = -\frac{1}{Y_{pO_2/X}} \mu_{max} \frac{pO_2}{k_{pO_2} + pO_2} X + k_L a (pO_2^* - pO_2) - \frac{F_{in}}{V} pO_2 \tag{4}$$

$$\frac{dV}{dt} = F_{in} \tag{5}$$

where: X is biomass concentration, [g/l]; S - substrate concentration, [g/l]; A - acetate concentration, [g/l]; pO_2 - dissolved oxygen concentration, [%]; pO_2^* - saturation concentration of dissolved oxygen, [%]; F_{in} - feeding rate, [l/h]; V - bioreactor volume, [l]; S_{in} - substrate concentration in the feeding solution, [g/l]; μ_{max} - maximum value of the specific growth rate, $[h^{-1}]$; k_i - saturation constants; k_La - volumetric oxygen transfer coefficient, $[h^{-1}]$; $Y_{i/X}$ - yield coefficients, [-]. For the parameter estimation problem real experimental data of the *E. coli MC4110* fed-batch cultivation process are used. The cultivation condition and the experimental data have been presented in [12].

For comparison of the performance of the presented here hybrid GA-SQP with pure GA and SQP a two cases are examined. In the first case (*Case 1*) the system (1)-(2) and (5) is considered. The estimated parameters are: μ_{max}, k_S and $Y_{S/X}$. In the second case (*Case 2*) the full system (1)-(5) is considered with unknown parameters: μ_{max}, k_S, k_A, k_{pO_2}, $Y_{S/X}$, $Y_{A/X}$, $Y_{pO_2/X}$ and pO_2^*. The objective function is presented as a minimization of a distance measure J between experimental and model predicted values, represented by the vector $\mathbf{y}$:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{m} \{[\mathbf{y}_{exp}(i) - \mathbf{y}_{mod}(i)]_j\}^2 \rightarrow min \tag{6}$$

where n is the number of data for each state variable m; $\mathbf{y}_{exp}$ - the experimental data; $\mathbf{y}_{mod}$ - model predictions with a given set of the parameters.

Algorithm parameters. Based on results in [11, 10], genetic algorithm operators and parameters for considered here parameter identification of fermentation process are as follows: A binary 20 bit representation is considered. The selection method used here is the roulette wheel selection. A double point crossover and a bit inversion mutation are applied. Crossover rate should generally be high - here it is set to 70%. Mutation is randomly applied with low probability - 0.01. A value of 97% for the rate of the selected individuals (generation gap) is accepted. Particularly important parameters of GA are the population size and number of generations. If there are too low number of chromosomes, GA has a few possibilities to perform crossover and only a small part of search space is explored. On the other hand, if there are too many chromosomes, GA slows down. The number of individuals is set to 100. The number of generations is 200 (pure GA) and 10 (hybrid GA-SQP). The division of the hybrid's time between the two methods influences the efficiency and the effectiveness of the search process. Numerous tests are performed to find the optimal division of the algorithm's time. The GA is run for 5, 10, 15, 20, 25, 30 generations before the SQP algorithm is started. The obtained results show that the optimal number of generations is 10. For 10 generations the GA reaches near optimum solution, which is a good initial point for the SQP algorithm. The use of 20 or 30 generations reflects mainly on the computational cost and has negligible improvement on the initial point.

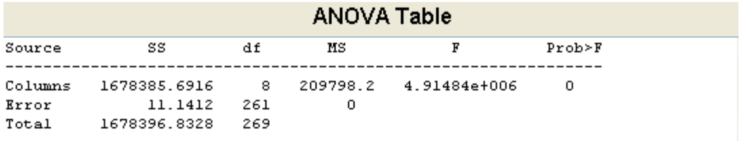
Results from parameter identification. All computations are performed using a PC/Intel Core 2 Quad CPU Q8200 @2.34GHz platform running Windows

XP, Matlab 7.5 environment. Initially the algorithms (GA, SQP and GA-SQP) were tested for parameter estimation of model (1)-(5) using generated data. The results were explicit: (i) The estimates of SQP were very sensitive to the initial points. The algorithm reached the value of J between 0.0063 and 0.013 according to the considered initial point. The computational time was 40-60 s. (ii) The best result of GA is J=0.0137 (for 5 runs). The computational time was 117.04 s. (iii) The result from GA-SQP was J=0.0063 for each run of the algorithm. The GA is run for 10 generations. The computational time varied between 55-75 s. In the second step the algorithms are used for parameter estimation of two considered models (*Case 1* and *Case 2*) using real experimental data. The experimental data were used without filtration or any processing. The idea was to test the algorithms in such hard real conditions. The numerical results (*Case 2*) from the parameter identification are presented in Table 1.

Table 1. Search parameters utilized in the different algorithms

Search parameter	**GA**	**SQP**	**GA-SQP**
μ_{max}	0.4780	0.4742	0.4741
k_S	0.0145	0.0148	0.0148
$1/Y_{S/X}$	2.0313	2.0137	2.0137
$1/Y_{A/X}$	8.6169	12.3012	9.3365
$1/Y_{pO_2/X}$	0.0340	0.0388	0.0368
k_A	54.3274	67.1266	50.9262
k_{pO_2}	0.0017	0.001	0.001
k_La	282.4246	300.0062	282.9080
pO_2^*	21.2696	21.2988	21.2988

In Table 1, considering GA and GA-SQP the average values of 30 runs are presented. One-factor ANOVA analysis is performed to see if the means of the obtained 30 groups are equal. The results are displayed in Fig. 1. The estimated parameter values of the algorithms are in admissible ranges [6, 14]. The parameters estimates obtained by the three algorithms are very close. The only exceptions are the parameters $Y_{A/X}$ and k_A. The acetate concentration during the cultivation process is very small in comparison with the concentrations of the other state variables. So, the influence of acetate error on the objective function is smaller than the other errors. In contrast, the influence of biomass error is tolerable and the result is almost equal estimates of μ_{max}, k_S and $Y_{S/X}$.



ANOVA Table

```
Source          SS          df        MS             F          Prob>F
---------------------------------------------------------------------
Columns   1678385.6916       8    209798.2    4.91484e+006       0
Error          11.1412     261          0
Total     1678396.8328     269
```

Fig. 1. One-factor ANOVA results

For the simple task (*Case 1*) the SQP and the hybrid GA-SQP algorithms obtain same values of the objective function. Moreover, the pure SQP has better convergence time. SQP enjoy global convergence guarantees and requires few function evaluations to locate a solution point. The pure GA has obtained almost the same J but for greater computational time - 88.1875 s. GA reaches the area near an optimum point relatively quickly but it took many function evaluations to achieve convergence. In *Case 2* when 9 parameters were estimated simultaneously the effectiveness of the hybrid GA-SQP is more evident. The result values of the objective function (Eq. (6)) and computation time are given in Table 2. The best result ($J = 6.5470$) is obtained using the hybrid GA-SQP after a total computation time of 72.1719 s. The GA-SQP hybrid technique merges a lot of features of the GA and the SQP optimality criterion. A combination of a genetic algorithm and a local search method speed up the search to locate the exact global optimum. It exhibits the robust global search capability of the GA while preserving the efficient local search capability afforded by SQP. Very close result was obtained by SQP with a "good" initial point. If the initial point is "bad" the algorithm falls in another local extrema with $J = 6.6822$. The solution depends on the choice of the start points as the pure SQP usually seeks a solution in the neighborhood of the start point. The obtained result from the pure GA is $J = 6.5617$ for longer computational time of 208.75 s. Since pure GAs consider a group of points in each search space in each generation, they are best suited for global search. But their main operations (i.e. reproduction, crossover, and mutation) are not very efficient for local search. The obtained through generations objective function values are presented in Fig. 2. As it can be seen the best performance show hybrid GA-SQP.

Table 2. Results of the search methods in Case 2

Criterion	**GA**	**SQP**	**GA-SQP**
J	6.5617	$6.5483^1/6.6822^2$	6.5470
CPU time, s	208.7500	$67.8594^1/66.3281^2$	72.17198

^{1} "good" initial point, ^{2} "bad" initial point

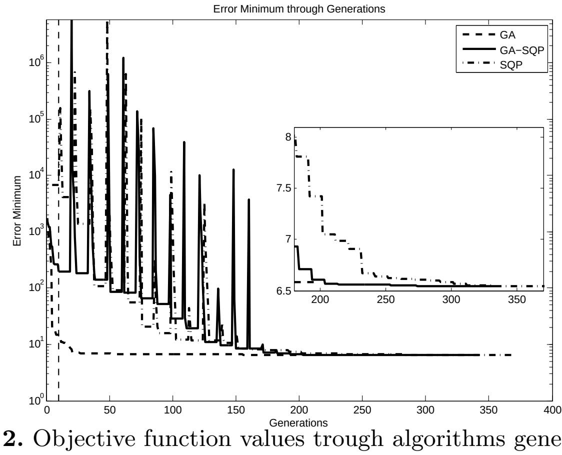


Fig. 2. Objective function values trough algorithms generations

A quantitative measure of the differences between modelled and measured values is another important criterion for the adequacy of a model. The graphical results of the comparison between the model predictions of state variables, based on hybrid GA-SQP algorithm estimations, and the experimental data points of the real *E. coli* cultivation are presented in Fig. 3.

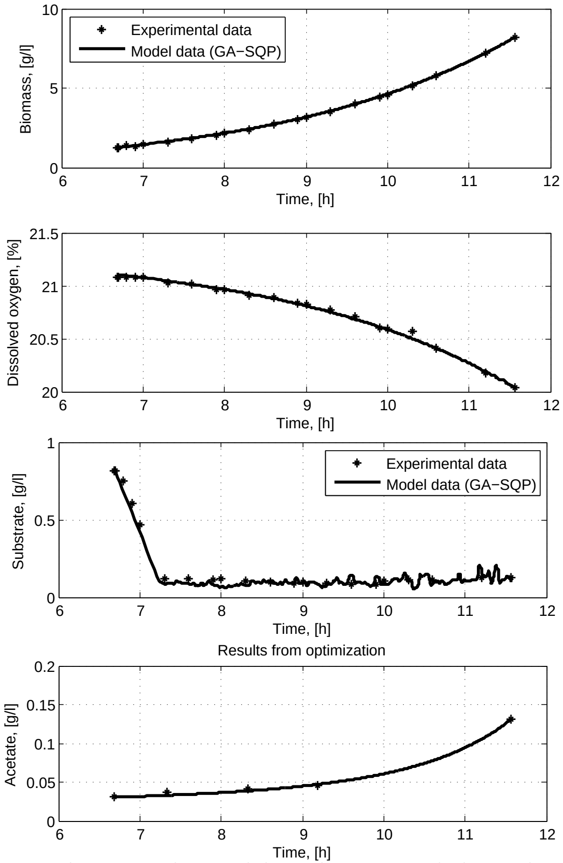


Fig. 3. Comparison between the model predictions and the real process variables

The presented graphics show a very good correlation between the experimental and predicted data.

4 Conclusion

In this paper a hybrid GA-SQP algorithm is proposed. In such a hybrid, applying a local search to the solutions guided by a genetic algorithm in the most promising region can accelerate convergence to the global optimum. The hybrid algorithm is compared with pure GA and SQP algorithms for parameter identification procedure. Algorithms performance is illustrated using a set of non-linear models of *E. coli MC4110* fed-batch cultivation process. As evident from graphical and numerical results, the proposed optimization hybrid algorithm performs very well. The algorithm takes the advantages of both GA's global search ability and SQP's local search ability, hence enhances the overall search ability and computational efficiency. The speed of convergence of the hybrid algorithm is superior to that of pure GA as well as the obtained objective function.

Acknowledgements

This work is supported by the National Science Fund Grants DMU 02/4 and DID 02-29.

References

1. Akpnar, S., Bayhan, G. M.: A Hybrid Genetic Aalgorithm for Mixed Model Assembly Line Balancing Problem with Parallel Workstations and Zoning Constraints, Engineering Applications of Artificial Intelligence, **24(3)** (2011) 449–457
2. Benjamin K. K., Ammanuel, A. N., David, A., Benjamin, Y. K.: Genetic Algorithm using for a Batch Fermentation Process Identification, J of Applied Sciences, **8(12)** (2008) 2272–2278
3. Byrd R. H., Curtis, F. E., Nocedal, J.: Infeasibility Detection and SQP Methods for Nonlinear Optimization, SIAM Journal on Optimization, **20** (2010) 2281–2299
4. Goldberg, D.: Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Weslcy Publishing Company, Massachusetts (1989)
5. Gill, Ph. E., Wong, E.: Sequential Quadratic Programming Methods, UCSD Department of Mathematics, Technical Report NA-10-03 (2010)
6. Levisauskas, D., Galvanauskas, V., Henrich, S., Wilhelm, K., Volk, N., Lubbert, A.: Model-based Optimization of Viral Capsid Protein Production in Fed-batch Culture of Recombinant *Escherichia coli*, Biopr&Biosys Eng, **25** (2003) 255–262
7. Mateus da Silva, F. J., Prez, J. M. S., Pulido, J. A. G., Rodrguez, M. A. V.: AlineaGA - A Genetic Algorithm with Local Search Optimization for Multiple Sequence Alignment, Appl Intell, **32** (2010) 164–172
8. Nocedal, J., Wright, S. J.: Numerical Optimization, Springer Series in Operations Research, Springer (2006)
9. Paplinski, J. P.: The Genetic Algorithm with Simplex Crossover for Identification of Time Delays, Intelligent Information Systems, (2010) 337–346
10. Roeva, O.: Improvement of Genetic Algorithm Performance for Identification of Cultivation Process Models, Advanced Topics on Evolutionary Computing, Book Series: Artificial Intelligence Series-WSEAS (2008) 34–39
11. Roeva, O.: Parameter Estimation of a Monod-type Model based on Genetic Algorithms and Sensitivity Analysis, LNCS, Springer, **4818** (2008) 601–608
12. Roeva, O., Pencheva, T. , Hitzmann, B., Tzonkov, St.: A Genetic Algorithms Based Approach for Identification of *Escherichia coli* Fed-batch Fermentation, Int J Bioautomation, **1** (2004) 30–41
13. Tseng, Lin-Yu, Lin, Ya-Tai: A Hybrid Genetic Local Search Algorithm for the Permutation Flowshop Scheduling Problem, Europen J of Operational Res, **198(1)** (2009) 84–92
14. Zelic, B., Vasic-Racki, D., Wandrey, C., Takors, R.: Modeling of the Pyruvate Production with *Escherichia coli* in a Fed-batch Bioreactor, Biopr&Biosys Eng, **26** (2004) 249–258