OPTIMIZATION OF GENE EXPRESSION WITH A GENETIC ALGORITHM

Angel Castellanos, Rafael Lahoz–Beltra

Abstract: *The central dogma of biology, in the simplest version, comprises two stages, transduction and translation, translating non-functional DNA information into an operational form represented by a protein. In this paper we simulated the optimization of the parameters that regulate genetic expression being the main contribution the proposal of the evolutionary surface of the parameters space. In particular we are referring to the production and degradation rates of mRNA and proteins. In addition, some methodological suggestions are made on how to study the regulation of genetic expression and on the different ways of reporting the results, either through bacterial agents or via differential equations. This work may be relevant in synthetic biology, bioinformatics or artificial life, as well as other areas of research.*

Keywords: *Optimization of space parameters, bacterial agents, genetic algorithms, gene expression regulation.*

ITHEA Keywords: *J.3 Life and Medical Sciences.*

Introduction

The regulation of gene expression is one of the fundamental evolutionary milestones for the maintenance of life. In broad terms, the expression of a gene comprises the translation of information from a gene (or DNA) into a functional molecule (protein or RNA). This translation is a process that must be subject to fine adjustment, since a lack of adjustment is usually related to pathological states, e.g. cancer [Lakatos et al., 2017]. Genetic regulation requires that in the course of evolution the cells adjust certain rates appropriately, the optimization of which is fundamental for the proper operation of the central dogma of biology (Figure 1). This dogma [Lahoz-Beltra, 2012] explains how non-functional DNA information is translated into a functional and operational molecule in the form of protein. The flow of information from DNA to proteins takes place at the 'hardware level' through a mechanism known as protein biosynthesis which includes two stages. A first stage is called transcription by which DNA information is translated into an intermediate molecule known as messenger RNA or mRNA. This step is followed by a second stage known as translation in which the information carried by the mRNA is translated into a final molecule, i.e. a protein. At the molecular level, and in a very simplified way, transcription and translation both require complex molecular machines, specifically RNA polymerase

and ribosomes, respectively. In the course of the transcription RNA polymerase, a class of proteins called enzyme, separates the two strands of DNA by exposing them as a template for mRNA synthesis. At this stage there are two fundamental parameters, the mRNA production or α_{mRNA} , i.e. the mRNA synthesis (mRNA/min) and the mRNA degradation or β_{mRNA} , i.e. a process whose duration (minutes) depends on the half-life of mRNA. In addition, and during the translation stage, the ribosomes will guide, in the order specified by the mRNA, the binding one after the other of the amino acids, i.e. the binding of the subunits from which will result the final protein. In this step there are two fundamental parameters. On one side, the protein production rate or $\alpha_{protein}$ (number of molecules per minute and per mRNA molecule) and on the other side, the protein degradation rate or $\beta_{protein}$. Therefore, an elementary model of genetic regulation will require the calibration of the four parameters described above. By optimizing these values the cells will control the amount of synthesized proteins. However, how did the cells optimize the value of these parameters during evolution? Assuming that the 'hardware' represented in Figure 1 is a 'molecular machine', and adopting an evolutionary computing approach, what would be the general appearance of the evolutionary surface of the parameter space? This work is a theoretical speculation about the evolutionary surface of the parameter space that regulates protein biosynthesis in the bacterium *E. coli*. The methodology introduced by us in this paper could be useful in the study of the optimization of gene regulation, one of the most relevant topics in molecular evolution. Aiming to facilitate the present study we model transcription and translation as two independent optimization problems, although coupled together. The classical models of genetic expression regulation are based on the use of differential equations [Alves and Dilao, 2005]:

$$\begin{cases} \dfrac{d[mRNA]}{dt} = \alpha_{mRNA} - \beta_{mRNA}[mRNA] \\ \dfrac{d[protein]}{dt} = \alpha_{protein}[mRNA] - \beta_{protein}[protein] \end{cases} \quad (1)$$

We will refer [mRNA] to the concentration of mRNA and [protein] to the concentration of the protein.

Methodology and modeling

In this paper gene expression regulation model adopts an elementary model, as it does not include details about RNA polymerase, cofactors or the role of synthesized protein in the repression of transcription. The model also does not include a sub-model that simulates the function performed by the ribosome. Therefore, it is a phenomenological model of genetic expression, simulating the optimization

of the genetic expression of a green fluorescent protein (GFP). We assume that transcription and translation are studied in a simple organism such as *E. coli* bacterium.

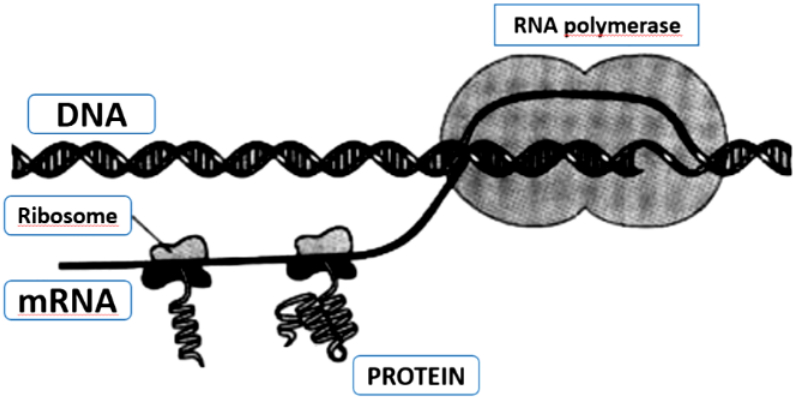

Figure 1. Central dogma of biology describes a flow of information DNA → mRNA → proteins (for an explanation see text).

GFP is a protein that is commonly used in synthetic biology projects, composed of 238 amino acids and with the feature that it emits green fluorescence when exposed to light. In 2012 [Klavins, 2012] simulated the central dogma of molecular biology using as an example the gene expression of the GFP protein. For this purpose the model of the central dogma was coded in Gro 4.0 cellular programming language (see Appendix), a language introduced by Klavins and co-workers [Jang et al., 2012]. In the aforementioned simulation the optimum rates or parameters (Figure 2) were set by the authors of the model with the following values: transcription rates were $\alpha_{mRNA} = 69.4$ mRNA/min, $\beta_{mRNA} = 3.69$ /min and translation rates $\alpha_{protein} = 3.0$ proteins per minute per mRNA, $\beta_{protein} = 0.01$ /min.

The purpose of our model was to study the genetic expression of the GFP protein using evolutionary computational methods. Our goal is to understand how natural selection was able to find the optimal transcription [Perez-Ortin et al., 2013] and translation rate values on which depends the regulation of protein biosynthesis. In this paper we question about the general appearance of the evolutionary surfaces of both production rates and degradation rates of mRNA and GFP. For this purpose, we take inspiration from models in which the optimization of some parameter plays an essential role in the dynamics of the phenomenon studied, such is the case of Max-Min quadratic equations in optimization problems. For example, in ecological informatics the DaisyWorld model [Nuño et al., 2010] simulates a planet whose inhabitants, two species of daisies, regulate the temperature of the planet. The model uses quadratic equations to simulate the optimal growth rate of daisies as a function of temperature. In theoretical genetics, quadratic expressions have been used previously to model the fitness landscape of gene-expression level [Bedford and Hartl, 2009]. In these examples, is generally used the vertex form of

a quadratic equation, i.e. $y = a(x-p)^2 + q$. The vertex has coordinates (p, q) being y a maximum (a<0) or minimum (a>0) when x = -p.

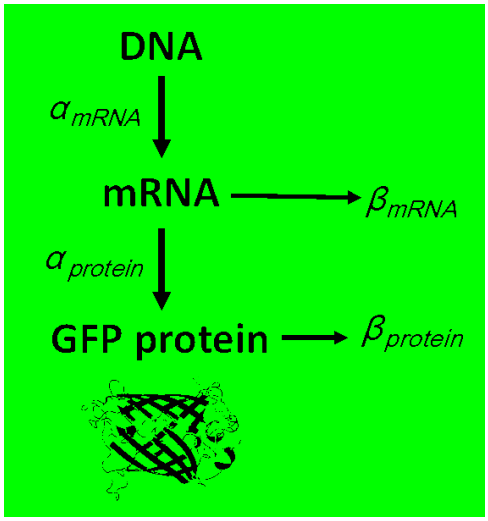

Figure 2. Diagram depicting the expression model of the GFP protein gene. The optimum parameter values were set in accordance with [Klavins, 2012]. At the bottom is shown the protein molecule in black and white (Retrieved May 21, 2018, from European Bioinformatics Institute, http://www.ebi.ac.uk/pdbe/entry/pdb/1ema).

On the basis of theoretical reasoning and simulation experiments previously conducted, we propose the following evolutionary surfaces. In the transcription step, the optimization of mRNA production and degradation rates was simulated in three different evolutionary scenarios (Figure 3):

$$F_1(x_1x_2) = \frac{2 - (0.00025(x_1 - \alpha_{mRNA})^2) - (0.1(x_2 - \beta_{mRNA})^2)}{2} \quad (2)$$

$$F_2(x_1x_2) = \frac{2 - (0.00025(x_1 - \alpha_{mRNA})^2) + (0.1(x_2 - \beta_{mRNA})^2)}{2} \quad (3)$$

$$F_3(x_1x_2) = \frac{2 + (0.00025(x_1 - \alpha_{mRNA})^2) + (0.1(x_2 - \beta_{mRNA})^2)}{2} \quad (4)$$

with the search domain represented in Table 1.

Table 1. **F(x_1,x_2) search domain**

	α_{mRNA}	β_{mRNA}
lower	5	0.50
upper	133	6.88

Notice how $F_1(x_1,x_2)$, $F_2(x_1,x_2)$ and $F_3(x_1,x_2)$ functions represent different optimization problems. In the first and second functions the optimal fitness search is a maximization problem, while for the third function the optimal fitness search is a minimization problem. In a similar way we simulate for the translation step the optimization of the production and degradation rates of the GFP protein. However, simulations were conducted only in two optimization environments, in one case as a maximization problem, in the other case as a minimization problem (Figure 4).

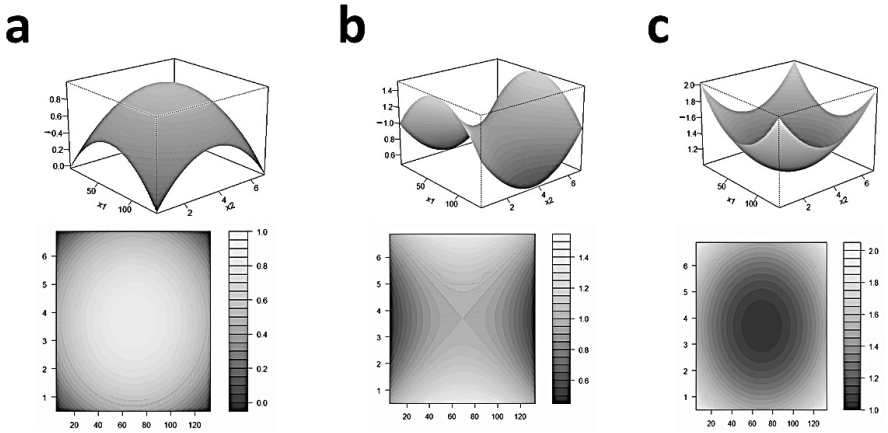

Figure 3. Evolutionary surfaces for the transcription parameters. Landscapes (a) $F_1(x_1,x_2)$, (b) $F_2(x_1,x_2)$ and (c) $F_3(x_1,x_2)$.

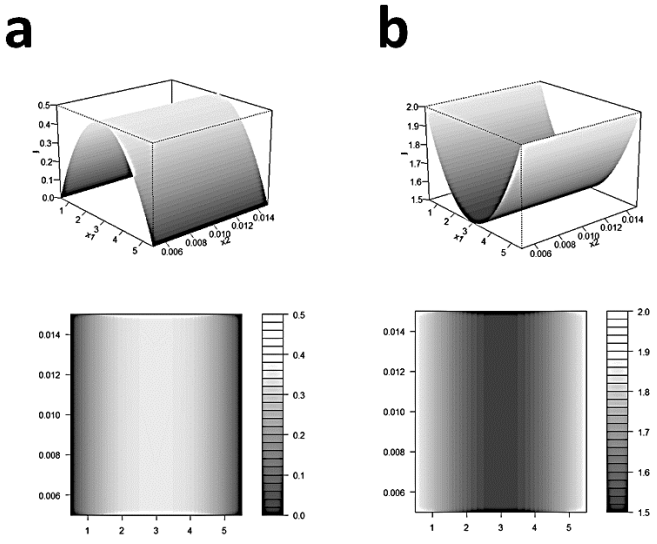

Figure 4. Evolutionary surfaces for the translation parameters. Landscapes (a) $F_4(x_1,x_2)$ and (b) $F_5(x_1,x_2)$.

The functions of the evolutionary surfaces and their search domain (Table 2) were as follows:

$$F_4(x_1x_2) = \frac{2-(0.16(x_1-\alpha_{protein})^2)-(40000(x_2-\beta_{protein})^2)}{2} \quad (5)$$

$$F_5(x_1x_2) = \frac{2+(0.16(x_1-\alpha_{protein})^2)+(40000(x_2-\beta_{protein})^2)}{2} \quad (6)$$

Table 2. **$F(x_1,x_2)$** ***search domain***

	$\alpha_{protein}$	$\beta_{protein}$
lower	0.5	0.005
upper	5.5	0.015

The optimization was conducted by means of a genetic algorithm, using the GA R-package v3.0.2 [Scrucca, 2013, 2017] with the following GA settings: type = real-valued, population size = 50, number of generations = 1000, elitism = 2, crossover probability = 0.8 and mutation probability = 0.1.

MMOGE: A method to study the optimization of genetic expression

The estimation of the parameters of a model, e.g. in this study, is one of the most common tasks in many disciplines, whether in synthetic biology, bioinformatics or in some models of artificial life. In the case we study in this paper the problem is simplified since the degradation rates can be easily estimated. This is because the values are known since they are the half-life of the molecules, e.g. the values of β_{mRNA} and $\beta_{protein}$ are sufficiently well known in the laboratories. Therefore, in the expressions (2), (3), (4) and (5) the problem is simplified to estimating the value of the production rates, i.e. α_{mRNA} and $\alpha_{protein}$, which doesn't mean it's a trivial problem. In this work, and taking into account the previous comments, we propose a methodology that we have referred to as MMOGE (*Max-Min Optimization Gene Expression*) and which comprises the following steps:

1. Consider the transcription and translation separately. We will assume that the optimal value of the production α and degradation β rates is fitted to a parabolic or Max-Min quadratic function which vertex is the maximum of the function (a<0 and x=-p).

2. By using some computer algebra system (CAS), we can tentatively set the width of the parabolic function, i.e. the search domain. For instance, in the problem discussed in this paper we use wxMaxima 16.04.2, adjusting the width of the quadratic function for the production rate of mRNA:

```
(%i2)   k:0.00025; x_opt:69.4;
-->     plot2d ([f(x), (1-k*(x-x_opt)^2)], [x,5,133], [y,0,1]);
```

Created with wxMaxima.

3. Based on the mathematical expressions of the parabolic function of each of the rates, the production rate and the degradation rate, we can combine them in a 3D function. The function obtained will have a single maximum that represents the optimum of both rates, e.g. expressions (2), (3), (4) and (5).

4. On the basis of the expressions of the previous step (3) and with a genetic algorithm, we can simulate the optimization of the genetic expression by natural selection.

5. We can finally 'test' the effect on the biological level of the optimized parameter values. The test can be conducted by simulation experiments, either by means of differential equations, or bacterial agents (e.g. Gro, Figure 6), etc.

Simulation results

In this paper we have shown how it is possible to simulate the evolution of the parameter values that regulate genetic expression in *E. coli* (Figure 6). The study uses a well-known example, such as the case of the gene of the GFP protein. Figure 5 shows one of the characteristic performance graphs of this type of evolutionary simulation experiments.

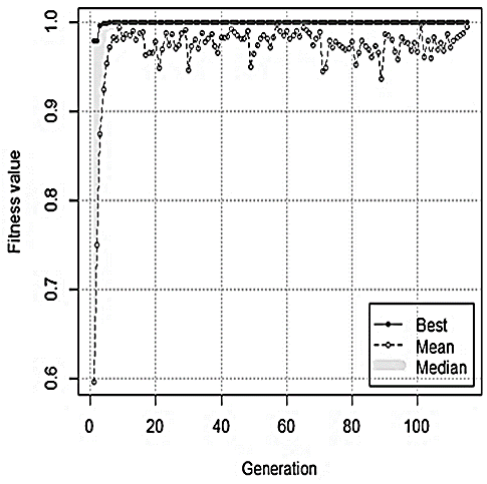

Figure 5. Performace graph for $F_1(x_1,x_2)$.

Tables 3 and 4 show the optimized values of the rates according to the genetic algorithm. In the case of transcription, $F_1(x_1,x_2)$ and $F_3(x_1,x_2)$ evolutionary surfaces reflect plausible environments, except for $F_2(x_1,x_2)$, where one parameter is not correctly optimized. By using this evolutionary surface we simulate a lower rate of mRNA degradation. It is at present known [Lakatos et al., 2017] that protein abundance due to poor genetic regulation, e.g. an excess of p53 protein, has a high correlation with the development of cancerous tumors. This situation is successfully simulated in our model, with the genetic expression evolving to a lower than normal mRNA degradation value. Moreover, the role of RNA today goes beyond protein biosynthesis by changing the classical paradigm of the central dogma of biology: RNA regulates gene expression, and can influence genome instability, e.g. by participating in the

survival of a cancerous tumor [Amirkhah et al., 2016]. One of the features of the model is the possibility that the results, the output, can be displayed in different formats. Figure 6 shows the output in a colony of bacterial agents (see Appendix) while Figure 7 shows the results through a system of differential equations.

Using the model described in this paper it is possible to perform more sophisticated experiments. In fact, there are several factors with an effect on transcription and translation that could be simulated [Milo and Phillips, 2015]. For instance, the antibiotic rinfampin has an effect on the beginning of the transcription, effect that could be simulated via α_{mRNA} parameter. For example, we could also simulate errors in the folding of a protein, increasing the translation speed or $\alpha_{protein}$.

Also, setting the values $\alpha_{mRNA} < \alpha_{protein}$ it is possible to simulate a 'collision' between the ribosome and RNA polymerase, resulting in a failure of the protein synthesis.Therefore, the optimal operation of the molecular machinery takes place if $\alpha_{mRNA} > \alpha_{protein}$. In summary, the proposed model opens up many possibilities for simulating genetic expression and the central dogma of biology.

Table 3. mRNA transcription rates obtained with the genetic algorithm

Evolutionary surface	mRNA production (α_{mRNA})	mRNA degradation (β_{mRNA})
$F_1(x_1,x_2)$	69.4001	3.6900
$F_2(x_1,x_2)$	70.5802	0.5007
$F_3(x_1,x_2)$	69.4008	3.6900

Table 4. Protein translation rates obtained with the genetic algorithm

Evolutionary surface	Protein production ($\alpha_{protein}$)	Protein degradation ($\beta_{protein}$)
$F_4(x_1,x_2)$	3.0000	0.0099
$F_5(x_1,x_2)$	3.0006	0.0099

Figure 6. Model of a bacterial agents colony simulated in Gro language (see Appendix, program written by [Klavins, 2012]).

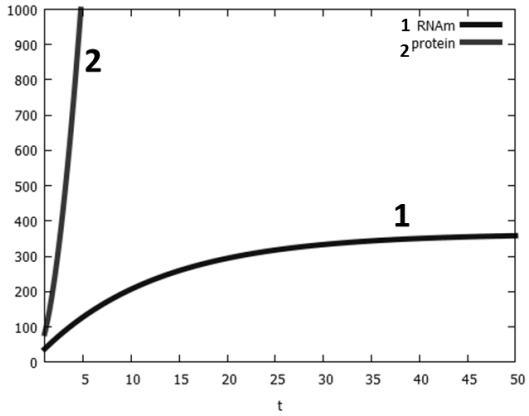

Figure 7. Simulation with differential equations (1) using the optimized values of the parameters according to the experiment described in this paper.

Appendix

The following program [Klavins, 2012] simulates in a colony of *E. coli* (Figure 6) the central dogma applied to the expression of the gene responsible for the GFP protein:

```
include gro

set ( "dt", 0.01 );
alpha_r := 69.4 / 2.35;              // mRNA / min / fL
beta_r := - log ( 0.5 ) / 3.69;      // 1/min
alpha_p := 3.0;                      // protein/min/fL/RNA
beta_p := 0.01;                      // 1/min

program gfp() := {
  mRNA := 0;
  gfp := 0;

  rate ( alpha_r * volume ) : { mRNA := mRNA + 1 };
  rate ( beta_r * mRNA ) :    { mRNA := mRNA - 1 };
  rate ( alpha_p * mRNA ) :   { gfp := gfp + 1 };
  rate ( beta_p * gfp ) :     { gfp := gfp - 1 };

};
```

```
set ( "gfp_saturation_max", 1000 );
set ( "gfp_saturation_min", 800 );

ecoli ( [ x := 0, y := 0 ], program gfp() );
```

Another possibility to show the output is the classic simulation with differential equations (Figure 7):

```
Model parameters
(%i4)   a11:29.53; a12:0.081; a21:3.0; a22:0.01;
Initial conditions
(%i5)   y0_RNAm: 10;
(%i6)   y0_protein: 10;
System of differential equations
(%i7)   EDO_RNAm: -a12*RNAm + a11;
(%i8)   EDO_protein: -a22*protein + a21* RNAm;
(%i10)  fll(x) := [first(x),last(x),length(x)]$ declare(fll,evfun)$
4-order Runge-Kutta method
(%i11)  puntos: rk([EDO_RNAm,EDO_protein ],[RNAm, protein], [y0_RNAm, y0_protein], [t, 0, 100, 0.1])$
(%i12)  %, fll;
(%i13)  AL: makelist([puntos[i][1], puntos[i][2]], i, 1, length(puntos))$
(%i14)  %,fll;
(%i15)  BL: makelist([puntos[i][1], puntos[i][3]], i, 1, length(puntos))$
(%i16)  %, fll;
Numerical solution curves
(%i17)  plot2d( [[discrete, AL],[discrete, BL]], [x, 1, 50],[style, [lines, 5]], [y, 1, 1000], [ylabel, " "], [xlabel, "t"], [legend, "RNAm", "protein"])$
```

Created with wxMaxima.

Bibliography

[Alves and Dilao, 2005] F. Alves, R. Dilao. 2005. A simple framework to describe the regulation of gene expression in prokaryotes. C.R. Biologies 328: 429-444.

[Amirkhah et al., 2016] R. Amirkhah, A. Farazmand, O. Wolkenhauer, U. Schmitz. 2016. RNA systems biology for cancer: From diagnosis to therapy. Systems Medicine, Methods in Molecular Biology (U. Schmitz, O. Wolkenhauer, eds.) vol. 1386. New York: Springer Science + Business Media.

[Bedford and Hartl, 2009] T. Bedford, D. L. Hartl. 2009. Optimization of gene expression by natural selection. PNAS 27 106(4): 1133–1138.

[Jang et al., 2012] S.S. Jang, K.T. Oishi, R.G. Egbert, E. Klavins. 2012. Specification and simulation of synthetic multicelled behaviors. ACS Synthetic Biology 1: 365-374.

[Klavins, 2012] E. Klavins, (2012, July). gro. The cell programming language. Retrieved from http://depts.washington.edu/soslab/gro/docview.html

[Lahoz-Beltra, 2012] R. Lahoz-Beltra. 2012. Cellular computing: Towards an artificial cell. International Journal "Information Theories and Applications" 19(4): 313-318.

[Lakatos et al., 2017] E. Lakatos, A. Salehi-Ryhani, M. Barclay, M.P.H. Stumpf, D.R. Klug. 2017. Protein degradation rate is the dominant mechanism accounting for the differences in protein abundance of basal p53 in a human breast and colorectal cancer cell line. PLoS ONE 12(5): e0177336. https://doi.org/10.1371/journal.pone.0177336.

[Milo and Phillips, 2015] R. Milo, R. Phillips. 2015. Cell Biology by Numbers. Retrieved from http://book.bionumbers.org/about-us/

[Nuño et al., 2010] J.C. Nuño, J. de Vicente, J. Olarrea, P. López, R. Lahoz-Beltra. 2010. Evolutionary daisyworld models: A new approach to studying complex adaptive systems. Ecological Informatics 5: 231-240.

[Perez-Ortin et al., 2013] J. Perez-Ortin, D. A. Medina, S. Chavez, J. Moreno. 2013. What do you mean by transcription rate? Bioessays 35: 1052-1062.

[Scrucca, 2013] L. Scrucca. 2013. GA: A package for genetic algorithms in R. Journal of Statistical Software 53/4: 1-37. https://www.jstatsoft.org/v53/i04/

[Scrucca, 2017] L. Scrucca. 2017. On some extensions to GA package: hybrid optimisation, parallelisation and islands evolution. The R Journal 9/1: 187–206. https://journal.r-project.org/archive/2017/RJ-2017-008.

Authors' Information

Rafael Lahoz–Beltra *– Department of Biodiversity, Ecology and Evolution (Biomathematics), Faculty of Biological Sciences, Complutense University of Madrid, 28040 Madrid, Spain; e-mail: lahozraf@ucm.es*

Major Fields of Scientific Research: Evolutionary computation, bioinspired algorithms.

Angel Castellanos *– Applied Mathematics Department. Universidad Politécnica de Madrid, Spain; Natural Computing Group, e-mail: angel.castellanos@upm.es*

Major Fields of Scientific Research: Artificial Intelligence, applied mathematics.