Cellular Genetic Algorithms for Identifying Variables in Hybrid Gene Regulatory Networks

Romain Michelucci, Vincent Callegari, Jean-Paul Comet, Denis Pallez

Cellular Genetic Algorithms for identifying variables in hybrid Gene Regulatory Networks

Romain Michelucci[0000-0001-6107-4394], Vincent Callegari, Jean-Paul Comet[0000-0002-6681-3501], and Denis Pallez[0000-0001-5358-8037]

Université Côte d'Azur, CNRS, I3S, Sophia Antipolis, France
firstname.name@univ-cotedazur.fr

Abstract. The hybrid modelling framework of gene regulatory networks (hGRNs) is a functional framework for studying biological systems, taking into account both the structural relationship between genes and the continuous time evolution of gene concentrations. The goal is to identify the variables of such a model, controlling the aggregated experimental observations. A recent study considered this task as a free optimisation problem and concluded that metaheuristics are well suited. The main drawback of this previous approach is that panmictic heuristics converge towards one basin of attraction in the search space, while biologists are interested in finding multiple satisfactory solutions. This paper investigates the problem of multimodality and assesses the effectiveness of cellular genetic algorithms (cGAs) in dealing with the increasing dimensionality and complexity of hGRN models. A comparison with the second variant of covariance matrix self-adaptation strategy with repelling subpopulations (RS-CMSA-ESII), the winner of the CEC'2020 competition for multimodal optimisation (MMO), is made. Results show evidence that cGAs better maintain a diverse set of solutions while giving better quality solutions, making them better suited for this MMO task.

Keywords: cellular genetic algorithm, epistatic and multimodal optimisation problem, RS-CMSA-ESII, hybrid GRN, chronotherapy, real-world application

1 Introduction

Studying the dynamics of gene regulatory networks (GRNs) aims to understand the various cellular processes and pathways that empower a living organism to carry out essential functions, such as metabolic processes and the ability to adapt to environmental disturbances. Modelling such GRNs allows novel and better cognisance of disease initiation and progression, opening new perspectives in pharmacological fields such as chronotherapy, which can be viewed as the practice of administering medication at specific times during the day, taking into account the body's natural rhythms and the varying effects of the treatment. By logically following the activation or inhibition of genes and proteins

under different conditions, biologist modellers can create models of these complex systems based on actual knowledge. That led to numerous modelling GRN frameworks such as *differential*, *stochastic* or *discrete* ones [22], each of them presenting its advantages and drawbacks. Whereas it is not too difficult to enumerate the different genes playing a role in a particular context as well as the known regulations between them, the common impediment remains the identification of the variables that govern the GRN dynamics.

In the present work, we consider *hybrid* frameworks [7] called hGRNs. They add to the discrete ones the time spent in each discrete state, allowing experimental observations to be represented as irregularly spaced time series of observable events. It has been shown that the hybrid model can exhibit these events in the same order and at the right time only if the dynamic variables that control the model behaviour satisfy a set of constraints. The design of these minimal constraints on the hGRN variables has been automated. An attempt has been made to use a continuous Constraint Satisfaction Problem (CSP) solver to extract solutions but faced difficulties when the number of variables increased [8]. Recently, [17] showed that the CSP, exhaustively characterising the set of solutions, can be expressed as a free optimisation problem (FOP) by indirectly handling constraints thanks to metaheuristics. The CSP was transformed into a non-separable, non-trivial, continuous, and single objective problem in which the search space increases exponentially with the number of genes in the hGRN. One limitation of this approach is that such algorithms are panmictic and can only identify one basin in the search space. From a modelling perspective, exhibiting a diverse sampling of biologically satisfactory solutions allows biologists to reason not only on one possible identification but also on a set of sensible ones. Therefore, this work focuses both on validating the previous approach on hGRNs involving more genes and complex dynamics and on the multimodal aspect of the identification problem. RS-CMSA-ESII is a new niching method for MMO that emerged as the most successful available method when robustness and efficiency are considered at the same time and does not make any assumptions such as distribution, shape, and size of the basins [2]. This CEC'2020 top niching-based algorithm is the logical choice to be tested as a baseline to gain more insights on its ability to find a set of solutions without having any assumptions on the modes. In the meantime, cGAs are well-known heuristics to tackle epistatic and multimodal tasks [4, 5] since the diversity maintenance is guaranteed thanks to the structure and ratio of the population, unlike RS-CMSA-ESII which employs mechanisms with different sub-populations running in parallel. So, this research aims to address the problem of the hGRN variables identification to obtain a diverse set of quality solutions for increasingly complex models while seeking to identify the most suitable method for achieving these goals.

To meet these objectives and based on the research hypotheses set out above, the article is organised as follows: Section 2 describes the hGRN continuous optimisation problem by detailing: (i) the definition of the hybrid model along with its dynamics, (ii) the experimental observations that serve as input, and (iii) how this problem has been treated as an FOP. Section 3 encompasses an overview

of RS-CMSA-ESII and cGAs from a multimodal perspective. Section 4 proposes experiments comparing CMA-ES, GA, multiple cGAs with varying ratios and structures, and RS-CMSA-ESII on three different hGRNs of increasing complexity. Experimental results and statistical tests are presented and discussed. Finally, conclusions are drawn in Section 5.

2 hGRN variables optimisation

2.1 Hybrid gene regulatory networks

Hybrid modelling of gene regulatory networks (GRNs) aims to describe the effect of regulations between genes in a biological system by taking into account the continuous time component. Traditionally, a GRN is a directed graph in which vertices express abstractions of one or multiple biological genes (v_1, v_2), and edges that act as either activation ($\rightarrow$) or inhibition ($\dashv$) represent regulations (Figure 1a). This static representation seems of limited interest since it does not integrate any dynamics. However, from Figure 1a, the corresponding discrete dynamics (Figure 1b) can be built. First, grey boxes are obtained from the previous GRN by enumerating all possible states $\mathbb{S}$: each grey square box identifies a *discrete state* $\eta \in \mathbb{S}$ defined by the level of the GRN genes. If we suppose that the maximum level of each gene v_i is 1, then the top right box is the state where each gene is expressed at its maximum level and is denoted by $\eta = (\eta_{v_1}, ..., \eta_{v_n})$. In Figure 1, this state is $\eta = (\eta_{v_1}, \eta_{v_2}) = (1, 1)$. From this first step, transitions between discrete states can be drawn (black arrows) and symbolise the discrete evolution of the concentration of the gene products. Although the obtained discrete state graph of Figure 1b is deeply interesting for logical reasoning about regulatory changes, it disregards temporal information, which is nevertheless crucial, for example, for optimising medical treatments by taking account of biological rhythms.

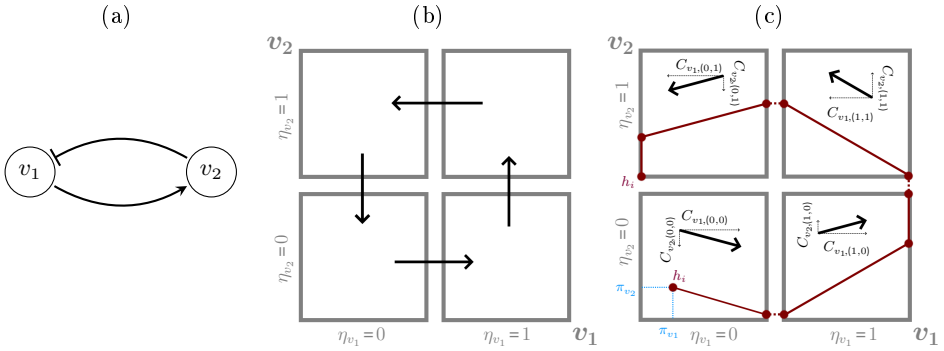

Fig. 1: Example of a GRN depicted as a directed graph (a), its discrete state graph (b), and a possible dynamic of its hybrid state graph (c) (taken from [17]).

The hybrid modelling framework adds the notion of temporal continuous evolution to the previous dynamics by adding linear continuous trajectories (red straight lines) to the discrete transitions of a GRN (pictured with dotted red lines in Figure 1c). On a trajectory, a point is called a *hybrid state* and given by its position π within a discrete state η. As an example, the initial hybrid state h_i in Figure 1c has the coordinates $\left((\eta_{v_1}, \eta_{v_2})^t, (\pi_{v_1}, \pi_{v_2})^t\right) = \left((0,0)^t, (0.25, 0.25)^t\right)$. To determine a complete trajectory through a set of discrete states, hGRN models require an initial hybrid state h_i and a vector of the evolution of concentrations in each discrete state, called *celerity vector*. This vector gives the direction and celerity of each gene $v \in V$ in a discrete state $\eta \in \mathbb{S}$, e.g. the celerity of v_1 in $\eta = (0,0)$ is denoted $C_{v_1,(0,0)}$. In the general case, the celerity of v in η is a floated value defined as $C_{v,\eta}$.

The aim is to identify celerity vectors to generate valid hGRN models of the biological system under study. Such a determination could help biologists make new interpretations about the possible dynamics of the system.

2.2 Biological knowledge

The identification process requires some input data, which allows the modeller to validate or not a possible valuation of continuous variables. While much work [10, 18, 20, 21] is based on gene expression data, our approach takes into consideration already-formalised information analysed by biologists derived from both biological data and expertise.

The formalism abstracts the knowledge extracted from biological experiments under the form of constraints on the global trajectory: it must (i) start from an initial hybrid state $h_i = (\eta_i, \pi_i)$, (ii) verify a triplet of properties in each successive discrete state $(\Delta t, b, e)$ where Δt expresses the time spent; b delineates the observed behaviours during the continuous trajectory; e specifies the next discrete state transition, and (iii) reach the final hybrid state $h_f = (\eta_f, \pi_f)$. Let us detail the biological knowledge (BK) used for the example of Figure 1c:

$$\{h_i\}\begin{pmatrix} 5.0 \\ noslide(v_2) \\ v_1+ \end{pmatrix};\begin{pmatrix} 7.0 \\ slide^+(v_1) \\ v_2+ \end{pmatrix};\begin{pmatrix} 8.0 \\ noslide(v_2) \\ v_1- \end{pmatrix};\begin{pmatrix} 4.0 \\ slide^-(v_1) \\ v_2- \end{pmatrix}\{h_f\}$$

$h_i = \left((0,0)^t, (\pi_{v_1}, \pi_{v_2})^t\right)$ represents both the initial and final state ($h_i = h_f$).

Starting from h_i, the time spent by the trajectory inside the discrete state $\eta = (0,0)$ is approximately 5 hours ($\Delta t = 5.0$). Within this state, the celerity should move towards the next discrete state of v_1 (v_1+) so as to increase the concentration level of gene v_1 until it reaches the right border without touching either the top or the bottom border ($noslide(v_2)$) and then jump into the neighbour state $\eta = (1,0)$. In this new discrete state, the trajectory evolves for 7 hours ($\Delta t = 7.0$) in the direction of $\eta_{v_2} = 1$ (v_2+) but, this time, the trajectory reaches the right border, which corresponds to the maximum admissible concentration of v_1 ($slide^+(v_1)$). This process continues until the trajectory reaches h_f. Any valuation of dynamic variables, i.e. celerity vectors, leading to a trajectory satisfying this BK is considered admissible.

2.3 Single objective and multimodal optimisation problem

Searching for celerity values that satisfy BK initially led to characterising the problem as a CSP and solving it by constraint-based programming [8]. On the one hand, this constraint-based programming method was able to exhaustively find the over-approximated sets of solutions, but as the number of dimensions increased, such a method was unable to extract even one particular solution.

A recent attempt [17] has recently formulated the problem as being single-objective by proposing an adequate fitness function consisting of three criteria and testing this approach on the hGRN model of Figure 1c (only two genes). In this preliminary study, the decision vector to be optimised consisted of finding the initial hybrid state h_i and all celerity values of all discrete states:

$$h_i, \{C_{v,\eta} | v \in V, \eta \in \mathbb{S}\} \tag{1}$$

Thus, for example, finding an admissible valuation of Figure 1c satisfying BK was equivalent to finding the optimal parameter set of:
$x = (h_i; C_{v_1,(0,0)}; C_{v_2,(0,0)}; C_{v_1,(1,0)}; C_{v_2,(1,0)}; C_{v_1,(1,1)}; C_{v_2,(1,1)}; C_{v_1,(0,1)}; C_{v_2,(0,1)})$.
In this previous work, the fitness function is defined as the sum of three distances, each corresponding to one of the criteria associated with BK:

$$f(x) = \sum_{\eta} d_{\Delta t}(tr, BK) + d_b(tr, BK) + d_e(tr, BK) \tag{2}$$

where $d_{\Delta t}(tr, BK)$ is the distance between the expected time given by BK (Δt) and the time spent in the current state by the considered trajectory; $d_b(tr, BK)$ represents the distance between the trajectory behaviour inside the discrete state and the property of BK; and $d_e(tr, BK)$ compares the expected next discrete state according to BK with the discrete state into which the considered trajectory enters. The function domain is $\left(\prod_{v \in V} [0, b_v]\right) \times [0, 1]^n \times \mathbb{R}^{|C|}$ where n is the number of genes and $|C|$ is the total number of celerities to identify, i.e. the length of the decision vector. The codomain is $\mathbb{R}^+$.

Minimising these three criteria led to the identification of admissible celerity values. However, the optimisation problem becomes increasingly complex when considering hGRN models with many genes. It implies more celerity values to identify and more complex interactions, leading to harder implicit constraints. The continuous CSP solver was unable to extract even one particular solution when considering a model with five genes, leading to 240 variables in the decision vector. Furthermore, the task is multimodal: it is interesting to find diverse solutions to provide biologists with evidence for different interpretations of hGRN dynamics. The approach proposed by [17] did not address this issue. The peculiarities of this optimisation problem are: (i) there is an infinite number of solutions that satisfy the BK constraints, and (ii) the optima solutions lie on a neutral landscape, i.e. a plateau. Indeed, solutions form a measure zero set due to the equality constraints on the time criterion in the fitness function. Therefore, the optimisation procedure requires the ability to sample, in a continuous landscape, global and local optima plateaus of measure zero. These considerations specific to this optimisation problem cannot be addressed only by panmictic

schemes. Therefore, the limits of the mentioned approach are tested by introducing experiments with well-known multimodal heuristic algorithms on higher dimensional hGRNs.

3 RS-CMSA-ESII and cGAs for MMO

RS-CMSA-ES [1] was designated the most successful niching method for the CEC'2013 MMO test suite. In this initial version, several parallel subpopulations, each following the evolution scheme of CMSA-ES [9], aim at finding distinct global minima. CMSA-ES is an adapted version of CMA-ES [14], diminishing the complexity of the adaptation process and implying fewer hyperparameters tuning. RS-CMSA-ES gathers several techniques and encompasses them as a new algorithm for MMO without making any assumption about the fitness landscape: taboo points (points from which the offspring of a subpopulation must maintain a sufficient distance, i.e. the centre of the fitter subpopulations and the previously identified basins), the normalised Mahalanobis distance, and the Ursem's hill-valley function [23]. The new variant RS-CMSA-ESII [2] introduces an update of the adaptation schemes for the normalised taboo distances, new termination criteria for subpopulation evolution, and an improvement of the time complexity thanks to (i) a new initialisation strategy of subpopulations, and (ii) a more accurate metric for the determination of critical taboo regions thanks to the properties of Mahalanobis distance. The RS-CMSA-ESII superiority over successful niching methods in static MMOs made it an ideal candidate for this study.

cGAs are well-known methods for addressing multimodal and epistatic problems [4, 5]. They are a subclass of GAs in which the population is structured in a specified topology, allowing individuals to interact only with their neighbours. The topological structure defines a connected graph where a vertex represents an individual, and an edge represents the possibility of interaction between two individuals: each individual, in this graph, can only mate with its neighbours. Therefore, in a cGA, the choice of the population topology and the neighbourhood are two parameters that guide the search and control the solutions' diffusion speed along the graph. The *radius* introduced in [5] directs the dispersion strength based on the chosen neighbourhood: the higher the radius, the more spread out a neighbourhood's pattern is, and so the harder a good solution will reach other individuals of the population because there will be more intermediate individuals to the most distant individual. Furthermore, [19] introduced the *ratio* measure controlling the balance between exploration and exploitation. It is defined as a trade-off between the radii of the neighbourhood and the population structure: reducing the ratio leads to the promotion of exploration. Overlapping neighbourhoods also help to explore the search space because the slow diffusion of solutions through the population allows exploration by preserving diversity [3, 4]. On the one hand, this leads cGAs to find several optima compared to GAs and to be well suited for complex problems. On the other hand, this is often at the expense of slower convergence towards global optima.

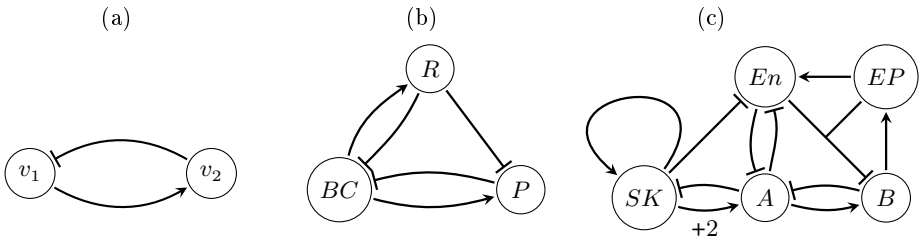

Fig. 2: Interaction graphs of the 2G (a), 3G (b), and 5G (c) hGRN.

name	Nb. genes	Decision vector len.	BK
example cycle (2G)	2	8	given in Section 2.2
circadian cycle (3G)	3	20	[7]
cell cycle (5G)	5	240	[6]

Table 1: Description of hGRN models.

In the following section, tests have been set up to compare the RS-CMSA-ESII performance along with cGAs to demonstrate which method is best suited to our multimodal task. Different structure and ratio values for cGA are experimented with to evaluate their performance. We compared all the results with standard panmictic metaheuristics on three hGRN models of increasing complexity to assess the suitability of their diversity mechanism for such MMO problems.

4 Experimental study

The three hybrid models of GRN are depicted in Figure 2 and described in Table 1 in terms of (i) the number of genes, (ii) the length of the decision vector to optimise, and (iii) constraints from BK utilised for evaluating candidate solutions.

4.1 Optimisation methods and parameters search

The comparison is carried out between (μ+λ) GA, CMA-ES, six synchronous cGAs with different ratios and neighbourhood structures, and RS-CMSA-ESII.

The two continuous metaheuristic implementations come from PyMoo [11], and each of the hyperparameters chosen is identical to those detailed in [17]. Their population size is also 500. Since we were interested in observing the influence of the cGAs parameters to find those most suitable for solving the different hGRN problems, multiple sets of parameters were tested (listed in Table 2). The names of the neighbourhoods follow the classical notation: the label Ln (linear)

name	population	neighbourhood	ratio
cGAL5	5x10	L5	0.279
cGAl9	10x10	L9	0.367
cGAL29	15x15	L29	0.719
cGAL41	21x21	L41	0.851
cGAL13	7x7x7	L13	0.607
cGAC9	7x7x7	C9	0.408

Table 2: Description of tested cGAs parameters.

for the neighbourhoods composed by the n nearest neighbours in a given axial direction (north, south, west and east) while the label Cn (compact) designates the neighbourhoods containing the $n-1$ nearer individuals to the considered one (in horizontal, vertical, and diagonal directions). The population size and the neighbourhood structure vary so that we can test (i) low ratio cGAs with a small population size and, conversely, (ii) high ratio cGAs with a larger population, both in a toroidal 2G square grid, and (iii) 3G neighbourhood structure. To ensure fair results, their implementation is also based on the standard GA implementation provided in PyMoo. RS-CMSA-ESII implementation is taken from [2] with the control parameters set to their default values.

Each experiment is run 50 times to obtain statistically significant results. The termination criteria chosen is the number of function evaluations (NFE): $100,000$ for 2G and 3G and $200,000$ for 5G. These values were chosen based on the relative complexity and the decision vector length.

4.2 Results

For each algorithm, problem dimension and at each generation, we compute the best candidate solution so far, repeat executions 50 times, and compute the Mean Best Fitness (MBF). The monotonic evolution of all algorithms is shown in the left column of Figure 3. It can be observed that (i), as expected, panmictic metaheuristics perform worse than cGAs in all cases since they reach a plateau faster and attain a higher fitness score after convergence; (ii) cGAL13, cGAL29, and cGAL41 stand out among the algorithms tested since, on the one hand, they have a slower convergence, and on the other hand, even when the maximum budget is attained, their curves show that the search process could have pursued its convergence; and (iii) RS-CMSA-ESII performs worse than CMA-ES.

In addition, Cumulative Distribution Function (CDF) curves are constructed on the right side of Figure 3 for each hGRN considered. Each CDF curve describes the probability of finding a solution at, or below, a given fitness score. For instance, in 3G, there is almost an 80% probability that a user will obtain a solution with a fitness score less than or equal to 10^{-4} with cGAL9 given

100,000 NFE. From these plots, (i) cGAs don't often find the overall best solution (the one with the lowest fitness score) but results are rarely unsatisfactory (> 1), (ii) in all cases, CMA-ES can deliver top results (satisfactory and precise solutions) as it is of poor performance (not solving the problem), (iii) RS-CMSA-ESII similarly to CMA-ES has mixed performance and does not find any single satisfactory valuation in 5G.

In MMO, the chi-square-like performance statistic and maximum peak ratio are common measures to identify a maximum number of optima (local and global). However, both of these measures assume the number and locations of the global optima are known *a priori*. This assumption does not hold in our case, so the scoring function used is introduced in [16] and defined as:

$$sc(P, \theta_l, \theta_u) = \sum_{B_j \in Bin_k(clust_\sigma(P), \theta_l, \theta_u)} w_j |B_j| \tag{3}$$

This alternative performance measure suggests the selection of a threshold interval $[\theta_l, \theta_u]$ covering all fitness score values considered interesting by an expert. θ_l is the ideal point while θ_u is an upper bound below which fitness values are judged satisfactory. In our case, $\theta_l = 0$ and $\theta_u = 10^{-2}$. 10^{-2} is a precision error

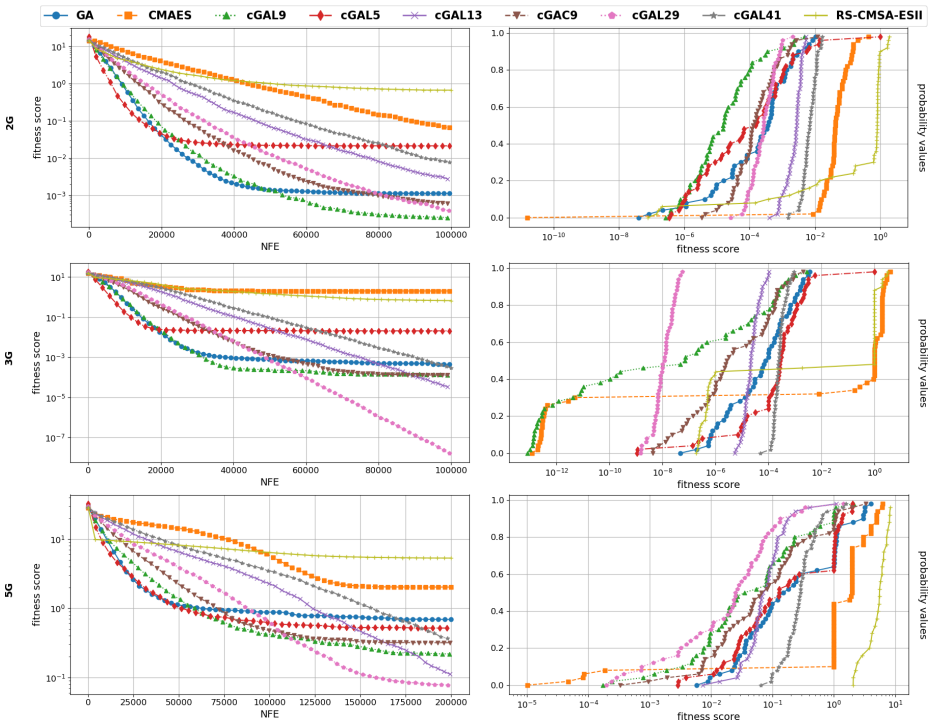

Fig. 3: Monotonic evolution of MBF values (left) and CDF curves of overall best results (right) for the three hGRNs.

coherent with biological expertise. For instance, a trajectory which would slide in a state during a fraction of seconds ($< \theta_u$) before going to the next discrete state is a satisfying trajectory despite BK stating $noslide(v)$. The score measurement uses density-based clustering with parameter σ to remove redundancy between candidate solutions clustered closely around the same local optimum. In this study, DBSCAN [13] is parametrised with $\sigma = 10^{-1}$ which is the maximum Euclidean distance between two samples for one to be considered as in the neighbourhood of the other. Equidistant binning is then used to adapt the distribution weights: more emphasis is put on higher quality optima than lower ones. The number of bins is kept at 16. This score assesses the combined quality of the found candidate solutions while it is not prone to be misled by redundancy. Table 3 shows numerical values for the mean scores where bold results highlight the best performance for each model dimension. The comparison indicates that small ratio cGAs (cGAL9 and cGAL5) are to be preferred for 2G, whereas cGAs with a higher ratio perform better in the 3 and 5G cases, as shown by cGAL41 and cGAL29. It should also be noted that, in 5G, the extrema ratio values (cGAL5 and cGAL41) are penalised for being too exploratory or exploitative. cGAC9 has interesting results in all three cases but never stands out.

Table 4 summarises statistics of the last population clustered: it contains only the fitness values of the best candidate solutions ($< \theta_u$) gathered around each distinct optima found by clustering. The best results (column by column) are shown in bold. The average of the mean and standard deviation of the clustered results is reported, as well as the overall minimum fitness scores (the reader can refer to the leftmost point of each corresponding CDF curve). When considering one particular run, it may appear that an algorithm did not find any solution below θ_u. In such cases, the maximum value θ_u is considered: this results in a normalised average with the ideal value being θ_l, and θ_u the nadir one. It can be observed that cGAL9 finds, on average, higher quality optima than other algorithms in 2 and 5G. In 3G, cGAL29 identifies satisfying solutions with a lower fitness score on average.

Algorithms	2G	3G	5G
GA	12.13	148.09	8.29
CMA-ES	3e-3	0.275	29.32
cGAL9	**19.99**	63.47	21.11
cGAL5	16.82	30.07	2.04
cGAL13	7.33	290.59	0.03
cGAC9	13.78	162.28	34.11
cGAL29	13.90	182.18	**49.81**
cGAL41	1.56	**311.07**	0.0
RSCMSAII	1e-2	0.275	0.0

Table 3: Overview of the average performance measurement over 50 runs.

Algorithms	2G		3G		5G	
	mean ± std	min	mean ± std	min	mean ± std	min
GA	1e-3 ± 2e-4	4e-8	4e-4 ± 2e-6	5e-8	99e-4 ± 94e-4	6e-3
CMA-ES	98e-4 ± 96e-4	**2e-11**	7e-3 ± 2e-13	1e-13	9e-3 ± 9e-3	**1e-5**
cGAL9	**8e-4 ± 8e-4**	3e-7	1e-4 ± 7e-6	**8e-14**	**85e-4 ± 7e-3**	2e-4
cGAL5	1e-3 ± 9e-4	3e-7	9e-4 ± 2e-4	1e-9	95e-4 ± 94e-4	3e-3
cGAL13	6e-3 ± 2e-3	4e-4	3e-4 ± 4e-4	5e-6	1e-2 ± 98e-4	7e-3
cGAC9	1e-3 ± 1e-3	3e-6	1e-4 ± 8e-6	4e-9	9e-3 ± 8e-3	3e-4
cGAL29	3e-3 ± 2e-3	3e-5	**9e-7 ± 1e-5**	1e-9	79e-4 ± 7e-3	2e-4
cGAL41	8e-3 ± 3e-3	2e-3	2e-3 ± 16e-4	5e-5	1e-2 ± 0	1e-2
RSCMSAII	8e-3 ± 1e-20	8e-8	7e-3 ± 2e-13	1e-13	1e-2 ± 0	1e-2

Table 4: Summary of clustered results.

4.3 Statistical analysis

A statistical validation campaign was conducted to evaluate the observed differences in the reported performance values of all algorithm pairs for each different hGRN. We consider two null hypotheses H_0^1 which states that the observed performance scores are equal, and H_0^2 which states that the average fitness scores obtained by clustering are similar. These null hypotheses are duplicated for each of the hGRN dimensions considered. To test them, we first employed the Friedman rank-sum test to assess whether at least two methods exhibit significant differences. The p-values for the null hypotheses show, at a α = 5% confidence level, that the differences are significant. The choice between parametric and non-parametric tests is made according to the independence of the samples (seeds are different), whether or not the data samples are normally distributed, and the homoscedasticity of the variances [12]. As neither normality nor homoscedasticity conditions required for the parametric tests application hold, the non-parametric Wilcoxon signed-rank test was performed. In a complementary way, to reduce the issue of Type I errors in multiple comparisons, the Bonferroni correction method was applied. [15] gives the score +1 (resp. -1) for the superior (resp. inferior) algorithm whenever the considered null hypothesis could be significantly rejected. A score of 0 is assigned when neither algorithm is significantly better than the other. Since we have three different case studies (2G, 3G, 5G), for each pair of algorithms and each null hypothesis, we sum the three obtained scores to estimate which one is globally better considering the three hGRNs. Table 5 (resp. Table 6) show these sums according to the pairwise Wilcoxon tests (resp. Bonferroni correction): a positive number for algorithm in line l shows that it was significantly better than the algorithm in column c (considering the three hGRNs). For example, according to the Bonferroni correction applied on H_0^1, we can state that cGAL29 is significantly better than RS-CMSA-ESII for the three study cases but compared to cGAL41, we can only say that it is globally better: cGAL29 may have scored +2 and cGAL41 +1 or cGAL29 may have scored +1 and cGAL41 0.

	CMA-ES		cGAL9		cGAL5		cGAL13		cGAC9		cGAL29		cGAL41		RSCMSAII	
GA	+2	+2	0	−2	0	+1	0	+1	−2	−1	−2	−1	0	+2	+2	+2
CMA-ES			−2	−2	−2	−2	−1	−1	−2	−2	−2	−2	−1	−1	0	+1
cGAL9					+3	+2	+1	+3	0	+1	0	0	+1	+3	+3	+3
cGAL5							0	0	−1	0	−1	−1	0	+2	+2	+2
cGAL13									−1	−3	−1	−3	0	+2	+2	+2
cGAC9											−1	−1	+1	+3	+3	+3
cGAL29													+1	+3	+3	+3
cGAL41															+2	+1

Table 5: Pairwise Wilcoxon statistical tests of H_0^1 (left) and H_0^2 (right).

	CMA-ES		cGAL9		cGAL5		cGAL13		cGAC9		cGAL29		cGAL41		RSCMSAII	
GA	+2	+2	0	**0**	**+1**	**0**	0	+1	**0**	**0**	**−1**	**0**	0	+2	+2	**+1**
CMA-ES			−2	−2	−2	−2	**−2**	**−2**	−2	−2	−2	−2	**−2**	**−2**	0	**0**
cGAL9					**+1**	**+1**	**0**	**+2**	0	**0**	0	**+1**	**0**	**+2**	**+2**	**+2**
cGAL5							0	**+1**	−1	**−1**	−1	**0**	0	+2	+2	**+1**
cGAL13									**0**	**−2**	−1	−3	0	+2	+2	**0**
cGAC9											**0**	**0**	**0**	**+2**	**+2**	**+2**
cGAL29													+1	+3	+3	+3
cGAL41															+2	**0**

Table 6: Bonferroni post-hoc analysis of H_0^1 (left) and H_0^2 (right) with bolded differences compared to Table 5.

If we analyse the conclusions supported by the tests, based on the acceptance or rejection of the above hypotheses, we arrive at the following findings: on the different tasks, cGAL9 and cGAL29 are more competitive in finding more optima than other algorithms with better fitness values on average. RS-CMSA-ESII lags as the panmictic algorithms maintain greater diversity in their population across different hGRN landscapes.

4.4 Visualisation

Figure 4 shows the diversity of solutions of cGAL9 tested on hGRNs with 2, 3 and 5 genes. Please note that three different graph types are modelled to emphasize the same phenomenon: the evolution of gene products concentration. In 2G (Figure 4a) and 3G (Figure 4b), the discrete states can be represented as squares and cubes. However, in 5G (Figure 4c), the choice has been made to represent the evolution of concentration (in y-axis) as a function of the time spent (in x-axis) for the different genes. This visually confirms that the application of evolutionary computation allows us to exhibit very different solutions, each consistent with BK.

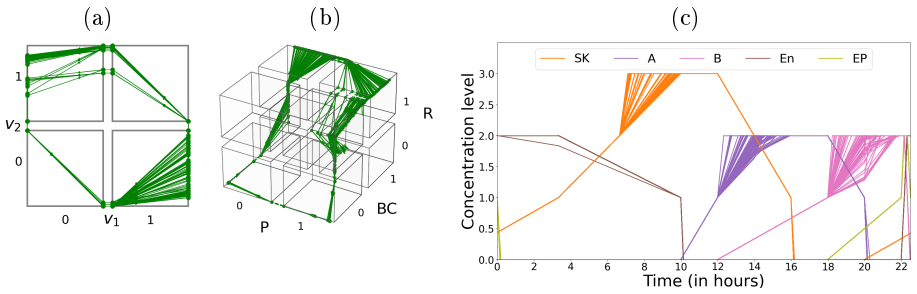

Fig. 4: Admissible trajectories obtained with cGAL9 on the 2G (a), 3G (b), and 5G (c) hGRN.

5 Conclusion

hGRN variable identification is framed as an ideal tool to help biologists develop hypotheses and facilitate the design of their experiments. This study proposes an improvement to [17] since (i) it shows that evolutionary computation can outperform constraint-based approach by dealing with higher dimensional models, the 5G cell cycle in this study, and (ii) it is now able to find a diverse set of optima solutions instead of a unique one. CGAs have shown superiority over the best available niching-based algorithm (RS-CMSA-ESII) by maintaining diversity within the population structure. Surprisingly, RS-CMSA-ESII does not ensure diversity in the results: only one solution is found. In our case, optima are located on a neutral landscape: there is an infinite number of solutions forming a null set. Therefore, for sampling a continuous landscape with global and local optima plateaus of measure zero, the mechanisms employed by RS-CMSA-ESII are not suitable. Because the Ursem's hill-valley test fails, it ensures that only one subpopulation at a time evolves, leading to a single solution. That entails the degenerate use of the metaheuristic, explaining the disappointing results of RS-CMSA-ESII. In the case of cGAs, maintaining diversity through population structure helps to preserve diversity in the parameter space and thus enables us to obtain a diversity in the phenotype space. Future works will consider the development of specific diversity mechanisms to better leverage the multimodality issue on a neutral landscape: the design of an appropriate *self-adaptive* cGA to obtain quality results while maximising the number of optima. At the same time, introducing larger biological systems will lead to applying *large-scale* optimisation.

Acknowledgments. This work has been supported by the French government, through the France 2030 investment plan managed by the Agence Nationale de la Recherche, as part of the "UCA DS4H" project, reference ANR-17-EURE-0004.

References

1. Ahrari, A., Deb, K., Preuss, M.: Multimodal optimization by covariance matrix self-adaptation evolution strategy with repelling subpopulations. Evolutionary Computation (2017). https://doi.org/10.1162/evco_a_00182
2. Ahrari, A., Elsayed, S., Sarker, R., Essam, D., Coello, C.A.C.: Static and dynamic multimodal optimization by improved covariance matrix self-adaptation evolution strategy with repelling subpopulations. IEEE Transactions on Evolutionary Computation (2021). https://doi.org/10.1109/TEVC.2021.3117116
3. Alba, E., Dorronsoro, B.: Solving the vehicle routing problem by using cellular genetic algorithms. In: European Conference on Evolutionary Computation in Combinatorial Optimization (2004). https://doi.org/10.1007/978-3-540-24652-7_2
4. Alba, E., Dorronsoro, B.: Introduction to cellular genetic algorithms. In: Cellular Genetic Algorithms (2008). https://doi.org/10.1007/978-0-387-77610-1_1
5. Alba, E., Troya, J.M.: Cellular evolutionary algorithms: Evaluating the influence of ratio. In: International Conference on PPSN (2000). https://doi.org/10.1007/3-540-45356-3_3
6. Behaegel, J., Comet, J.P., Bernot, G., Cornillon, E., Delaunay, F.: A hybrid model of cell cycle in mammals. In: 6th International Conference on Computational Systems-Biology and Bioinformatics (2015). https://doi.org/10.1142/S0219720016400011
7. Behaegel, J., Comet, J.P., Folschette, F.: Constraint identification using modified Hoare logic on hybrid models of gene networks. In: Proceedings of the 24th Int. Symposium TIME (2017). https://doi.org/10.4230/LIPIcs.TIME.2017.5
8. Behaegel, J., Comet, J.P., Pelleau, M.: Identification of dynamic parameters for gene networks. In: Proceedings of the 30th IEEE Int. Conf. ICTAI (2018). https://doi.org/10.1109/ICTAI.2018.00028
9. Beyer, H.G., Sendhoff, B.: Covariance matrix adaptation revisited - the cmsa evolution strategy -. In: International Conference on PPSN (2008). https://doi.org/10.1007/978-3-540-87700-4_13
10. Biswas, S., Acharyya, S.: Neural model of gene regulatory network: a survey on supportive meta-heuristics. Theory in Biosciences (2016). https://doi.org/10.1007/s12064-016-0224-z
11. Blank, J., Deb, K.: pymoo: Multi-objective optimization in python. IEEE Access (2020)
12. Eftimov, T., Korošec, P.: Statistical Analyses for Meta-Heuristic Stochastic Optimization Algorithms: GECCO Tutorial (2020). https://doi.org/10.1145/3377929.3389881
13. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: kdd (1996)
14. Hansen, N., Auger, A.: Cma-es: evolution strategies and covariance matrix adaptation. In: Proceedings of the 13th annual conference companion on Genetic and evolutionary computation (2011). https://doi.org/10.1145/2001858.2002123
15. Kronfeld, M., Dräger, A., Aschoff, M., Zell, A.: On the benefits of multimodal optimization for metabolic network modeling. In: German conference on bioinformatics (2009)
16. Kronfeld, M., Zell, A.: Towards scalability in niching methods. In: IEEE CEC (2010). https://doi.org/10.1109/CEC.2010.5585916
17. Michelucci, R., Comet, J.P., Pallez, D.: Evolutionary continuous optimization of hybrid gene regulatory networks. In: EA 2022. https://doi.org/10.1007/978-3-031-42616-2_12

18. Mitra, S., Biswas, S., Acharyya, S.: Application of meta-heuristics on reconstructing gene regulatory network: A bayesian model approach. IETE Journal of Research (2021). https://doi.org/10.1080/03772063.2021.1946433
19. Sarma, J., De Jong, K.A., et al.: An analysis of local selection algorithms in a spatially structured evolutionary algorithm. In: ICGA. pp. 181–187. Citeseer (1997)
20. da Silva, J.E.H., Betnardino, H.S., Helio J.C., B., Vieira, A.B., Luciana C.D., C., de Oliveira, I.L.: Inferring gene regulatory network models from time-series data using metaheuristics. In: IEEE CEC (2020). https://doi.org/10.1109/CEC48606.2020.9185572
21. Sun, J., Garibaldi, J., Hodgman, C.: Parameter estimation using meta-heuristics in systems biology: A comprehensive review. IEEE/ACM Trans. Comput. Biology Bioinform. (2012). https://doi.org/10.1109/TCBB.2011.63
22. Tenazinha, N., Vinga, S.: A survey on methods for modeling and analyzing integrated biological networks. IEEE/ACM Trans. Comput. Biol. Bioinform. (2011). https://doi.org/10.1109/TCBB.2010.117
23. Ursem, R.K.: Multinational evolutionary algorithms. In: Proceedings of CEC (1999). https://doi.org/10.1109/CEC.1999.785470