MEMÓRIAS
DA
ACADEMIA DAS CIÊNCIAS
DE
LISBOA

CLASSE DE CIÊNCIAS

TOMO XLV

From Charles Darwin to evolutionary genetic algorithms

RUI DILÃO

ACADEMIA DAS CIÊNCIAS
DE LISBOA

LISBOA • 2018

From Charles Darwin to evolutionary genetic algorithms

Rui Dilão

Abstract

The concepts of evolution and natural selection escaped from the context of biology and entered mathematics, physics and computer science. These concepts inspired powerful optimization methods, enabling the analysis and calibration of complex mathematical models of all kinds of phenomena and depending of hundreds of parameters. These mathematical models often describe evolutionary processes in biology but are also associated with mathematical applications of difficult engineering problems. These new class of methods are called evolutionary genetic algorithms and are one of the best available techniques to analyze ill posed problems in general associated with irregular data sets.

INTRODUCTION

The theory of evolution, synthesized by Charles Darwin in 1859, [Dar], introduced an ensemble of arguments that explains the evolution, the variability and the environmental adaptation of all life forms.

Life forms evolve and adapt in an ever-changing environment and are very robust to these changes (resilience). In the turn of the nineteenth to the twenty century it has been observed that life is anchored in some form of "genetic" information contained in organisms, and occasional or random events occurring at the molecular level have an important role in evolution and adaptation.

One of the mechanisms that has an important role in biology, in physics and in engineering is randomness. Evolution occurs due to random environmental forces and random molecular interactions that occur in nature. Then, selective mechanisms, in some cases associated with environmental conditions as temperature, humidity, etc., choose the fittest. In 1953 Miller [Mil] showed with an experiment that the building blocks of life, the aminoacids, spontaneously form in an aquatic environment subjects to random extreme events (heat and electric discharges). More recently the same observations where corroborated by analyzing the chemical composition of volcanic steams [Joh].

To better understand the role of randomness in evolution, we present a mathematical model aiming to describe the first steps of the embryonic development of the fruit fly *Drosophila melanogaster*. This mechanism is shared by all insects and arthropods and, as we shall see, give some clues on the molecular processes that are in the origin of the complexity and evolution of organisms. This mathematical model takes the form of a mixed system of ordinary and partial differential equations depending on a large number of unknown parameters that are very difficult if not impossible to measure. The mathematical model is based on the assumption of randomness of the forces occurring at the molecular level.

Then, we calibrate the parameters of the model with the available experimental data, making predictions about the behavior of the biological system. For the calibration technique, we developed an *evolutionary genetic algorithms* enabling to fit experimental data (noisy) of very complex systems to mathematical models with an arbitrarily large number of parameters. These evolutionary genetic algorithms are based on the random choice of parameter configurations with subsequent selective procedures. The biological problem analyzed here and the calibration techniques were both inspired by the Darwin evolutionary reasoning.

The success of a mathematical model in the description of developmental processes depends on the possibility of finding the set of realistic parameters that best fits the dynamic characteristics of the systems under analysis. The dynamic properties of the mathematical model have a predictive nature and enable to interfere with the biological system under controlled conditions.

THE DARWIN LEGACY

The ideas of evolution, natural selection, common descent and struggle for life are commonly associated with the name of Charles Darwin. In fact, in 1809, Lamark first used the concept of evolution to discuss the altering of a species into another (transmutation of species). Latter, in 1844, Chambers wrote: "new species were produced from previous ones in a progressive sequence leading to humans". In 1803, Malthus introduced the concept of struggle for life in the context of the dynamics of tribal groups. For Darwin, it was the competition inside the species and its relations with the environment that leads to the "survival of the fittest". In 1850, Spencer, a philosopher of sciences, introduced the generic expression "struggle for life".

The concept of natural descent appeared for the fist time in 1932 in the Darwin notebooks in the form of a diagram called the tree of life. In this diagram, it is suggested that different organisms could have evolved from a common ancestor. For the general public, this idea appeared for the first time in print in 1858 by Wallace.

The contribution of Darwin was to frame all these concepts together and to give a reasonable explanation of evolution, funded in observations [Dar]. In the Galápagos islands, Darwin found new species that have a common ancestor with other continental species. Darwin argued that natural selection exists in regions of the globe where geographic barriers were somehow imposed, and geographic migrations from a common ancestor are in the origin of the differentiation between individuals, leading to speciation. It was necessary to wait almost 100 years to make the first experiments in laboratory on evolution and natural selection (Luria-Delbrük experiment, 1943, [Lur]), showing that evolution and selection are separated and independent processes, confirming the Darwinian point of view.

THE FORMATION OF *DROSOPHILA* SEGMENTS

One of the characteristics of arthropods (insects, spiders, centipedes, shrimps, trilobites, etc.) is the formation, during the first hours of development, of protein segments along the embryo. These segments are bands formed by proteins that are expressed along the antero-posterior axis of the embryo. In Figure 1, we show the band structure of some of the proteins of the embryo of *Drosophila*.

The number of protein segments changes from species to species and it is believed that this feature is inherited from some ancestral arthropod. The importance of these segments, formed during the first hours of development, is due to the fact that their positions determine the body plans of the adult larvae at a later stage of development, Figure 2.

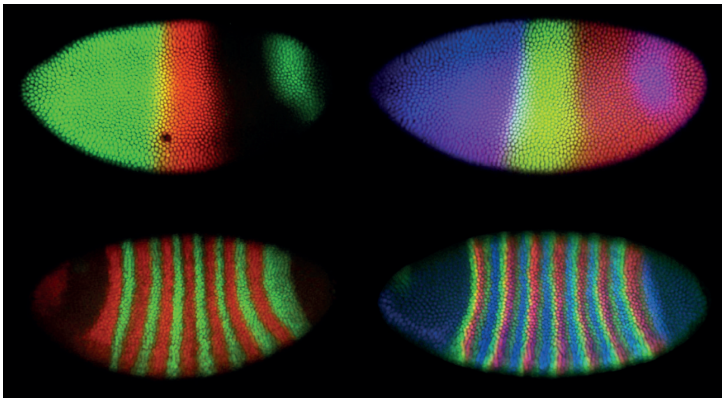

Figure 1
Protein segments in the embryo of the fruit fly *Drosophila melanogaster*. (Data from the FlyEx database). On the two top figures we see the segments of the gap-gene family of proteins. On the two bottom figures, we show the segments of the pair-rule and the segment-polarity families of proteins. These bands form during the first two hours after fertilization of the *Drosophila* egg. In *Drosophila*, the pair-rule family shows 7 bands and, for example, in some centipedes this family develops 9 bands. The small dots in the figure show the localization of nucleus and the different color are the markers of different proteins. At this phase of development, there are no cellular membranes and the nucleus of the embryo are not isolated from each other by the usual cellular membrane. This structure is called a syncytium. In animals, during development the first stages of development, there is also a syncytial phase.

The biological events leading to the formation of the protein patterning in *Drosophila* are the following ([Dri] and [Nüs]): In the Drosophila egg, some mRNAs of maternal origin (*bicoid, nanos, caudal, hunchback, etc.*) are placed near the poles of the oocyte by the mother's ovary cells, defining the antero-posterior axis of the embryo. For example, the regions of deposition of *bicoid* and *nanos* mRNAs determines the regions of the embryo where the head and the posterior organs of the larva will latter be developed. Initially, the oocyte has only one nucleus, but fertilization triggers nuclear duplication by mitosis without the formation of cellular membranes (syncytium), Figure 1. The formed nucleus distribute near the internal wall of the embryo. Simultaneously, fertilization triggers the translation of the deposited maternal mRNAs to proteins. These proteins are localized near the external nuclear membranes of the recently formed nucleus.

After fertilization, the first 13 nuclear divisions occur without the organization of cellular membranes around the nucleus (the embryo is a syncytium), giving rise to the syncytial blastoderm. The cytoplasmic membranes only become completely formed three hours after fertilization, in the interphase following the 14th mitotic cycle, just before the onset of gastrulation.

During the syncytial stage, the transcribed zygotic genes are divided in three main families: gap, pair-rule and segment-polarity genes. The proteins resulting from their expression define broad segmentation patterns along the antero-posterior axis of the embryo as shown in the two first pictures of Figure 1. The proteins with origin in the maternal mRNAs form (steady) gradients along the antero-posterior axis of the embryo. In the beginning of cleavage cycle 14, proteins of maternal origin act as transcription factors (activators or inhibitors) for gap-genes, pair-rule and segment polarity genes.

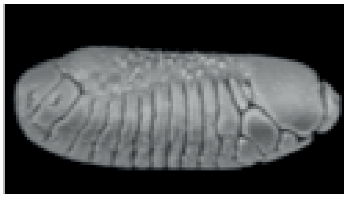

Figure 2
Segmentation of the *Drosophila* larvae. These segments begin to appear 7 hour after fertilization and are completely formed 15 hours after fertilization. Their positions correspond to the localization of the expressed pair-rule proteins, expressed during the first two hours of development (Figure 1). This suggests that the body plan of *Drosophila* is established very early during development.

There are several models aiming to describe proteins steady gradients in *Drosophila* early development. Some models are based on the hypothesis of protein diffusion along the antero-posterior axis of the embryo, [Hou, Alv2]. Other models are based on the diffusion of mRNA of maternal origin, [Dil1, Dil2, Dil3, Dil4]. The protein diffusion hypothesis is sometimes justified by the absence of cellular membranes during the first 14 cleavage cycles of the embryo, and has been proposed by Nüsslein-Volhard and co-workers in the late eighties, [Dri]. However, this hypothesis implies several characteristics for the embryo and for the protein dynamics that have never been observed. The mRNA diffusion hypothesis is supported by the recent observation of the gradient of the bicoid mRNA of maternal origin, [Spi]. In order to decide which is the mechanism that best describe the segmental patterning as shown in Figure 1, Dilão and co-workers introduced several mathematical models aiming to the describe the patterns of Figure 1. These models are based on the hypothesis of diffusion of the maternal origin mRNA, together with localized protein production without diffusion, [Dil1, Dil2, Dil3, Dil4].

In order to test the pattern formation mechanisms in *Drosophila*, some details are necessary in the description of models. Optimally, these details should include measurements of protein and mRNA formation rates, degradation rates and diffusion coefficients. Also direct measurements of protein concentrations would be desirable. However, the values of these parameters change from embryo to embryo and it is not technically possible to measure simultaneously in the same embryo all these parameters.

Most of the biological information on patterning is in the form of the fluorescent images as shown in Figure 1. The relationship between different proteins, mRNAs and genes are purely qualitative. For example, in Figure 3, we show, the biological information of some of the known interaction between the proteins of maternal origin, the mRNAs and the gap-genes, in Drosophila early development.

In the double graph of Figure 28, the green-solid represent activation interactions and the red arrows inhibitory mechanisms. The symbols associated with the vertices of this double-graph represent a particular substance into one of its forms: gene, mRNA or protein. A particular formalism has been developed in order to transform graphs describing genetic regulatory networks into a mathematical model, [Alv1] and [Dil3].

For the particular case of the activation of Bicoid protein (BCD) by bicoid mRNA (bcd), represented by the sub-graph with two vertices at the top left of Figure 3, we have assumed that bicoid mRNA diffuses along the antero-posterior axis of the embryo. The mathematical model associated to this sub-graph is described by the partial differential equation ([Dil2]),

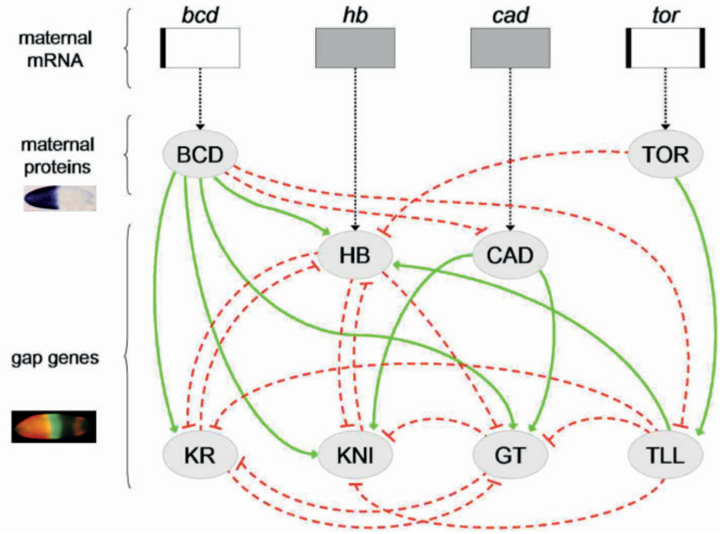

FIGURE 3
Double graph of interactions describing the genetic network associated with the maternal and gap-gene phase of Drosophila early development. Solid-arrows represent activation of a gene by a protein and the red-dotted arrows show inhibitory interactions. Black arrows represent translation of a protein from the corresponding mRNA (Figure from [Alv2]).

$$\frac{\partial R}{\partial t} = -dR + D\frac{\partial^2 R}{\partial x^2}$$

$$\frac{\partial B}{\partial t} = aR$$

where $R(x, t)$ and $B(x, t)$ are the concentration of bcd and BCD, respectively. These concentrations are defined along the antero-posterior axis of the embryo ($x \in [0,L]$), and we assume zero-flux boundary conditions. The embryo length is $L = 0.5$ mm. In the beginning of the developmental process there is the deposition of bcd at some (unknown) region of the antero-posterior axis of the embryo and this concentration of bcd ($R(x, 0)$) is also unknown. This deposition region is the interval $[I_1,I_2]$, somewhere along the embryo with length L. In this model, the Bicoid protein does not diffuse, it is translated by the ribosomal machinery that exists outside the nuclear membranes. The diffusion of bicoid mRNA is driven by Brownian motion (molecular chaos) eventually along microtubules distributed along the embryo, [Dil2].

The model equations presented above with the boundary and initial conditions have seven free parameters, but theoretical analysis shows that it is only possible to determine four independent parameters, ([Dil2]). Anyhow, all the parameters are unknown and the experimental data is in the form of fluorescent signal proportional to protein concentration. The technique to calibrate the model equations with the experimental data is to use evolutionary genetic algorithms — essentially we must choose the set of parameters that best fit the theoretical model. If the fit is good, we believe that the model is sufficiently accurate to make predictions about the system.

In Figure 4, we show the fit of the model solutions with the experimental data. The model predictions are

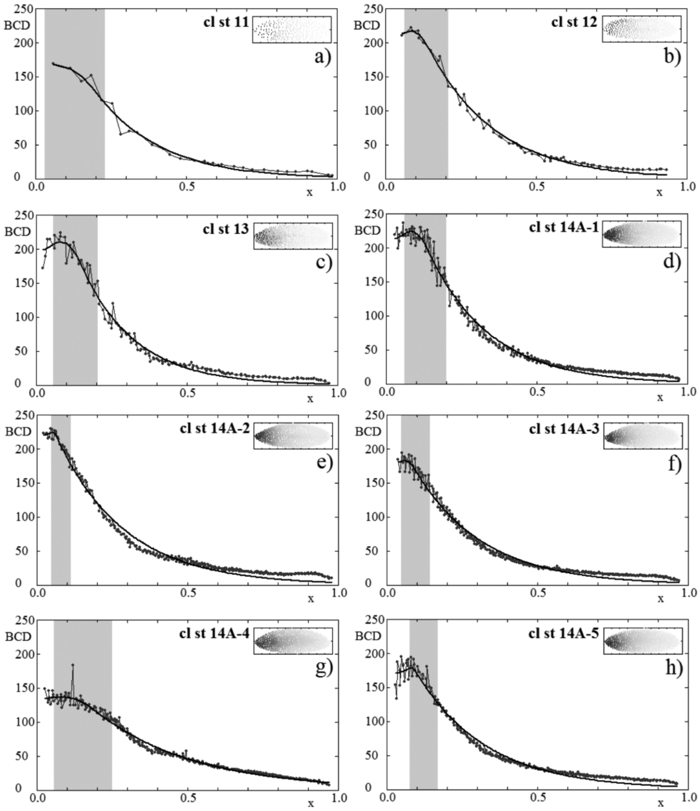

FIGURE 4
In blue, we show the experimental distribution of BCD (Bicoid protein) along the antero-posterior axis of the embryo of Drosophila, for different embryos in different stages of development. The small two-dimensional distribution within each frame is the two-dimensional gradient of protein BCD. The blue graph is the antero-posterior distribution taken along the middle axis of the embryo. The black line is the best steady solution of the model equations and the grey region is the region where bcd mRNA were initial deposited by the mother ovary cells. Both the grey regions and the model parameters were determined with an evolutionary genetic algorithm, [Dil1]. The agreement between the diffusion model and the experimental data is very good, with relative errors in the range 5-8%. This Figure is reprinted from [Dil2].

denoted by the black lines and are the steady-state solution of the partial differential equation shown above. Each graph represents the distribution of the protein Bicoid measured in vitro and in a different embryo. Each picture corresponds to a different choice of the parameters. We don't list the choice of parameters but in the following we describe the evolutionary genetic algorithm approach used to determine the parameter values.

EVOLUTIONARY GENETIC ALGORITHMS

The determination of the parameters that best fit the noisy data (blue lines) shown in Figure 4 is called an optimization problem. In this more technical context, we introduce now some concepts borrowed from biology. The reason for this is due to the fact that we have no guesses about parameters values and the data is not continuous, preventing us from using smoothing techniques. This is a situation very similar to the natural mechanisms leading to biological evolution: there are no preferential choices for the generation of diversity and the selection acts only *a posteriori*.

As we have to choose the set of parameters that characterize our particular problem, we call a population to the compact set of all the possible parameter values. In the model equations described above, we have seven parameters, and therefore the population is represented by a compact subset of R^7. An individual of a population is a choice of a point of this set. Within this set of points, we can reproduce, mutate, swarm, etc., provided the new or modified parameter values remain in the compact set. So the question is to know which individual (set of parameters) best fits the experimental data.

A genetic algorithm is a procedure to randomly choose individuals of the population that best fit the model with the experimental data. The selection of the best candidates is done by a criterion that can be, for example, the minimization of a least square deviation function. This procedure to be effective must be implemented in such a way that convergence to a best local or global solution is achieved.

The simpler form of a genetic evolutionary algorithm is the Covariance Matrix Adaptation Evolutionary Strategy, developed by Hansen, [Han]. The implementation of this algorithm follows the following steps:

1 – Choose an initial individual (the Darwinian ancestral) p_0 of the population (the full set of parameter values for the model) and let $C = I_n$ be a covariance matrix, where I_n is the $n \times n$ identity matrix.

2 – From the multivariate Gaussian distribution with covariance matrix C and mean value p_0, sample λ offsprings. For each offspring or set of parameter values calculate the solution of the model equations and then calculate the fitness function, a chi-squared distributions, for example. From the best μ (μ<λ) offsprings, according to the fitness function, recalculate a new mean value p_0 and a new (unbiased estimator of the) covariance matrix C.

3 – Repeat these steps several times. After several iterations, the best individual ever found is a candidate for the best choice of the model parameters.

In general, the genetic algorithm just described has a fast convergence for a local or global optimal solution. The fits of figure 4 have been obtained with a similar technique.

This algorithm admits generalizations for the cases where there are conflicting goals, [Dil1, Dil4], and the solutions of the optimization problem is not unique. In the context of multiple goals, we have a multi-objective or Pareto type optimization problem and evolutionary genetic algorithms are a good choice for an unbiased search of realistic solutions of multi-parameter models. For example, in Figure 5, we show, several (non-unique) Pareto optimal solutions for the distribution of proteins Bicoid and Caudal in the embryo of *Drosophila*. In each picture, we show a different instance of the Pareto optimal solution of the model equations for the same experimental data set. In this case, the model equations do not minimize simultaneously the two experimental data sets. This is a characteristics of multi-objective optimization techniques where non-unicity is the rule. Obviously, this is a feature of all the living organisms where individuals of the same species are characterized by different sets of parameters (Pareto optimal), with different performances relative to different properties.

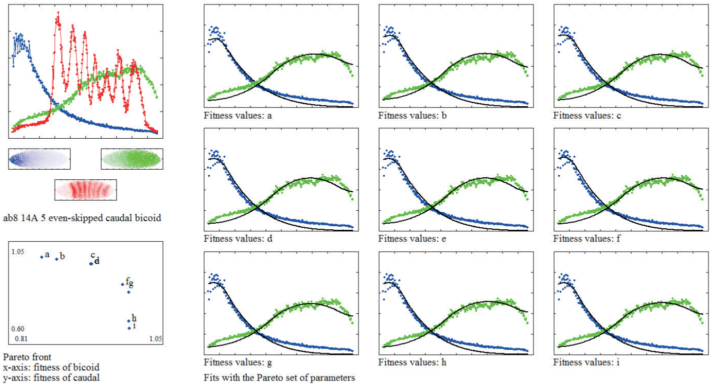

FIGURE 5
In the picture on the top-left we show the concentration of the proteins Bicoid (blue), Caudal (green) and Even-skipped (red) along the antero-posterior axis of the embryo of *Drosophila*. On the nine figures on right we show Pareto bi-objective optimal solutions of the model equations for the same experimental data, [Dil1]. Each picture corresponds to a different set of parameter values on the Pareto front of this bi-objective optimization problem. In the lower left we show the Pareto front (set of optimal solutions) on the Bicoid-Caudal fitness space. The parameter values have obtained with a evolutionary genetic algorithm for multi-objective optimization problems, [Dil1].

CONCLUSIONS

Evolutionary genetic algorithms are a class of probabilistic algorithms that can be used to search for solutions of optimization problems, without having any *a priori* knowledge about the solutions or about its regularity properties. These algorithms mimic the evolution-selection duality reasoning as proposed by Darwin. Here, we have shown how to use evolutionary genetic algorithm ideas to solve uni-objective and bi-objective Pareto type optimization problems from developmental biology. On the other hand, the biological problems presented here have its roots in the work of Darwin.

REFERENCES

[Alv1]- F. Alves and R. Dilão, A simple framework to describe the regulation of gene expression in prokaryotes, *Comptes Rendus Biologies*, **328** (2005) 429-444.

[Alv2]- F. Alves and R. Dilão, Modeling segmental patterning in Drosophila: Maternal and gap genes, *J. Theor. Biol.* 241 (2006) 342-359.

[Dar]- C. Darwin, On the Origin of Species by means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life, London: John Murray, 1859.

[Dil1]- R. Dilão, D. Muraro, M. Nicolau and M. Schoenauer, Validation of a morphogenesis model of Drosophila early development by a multi-objective evolutionary optimization algorithm. In: C. Pizzuti, M. D. Ritchie and M. Giacobini (Eds.): EvoBIO 2009, Lec. Notes in Computer Science, vol. 5483 (2009) 176-190.

[Dil2]- R. Dilão and D. Muraro (2010), mRNA diffusion explains protein gradients in Drosophila early development, *J. Theor. Biol.* 264 (2010) 847-853.

[Dil3]- R. Dilão, Muraro D (2010), mRNA diffusion explains protein gradients in Drosophila early development, *PLoS ONE*, **5** (5) (2010) 1-10 (e10743).

[Dil4]- R. Dilão and D. Muraro, Calibration and validation of a genetic regulatory network model describing the production of the protein Hunchback in Drosophila early Development, *Comptes Rendus Biologies*, 2010, in press.

[Dri]- W. Driever, C. Nüsslein-Volhard, A gradient of Bicoid protein in Drosophila embryos, *Cell* 54 (1988) 83-93.

[Han]- N. Hansen, The CMA Evolution Strategy: A Tutorial, 2008.

[Hou]- B. Houchmandzadeh, E. Wieschaus and S. Leibler, Establishment of developmental precision and proportions in the early Drosophila embryo, *Nature* 415 (2002) 798-802.

[Joh]- A. P. Johnson, H. J. Cleaves, J. P. Dworkin, D. P. Glavin, A. Lazcano, J. L. Bada, The Miller Volcanic Spark Discharge Experiment. *Science* 322 (2008) 404.

[Lur]- S. E. Luria and M. Delbrück, Mutations of Bacteria from Virus Sensitivity to Virus Resistance, *Genetics* 28 (6) (1943) 491–511.

[Mil]- S. L. Miller, A Production of Amino Acids Under Possible Primitive Earth Conditions, *Science* 117 (1953) 528-528.

[Nüs]- C. Nüsslein-Volhard, Coming to life. How Genes Drive Development. Yale University Press, New Haven, 2006.

[Spi]- A. Spirov, K. Fahmy, M. Schneider, E. Frei, M. Nooll and S. Baumgartner, Formation of the bicoid morphogen gradient: an mRNA gradient dictates the protein gradient, D*evelopment* 136 (2009) 605-614.

(Comunicação apresentada à Classe de Ciências
na sessão de 7 de maio de 2009)