CelloS: a Multi-level Approach to Evolutionary Dynamics

Camille Stephan-Otto Attolini[1], Peter F. Stadler[12], and Christoph Flamm[1]

[1] Institut für Theoretische Chemie,
Universität Wien, Währingerstraße 17, A-1090 Wien, Austria
{camille,studla,xtof}@tbi.univie.ac.at
[2] Lehrstuhl für Bioinformatik, Institut für Informatik, Universität Leipzig,
Kreuzstraße 7b, D-04103 Leipzig, Germany.
{camille,studla}@bioinf.uni-leipzig.de

Abstract. We study the evolution of simple cells that are equipped with a genome, a rudimentary gene regulation network at transcription level and two classes of functional genes: motion effectors allow the cell to move in response to nutrient gradients while nutrient importers are required to actually feed from the environment. The model is inspired by the protist *Naegleria gruberi* which can switch between a feeding and dividing amoeboid state and a mobile flagellate state depending on environmental conditions. Simulation results demonstrate how selection in a variable environment affects the gene number and efficiency so that the cells can rapidly switch from one expression regime to the other depending on the external conditions.
Keywords: Artificial Cells, gene regulation, evolution, *Naegleria gruberi*

1 Introduction

A non-trivial task in Artificial Life research is to devise genotype-phenotype maps, i.e., relations between genomic sequence information and the shape, structure, and behavior of the organism that is encoded by the genome. The difficulties stem from the complexity of even the simplest cells, which precludes a representation of an entire cell at the molecular level. On the other hand, at present there are no established "intermediate-level" theories that would provide consistent but simplified representations of cellular processes (energy metabolism, biomass production, cell division, sensory responses, intracellular transport, gene expression, etc.). One therefore has to resort either to simulations based on a large number of *ad hoc* assumptions, or to the construction of minimal models based on biophysical and biochemical principles.

The process of RNA folding, for example, can be viewed as a minimal model of a genotype-phenotype map. Here, the sequence of the RNA molecule acts as the genotype (the sequence information is actually heritable in *in vitro* selection (SELEX) experiments [16]), while the (secondary) structure of the molecule is interpreted as the phenotype (SELEX experiments indeed often demonstrate a

strong structure dependence of the selected nucleic acids). Detailed investigations of the RNA model lead to the development of important concepts, such as neutral networks percolating sequence space, the phenomenon of shape space covering, and the importance of accessibility for phenotypic evolution [20, 7]. The structure of the genotype-phenotype map determines the structure of the fitness landscape [21] which in turn determines the dynamics of an evolving population. The high degree of neutrality of the RNA folding map, for example, explains punctuated equilibria in the absence of external events [8, 13], leads to a selection for robustness against mutations [22] and influences evolvability [4].

Concepts such as epistasis and phenotypic plasticity easily translates into this RNA folding metaphor [6], however, important characteristics of the genotype-phenotype maps of biological organisms, do not have a counterpart in this framework:
While genotype and phenotpye are embodied in the same physical entity in the RNA model, there is a rather strict separation between genomic information and functional molecules in all biological organisms. This allows an organism to exist in different internal states (that depend on its *individual* history) which may cope with environmental conditions in different ways. Regulatory networks are at the core of the mechanism by which cells individually adapt to changing conditions, see e.g. [9, 3]. The majority of the artificial gene regulation models used today [1, 5, 11, 19] are based on the well established "operon model" of gene expression [14], which divides the genes into two classes: (i) the transcription factors capable of binding to the DNA thereby modulating the expression of downstream located genes; and, (ii) structural proteins which perform some functions different from the regulation of the gene expression. In the simplest case, regulatory networks arise when transcription factors also enhance or inhibit the expression of other transcription factors. (Note that such models still ignore crucial regulation mechanisms of real cells such as signal transduction networks and post-transcriptional gene silencing.)

The `CelloS` model described in this contribution combines a simple computational cell model, the extended Potts model (see [18] and references therein), with an artificial genome and a minimal model of gene expression [19]. This combination allows us to study the coupling of the environmental dynamics to the cell internal dynamics of gene expression within the framework of an evolving cell population.

Our approach is motivated by the cell differentiation of the amoeba *Naegleria gruberi*, which is capable of changing cell shape, from a crawling amoeba to an asymmetric elongated cell, and of growing flagella when nutrients are scarce. It has been shown [10] that all proteins necessary for the differentiation are synthesized *de novo*, i.e., due to transcriptional regulation. The initiation of morphological changes require the synthesis of sufficient amounts of proteins, i.e., a significant investment. The transformation is temporal and the organism returns back to the amoeba state when nutrients are again available. *N. gruberi* divides in the amoeba state only, while the flagellate state is much more mobile and hence better suited to explore novel nutrient sources.

2 The model

The basic tool for our simulations is the Potts model with some extensions [17] on a 2D lattice. A *cell* C is a maximal connected subset of the lattice such that all lattice points in C have the same type or "color" u. Cells interact with each other with strength J_{uv} at neighboring lattice points depending on their types u and v. A special type 0 denotes empty lattice sites. Each cell is characterized by its energy

$$E_C = \sum_{i \in \partial C} \sum_{j \in N(i) \setminus C} J_{u_C, u_j} + \lambda(\mathrm{vol}(C) - V)^2 \tag{1}$$

where $\mathrm{vol}(C)$ is the volume of the cell, i.e., its number of lattices points, V is a user-defined target volume, and λ is a compressibility parameter. The double sum runs over all lattice edges that point from the boundary (surface) of the cell C to other cells or into the environment. The enviroment contains a nutrient with concentration c_i at lattice point i.

Cell motion is implemented by a simple Metropolis Monte Carlo step in which a cell attemps to modify its boundary at lattice point $i \in \partial C$ by changing the type of an adjacent site i' to its own type, or by changing one of its boundary sites to 0. The cells feel the gradient in the nutrient by evaluating $c_{i'} - c_i$. The transition probability is $\min\{1, \exp(-(\Delta E_C + \mu_0(c_{i'} - ci) + H_\partial)/T)\}$ where H_∂ is the energy cost of deforming the cell's boundary, μ_0 describes the reactivity of the cell to changes in the nutrient concentration, and T a temperature-like parameter. Note that cell motions are internally driven and hence consume energy rather than the result of molecular Brownian motion. Our cells have a finite life expectancy and require energy to stay alive. This is modeled by a "battery" that is used up when enzymes are synthesized or the cell moves. When the "battery" is empty, the cell dies and the corresponding lattice sites are reset to 0.

Each cell on the lattice contains an RNA sequence of length 1000 that represents its genome and contains the information necessary to decode the cell's behavior. This genome can encode two types of effector molecules (corresponding of course to proteins in *N. gruberi*, but modeled as RNAs here for computational convenience) and a simple regulation mechanism. The "genetics" of the `CelloS` model is summarized in Fig. 1.

A short signal sequence (corresponding e.g. to the TATA box in real cells) marks the beginning of a "coding region" on the genomic sequence. We use the signal `GC` and define a gene the following 40 nucleotides. This subsequence is folded into its secondary structure using the `RNAfold` program of the `Vienna RNA Package` [12]. The structure is then compared with two target shapes for the "motion effectors" and the "nutrient importers", which are kept fixed throughout the simulation. The closer target shape determines the function of the gene, while the number of base pairing differences measures the gene's efficiency.

In the current implementation we keep the gene regulation network fixed. In order to implement the switching between the motion effectors and nutrient importers we use the simple negative feedback system shown in Fig. 1 (bottom).

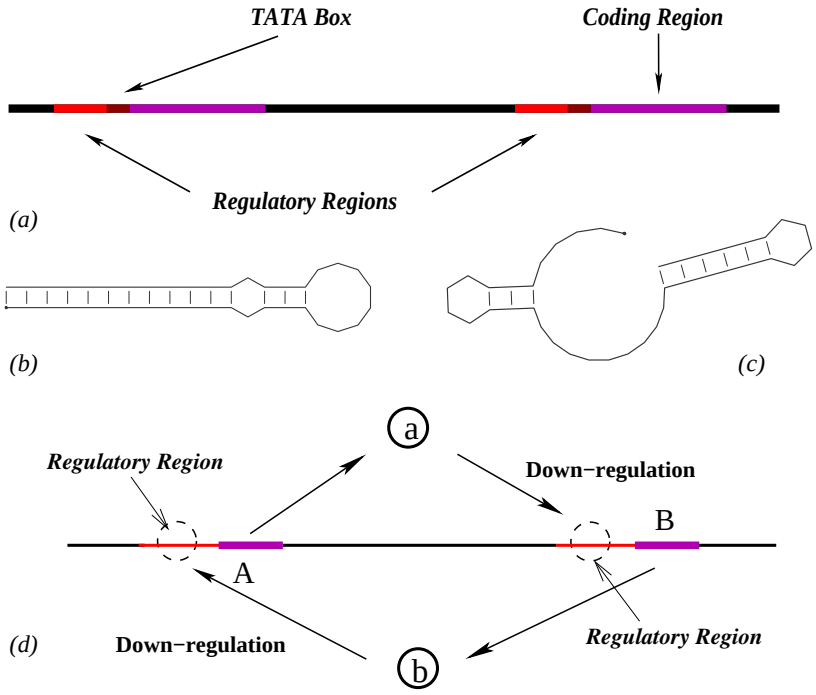

Fig. 1. Genetics of the `Cellos` model. (a) Genomic organization. Two classes of functionally different RNAs are distiguished by archetypic shapes: motion effectors (b) and metabolic effectors that act as nutrient importers (c). Panel (d) summarizes the logic of regulation in `Cellos`: expression is down-regulated when an RNA from the other function-class binds to the regulatory region of the gene.

The differential equations for this scheme are:

$$\begin{aligned} \frac{\partial G_A}{\partial t} &= \gamma_A \cdot k \frac{1}{1+G_B^3} - d \cdot G_A \\ \frac{\partial G_B}{\partial t} &= \gamma_B \cdot k \frac{1}{1+G_A^3} - d \cdot G_B \end{aligned} \tag{2}$$

where G_A and G_B are the concentrations of the two types of gene products, γ_A and γ_B are their efficiencies, and k and d fixed constants. A 4th order Runge-Kutta method is used to numerically integrate these differential equations.

Once the genome is decoded, the concentrations of the gene products are computed. The cell is then able to feed depending on the available nutrient in the environment provided it expresses nutrient importers, and to move if motion efforts are expressed. The battery level B is decreased depending on the gene products that are produced and it is recharged if the cell is in a food source:

$$B' = B - c_0(G_A + G_B) + \phi_0 G_B \tag{3}$$

The parameteres c_0 and ϕ_0 describe the ratio of nutrients obtained from the environment against the cost of producing the importers and motion effectors, respectively. The mobility of the cell depends on the concentration of expressed motion effectors which is reflected in a modified transition probability for changing the cells boundery by replacing the constant μ_0 with $\mu_0 \cdot G_B$.

The products of metabolic genes play two different roles: first, they recharge the battery of the cell; and second, they increase the cell's target volume. Once a cell has doubled its normal size, it divides by fission copying its genome to the new cell. This process is usually inaccurate, producing mutations in the new RNA string. In this model every replication implies one random point mutation in the genome.

Food sources are depleted when cells feed from them. Once a source is empty, it is replaced by a new one in a randomly chosen spot of the lattice. This way, cells are forced to switch between the metabolic and movement states, reinforcing the selection of only those capable of doing so.

Individual cells with very similar genomes belong to the same *species*. The definition of species in our model is similar to that proposed by Kenneth and Risto in [15]. Each gene in the population has a unique historical number. Every time a mutation creates a new one or changes the type of an old gene, this global variable is increased and assigned to the new gene. In order to compare two genomes, we use a linear combination of the number of excess (T) and disjoint (D) genes, and the average efficiency difference between common genes (W). If the result of

$$\delta = \frac{c_1 T}{N} + \frac{c_2 D}{N} + c_3 \cdot W \tag{4}$$

is below a threshold value, the new cell is assigned to the same species as the old one. Whenever a new species is created, a genome is set to represent the whole species. Every time a new cell is born, its genes are compared to all species' genomes and included in the first one for which the distance is below the threshold.

3 Simulation Results

Some fixed parameters are used in all simulations: we use a lattice of 200×200 sites with periodic boundary conditions, $J_{x,0} = 11$ for the contact with an empty site, $J_{ab} = 37.5$ for the contact between different cell types, and $J_{aa} = 35$ for the contact with a cell of the same species. Furthermore $T = 3$, $H_\partial = 0.8$, $\mu_0 = 5000$, $c_0 = 0.4$, $V = 30$, $\lambda = 5$.

Figure 2 shows the evolution of the system for two different simulation runs. In the first four images three food sources were available for the cells to eat. Population size changes depending on the conditions. The last two images were produced with only one food source.

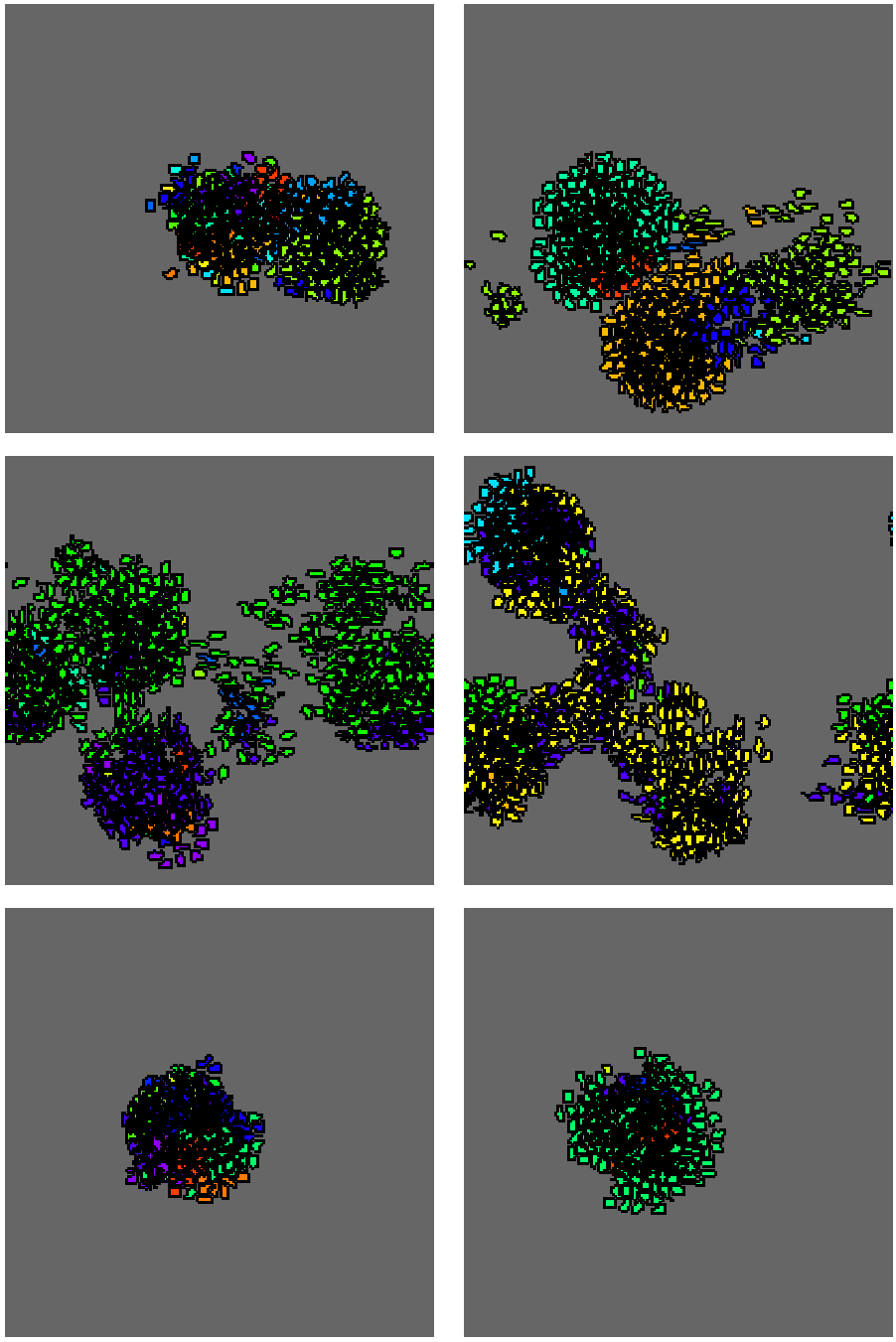

Fig. 2. First four images for three food sources. Last two with only one source.

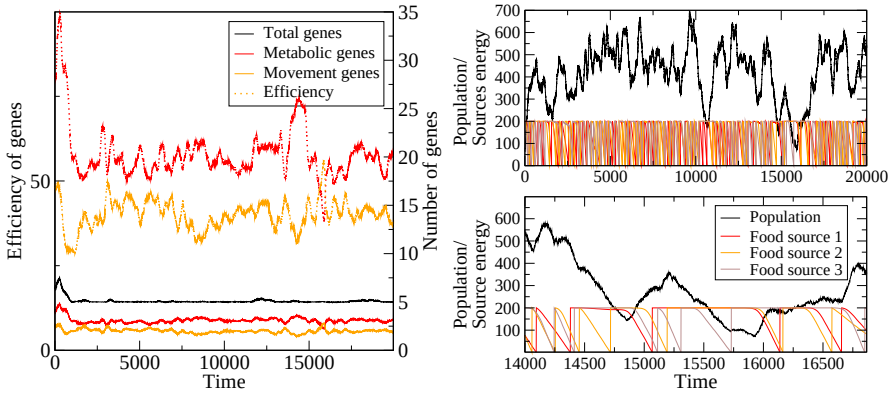

Fig. 3. Left: Number and efficiency of genes in a simulation with parameters: number of food spots 3, mean life 600, volume increment per generation 0.6. Right: Population and energy of the sources. The zoom in the bottom shows how population grows only when cells are feeding from the sources.

3.1 Genome structure

We measure the impact of the external conditions in the genome by looking at the number of metabolic and movement genes, their efficiencies and the effectors expression inside and outside a food source.

The regulatory network we are using, imposes a well defined range in which gene efficiency must lay in order to obtain the necessary switch between states. In our simulations it is clear how these numbers are controlled by natural selection when the genome is mutating randomly. In Fig. 3 it can be seen how after a period of adjustment, the population falls in a regime where gene number and efficiency are inside a small interval for both kind of genes.

The population grows depending on the availability of nutrients. Every time a food source is depleted, cells must migrate to the next one. This periods are usually reflected in a diminution of the population and increase in the average number of movement genes in it. The second panel in Fig. 3 shows the energy of the sources and the change in the number of cells. Source energy staying at its maximum means that there are no cells feeding from it. This is clearly related with a decrease of the population size (Fig. 3).

This combination of gene types allows a switching in their products expression depending only in the presence or absence of food from the environment. Figure 4 shows this behavior for a single cell with the right number of genes.

One special case of study, is when there is only one food spot of infinite life in the lattice. Cells that are in the spot are thrown out of it by the newborns. Even when there is no need of traveling long distances, the fact that cells have to be constantly coming back into the source makes the presence of movement genes indispensable. At the same time, since food is easily available, there is no need to increase the efficiency of metabolic genes. Battery may be refilled

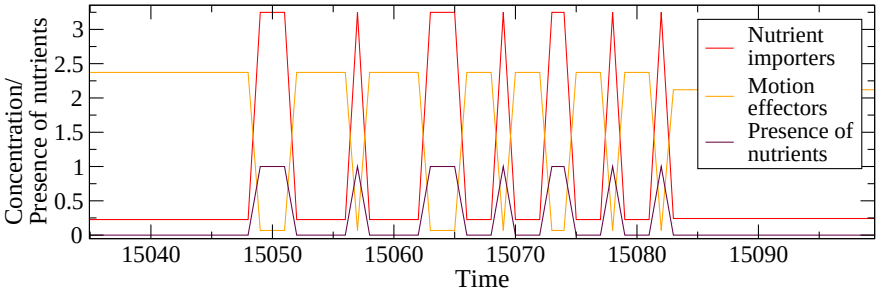

Fig. 4. Switching of gene products expression depending on the presence/absence of nutrients. Same parameters as in the previous figure.

slowly without killing the cell since the time they spent outside the food source is usually very short.

3.2 Phylogenetics

With our simple definition of species the number of species depends directly on the volume increase per generation. Phylogenetic trees can be recorded based on the speciation events, see Fig. 5 for a characteristic example. The Darwinian evolution is dominated by one or a few species at any given point in time. The coexistence of distinct lineages over longer times is comparably rare. In some runs one of the initial species survives until the end of the run, failing to find any important improvement in phenotype via mutations.

4 Concluding Remarks

In this first (and very simple) implementation of the model, we observe the response of the genome to variable environmental conditions. After an initial phase of selection the number of genes stays approximately constant. The cells can then use their gene regulation network to cope with environmental changes. Population dynamics also reflect the presence or absence of nutrients, together with an increase of the number and/or the efficiency of movement genes. We found that, at least in our simple environment, it is not important to have a large number of genes, but to have the right amount of them depending on the environmental inputs and the regulatory network modifying their products' expression.

Since the mechanism of the regulation of gene expression in the current implementation of the `CelloS` model can itself not be a target of evolution, we plan to add transcription factors as a third class of gene products to the artificial genome. This will allow the cells to find innovative regulatory strategies based on post transcriptional interaction. A fruitfull route will then be to study the mixing of regulatory strategies under sexuall reproduction of the cells.

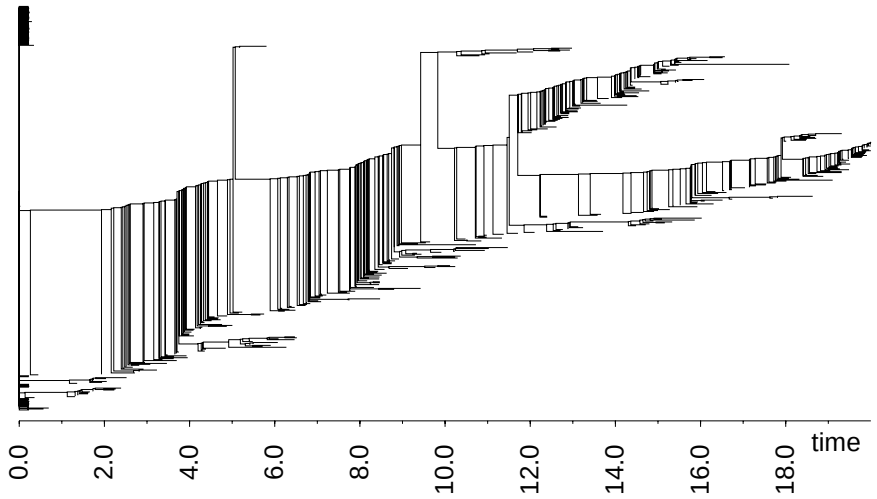

Fig. 5. Phylogenetic tree for a run with three food spots. Nodes in the tree represent the dissapearance of a species, while saddles stand for the split of two of them. Time unit is 1000 simulation steps.

Extending the set of mutation operators from point mutation to gene duplication and horizontal gene transfer, turns `Cellos` into a tool for generating test data for phylogenetic reconstruction methods. Comparing the simulated evolutionary scenario with the reconstructed one will allow to evaluate the performance of such methods.

The environmental dynamics can also be improved by switching to an artifical chemistry like the Toy Chemistry Model [2]. This forces for an additional decoding layer in the internal structure of the cells, which links our representation of the nutrient importers to organic molecules in the environment. Improvements of the `CelloS` model along these lines are under way.

Acknowledgments

This work was supported in part by Consejo Nacional de Ciencia y Tecnología, CONACyT, the DFG Bioinformatics Initiative (BIZ-6/1-2), and COST Action D27. We thank the Aegean Sea for its stimulating effect on this work.

References

1. W. Banzhaf. On the dynamics of an artificial regulatory network. In W. Banzhaf, T. Christaller, P. Dittrich, J. T. Kim, and J. Ziegler, editors, *Advances in Artificial Life*, volume 2801 of *LNCS*, pages 217–227, Heidelberg, Germany, 2003. Springer-Verlag. Proc. ECAL03.
2. G. Benkö, C. Flamm, and P. F. Stadler. A graph-based toy model of chemistry. *J. Chem. Inf. Comput. Sci.*, 43:1085–1093, 2003.

3. A. Deckard and H. M. Sauro. Preliminary studies on the in silico evolution of biochemical networks. *ChemBioChem*, 5:1423–1431, 2004.
4. M. Ebner, M. Shackleton, and R. Shipman. How neutral networks influence evolvability. *Complex.*, 7(2):19–33, 2001.
5. P. Eggenberg. Evolving morphologies of simulated 3D organisms based on differential gene expression. In P. Husbands and I. Harvey, editors, *Proc. ECAL97*, pages 205–213. The MIT Press/Bradford Books, 1997.
6. W. Fontana. Modelling 'evo-devo'with RNA. *BioEssays*, 24:1164–1177, 2002.
7. W. Fontana and P. Schuster. Shaping space: The possible and the attainable in RNA genotype-phenotype mapping. *J. Theor. Biol.*, 194:491–515, 1998.
8. C. V. Forst, C. M. Reidys, and J. Weber. Evolutionary dynamics and optimization: Neutral Networks as model-landscape for RNA secondary-structure foldinglandscapes. In F. Morán, A. Moreno, J. Merelo, and P. Chacón, editors, *Advances in Artificial Life*, volume 929 of *LNAI*, pages 128–147. ECAL95, Springer, 1995.
9. P. François and V. Hakim. Design of genetic networks with specified functions by evolution *in silico*. *Proc. Natl. Acad. Sci. USA*, 101(2):580–585, 2004.
10. C. Fulton and C. Walsh. Cell differentiation and flagellar elongation in *Naegleria gruberi*. *J. Cell Biol.*, 85:346–360, 1980.
11. N. Geard and J. Wiles. Structure and dynamics of a gene network model. In R. Sarker, R. Reynolds, H. Abbass, K. C. Tan, B. McKay, D. Essam, and T. Gedeon, editors, *Proc. CEC2003*, pages 199–206. IEEE Press, 2003.
12. I. L. Hofacker. Vienna RNA secondary structure server. *Nucl. Acids Res.*, 31:3429–3431, 2003.
13. M. A. Huynen, P. F. Stadler, and W. Fontana. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. (USA)*, 93:397–401, 1996.
14. F. Jacob and J. Monod. On the regulation of gene activity. *Cold Spring Harbor Symp. Quant. Biol.*, 26:193–211, 1961.
15. S. O. Kenneth and M. Risto. Efficient Reinforcement Learning through Evolving Neural Network Topologies. *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO-2002, 2002.
16. S. Klug and M. Famulok. All you wanted to know about SELEX. *Mol. Biol. Reports*, 20:97–107, 1994.
17. A. F. Marée and P. Hogeweg. Modelling *Dictyostelium discoideum* Morphogenesis: the Culmination. *Bull. Math. Biol*, 64:327–353, 2002.
18. R. M. H. Merks and J. A. Glazier. A cell-centered approach to developmental biology. *Physica A*, 2005. in press.
19. T. Reil. Dynamics of gene expression in an artificial genome – inplications for biological and artificial ontogeny. In D. Floreano, J.-D. Nicoud, and F. Mondada, editors, *Proc. ECAL99*, volume 1674 of *Lecture Notes in Computer Science*, pages 457–466, Berlin, 1999. Springer-Verlag.
20. P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: A case study in RNA secondary structures. *Proc. Roy. Soc. Lond.*, B225:279–284, 1994.
21. P. F. Stadler. Fitness landscapes arising from the sequence-structure maps of biopolymers. *J. Mol. Struct. (THEOCHEM)*, 463:7–19, 1999. Santa Fe Institute Preprint 97-11-082.
22. E. van Nimwegen, J. P. Crutchfield, and M. A. Huynen. Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci. USA*, 96:9716–9720, 1999.