Discrete Developmental Genetic Regulatory Networks for the Evolution of Cooperation

Moritz Buck and Chrystopher Nehaniv
Adaptive Systems Research Group, School of Computer Science, University of Hertfordshire
College Lane, Hatfield, Herts, AL10 9AB, U.K.

Abstract

Studying the transition between single-celled organisms and multicellular ones is vital for the understanding of the complexification of life on earth. During that major transition in evolution, the notion of individual changed, as the unit of selection, hence of fitness, switched from the single cell to the higher level multicellular organism.

One characteristic which is believed to be closely related to this transition is the emergence of cooperation in colonies of closely related microorganisms.

We developed during our research a new type of Boolean Artificial Developmental Genetic Regulatory Network, which is used to control artificial agents which are evolved in a two-level fitness driven genetic algorithm. This evolution is used to study the particular trade-offs in the transition to higher level individuality during the emergence of cooperation in cell colonies as against competition with the selfish tendencies of their simpler constituent cells, as well as the potential for emergence of higher level cooperation despite the competition between two levels of selection in evolutionary dynamics.

Motivation

Cooperation between different cells is a necessary property for the emergence of complex multicellular living entities. Without any signalling or environmental gradients, differentiation does not seem possible. In the first microbial mats and colonies, communication was a key element for the cooperation which eventually led to the first multicellular organisms (Michod, 2005). But how and under which conditions did such cooperation arise?

The major transitions in evolution are described as those evolutionary events where increase in complexity is discontinuous (Maynard Smith and Szathmáry, 1995). Some of these major transitions include a transfer of fitness from one level of selection to another. The fitness of an individual is the measure of its reproductive success. This is not usually directly measurable, but it can be approximated by diverse measures such as reproductive rate, chances of survival, reproductive rate of offspring, etc. However all these measures are mere approximations; many more factors enter into natural fitness, reflecting the capacity to adapt to the environment, as well as the environment and ecosystem themselves.

When the early single-celled organisms started to aggregate and cooperate the unit of selection had to change: no longer the single cell but the agglomerate, the colony was the individual (Margulis, 1982). The appropriate criterion of fitness, which had been the reproductive success of the single cell, had to be supplanted by the fitness of the colony itself. This was probably helped by kinship selection: cells which where close in space had a high probability of being very similar genetically, therefore fitness was linked to the survival of that specific genetic strain. The evolution of cooperation makes this transfer of fitness possible. But in a situation where selection was mostly at the level of a single cell, how did the switch from the selection at the cell level to the selection at a colony level arise?, and how could the higher level organism defend themselves against defection and competition at lower levels? The key conceptual breakthrough in understanding such transitions to acknowledge that within a population of potential constituent cells of any higher level individual, Darwinian evolution is operating (Buss, 1987). Hence it is appropriate to investigate the trade-off between heritability of fitness at the different levels (Michod, 1999).

In biology, genetic regulatory networks (GRNs) are responsible for a large part of the complexity of living systems. The genetic material in the cells does not only code for the proteins but also encodes the network that directs when and how these cellular tools are to be produced. These follow the rules of cellular chemistry in turn to interact with the genetic material and to regulate protein biosynthesis, or interact with the environment (including other cells), or may serve to actively respond to and regulate cell- and organism internal activity and changes (Arthur, 2000).

In this article, a new discrete GRN model will be used to study the evolution of cooperation in multicellular systems evolving under the constraint of a two-level fitness function, the two levels representing two different levels of selection.

Methodology

The general set-up is a genetic algorithm modelling the evolution of cooperating colonies of cells. The individuals will be colonies of genetically identical cells controlled by a new discrete developmental genetic regulatory network model. The evolution will be driven by a two-level fitness, where for a given GRN, one level is linked to individualistic behaviour of cells and one to colony-level organization.
GRN model

In biology GRNs are very important. In a simplified view, the genome is divided into genes, each gene has two parts – one part encoding the product of that gene, a protein (proteins are the work horses of the cellular machinery), and a regulatory section. The regulatory part controls when the protein is to be produced. Some proteins, or protein complexes, bind to the regulatory sites in the regulatory section to increase, reduce, start, or stop protein production by the gene (Davidson, 2001; Schilstra and Bolouri, 2002; Watson et al., 2003).

Our GRN model is inspired by the biological model (Watson et al., 2003; Davidson, 2001). The genome of the cells is an array of genes; each of those genes produces a certain protein and is regulated according to its regulatory region. This GRN model is similar to the GRN model used in other diverse work (Buck and Nehaniv, 2006; Knabe et al., 2006; Quick et al., 2003), but rather than allowing continuous values, has been discretized here. A main point about this new model is that it is boolean; there are no protein concentrations or numbers: either the protein is present in the cell or not, and either the protein is being produced or it is not. This discrete version is an instantiation of S. Kauffman's boolean network models of genetic regulatory control (Kauffman, 1969) extended to the multicellular case. Moreover, we study evolved rather than random networks.

The regulatory part of the gene is composed of a certain number of cis-sites, which are in turn composed of a certain number of binding sites. Each binding site has a boolean value assigned to it, true if the protein associated with that specific binding site is present in the cell, and false otherwise. They are two types of cis-sites, activatory or inhibitory. The boolean value assigned to each cis-site is the logical AND of the boolean values assigned to each of its binding sites, and this value is negated (logical NOT) if the cis-site is inhibitory. The expression value of a gene is the logical OR of all the values of its cis-sites. A gene can also be of two different types: default on or default off. A “default on” gene produces its protein if its expression value is true, a “default off” one if its expression value is false. If more than one gene could potentially produce the same protein, that protein is produced if one or more of these genes produce the protein. (This is the same as if those genes were one “hyper-gene” with the combination of their regulatory parts with the “default off” genes negated.)

Formally, we have for a protein binding site \( i \), the binding value \( b_i \),

\[
b_i = \begin{cases} 
\text{true} & \text{if binding protein is present} \\
\text{false} & \text{if binding protein is not present}
\end{cases}
\]

the expression value \( c_j \) of a cis-site \( j \),

\[
c_j = \begin{cases} 
\bigwedge_{i} b_i & \text{if } j \text{ is activatory} \\
\neg \bigwedge_{i} b_i & \text{if } j \text{ is inhibitory}
\end{cases}
\]

where the conjunction is over all binding sites within the cis-site; and the final protein production \( p_k \) of the gene \( k \),

\[
p_k = \begin{cases} 
\bigvee_{j} c_j & \text{if } k \text{ is default off} \\
\neg \bigvee_{j} c_j & \text{if } k \text{ is default on}
\end{cases}
\]

where the disjunction is over all cis-sites within the gene.

Using the fact that any boolean function can be written in disjunction normal form, it can be shown that this system is complete in the sense of combinatorial logic: given a boolean vector of size \( n \) (the vector of the presence/absence of \( n \) proteins in the cell) there always exists at least one network computing every one of the \( 2^n \) possible boolean functions. It is not completely clear whether the analogues of this can be proved for previous GRN models. This model has also some other interesting characteristics. It is quite robust at least in principle; for example, if one gene is duplicated, then the function represented by the network is not altered, which is not the case for networks in most continuous GRN models (Buck and Nehaniv, 2006; Knabe et al., 2006). This features is very interesting in biological models where duplication of genes has an important role to play (Ohno, 1970). Also this model, even if discrete, is relatively biologically plausible. In biology, expression of genes is seldom continuous, proteins have different levels of expression, or at least the effect of the expression is discontinuous. Biological switches are omnipresent in biological systems, especially notably in developmental ones (Arthur, 2000; Davidson, 2001)

Encoding

These genetic regulatory networks are encoded in a linear genome as the basis of heredity for “biologically inspired” evolution. The genome is a variable length boolean array.
These different visible states are independent of the states are controlled by another protein; if it is present the cell is in that state; if it is not, the cell is in one of two “cooperative” states. These states are controlled by another protein (i.e. by setting its value to one representing a low level single cell goal. Since these goals are exclusive, they are in competition. The fitness here has the particularity of actually being two fitness functions representing two levels of selection, one trying to reach a high level goal requiring cooperation and one representing a low level single cell goal. Since these goals are exclusive, they are in competition.

The lower level fitness for a cell is simply to stay as long as possible in the “individualistic” state. We check for each cell in the grid the fraction of simulation time that it has stayed in that state. If \( t_{\text{ind}}(i) \) is the time cell \( i \) has spent in the “individualistic” state, the “individualistic” fitness \( F_{\text{ind}} \)
of a GRN in a certain simulation is then

\[ F_{\text{ind}} = \frac{\max_{t=1}^{t_{\text{sim}}} t_{\text{ind}}(i)}{t_{\text{sim}}}, \]

where \( t_{\text{sim}} \) is the length of a simulation.

The higher level goal is to create a checkerboard pattern with the “red” and “green” cells. At each time step of a simulation, for each cell of the grid in a “cooperative” state we check its neighbourhood locally for appropriate checkerboard patterning: for each of the neighbouring cells which is in a different cooperative (but not individualistic) state, the cell gets a score of 0.25 (remark: 0.25 is 1 divided by the number of neighbours 4). So at each time step each cell can get a score between 0 and 1. Those scores are then summed for each time step over all cells and normalized to 1. Formally, let \( n_i(j, t) \) be equal to 0.25 if the \( j^{th} \) neighbour of cell \( i \) is in the opposite cooperative state to cell \( i \) at time \( t \), else 0. Then \( f_{\text{group}}(i) \), the fitness of cell \( i \), is

\[ f_{\text{group}}(i) = \left\{ \begin{array}{ll} \frac{1}{t_{\text{sim}}} \sum_{t=1}^{t_{\text{sim}}} \sum_{j=1}^{n_{\text{cells}}} n_i(j, t) & \text{if } i \text{ cooperative} \\ 0 & \text{if } i \text{ individualistic} \end{array} \right. \]

hence the higher level fitness of the GRN after a simulation \( F_{\text{group}} \) is the average of \( f_{\text{group}} \) over the colony

\[ F_{\text{group}} = \frac{1}{n_{\text{cells}}} \sum_{i=1}^{n_{\text{cells}}} f_{\text{group}}(i), \]

where \( n_{\text{cells}} \) is the total number of cells in the grid.

The final fitness of a GRN will be the maximum between the higher level and the lower level fitness weighted by \( \alpha \in [0, 1] \), a parameter weighting the advantage of being individualistic. So the fitness \( F \) of a GRN lies in the interval \([0, 1]\) and is

\[ F = \max(F_{\text{group}}, \alpha \cdot F_{\text{ind}}). \]

### Experimental Investigation

In a first series of experiments, we verify whether this kind of set-up can evolve the kind of networks we wish to get: cells cooperating to synchronize their states so as to set up a checkerboard pattern. So a series of 10 GAs are run (mutation rate: 0.002, cross-over rate: 0.5, population size: 125, starting genome size: 1000, generations: 500, tournament size: 25, size of the grid: \( 6 \times 6 \), length of simulation: 30 time steps) with the \( \alpha \) parameter set to zero, meaning that there is no contribution of to fitness from the individualistic fitness, the evolution is only driven by the high level fitness. This might make it easier to evolve the wanted cooperation.

Parallel to this series of experiment we compute 10 series of fitnesses for 500 (number of generation) times 125 (the size of the population) randomly generated genomes — which is equivalent to the number of genomes computed in a run of the GA, so to compare in a qualitative manner the effect of using evolution.

After that series of experiments, a second set of experiments is launched with increasing \( \alpha \), so as to assess the effects of reducing the advantage of cooperation: 10 GAs for each value of \( \alpha \) (the other parameters are kept to the same) are run. We will study (1) the maximum fitness reached by these evolutionary runs, and (2) the proportion of these runs having reached a maximum fitness higher than \( \alpha \) (meaning that the GRNs evolved have reached a certain level of cooperation above that available to colonies of non-cooperative individualistic cells).

### Results

In the first set of experiments we managed to achieve the expected result. GRNs able to set up the desired checkerboard patterns evolved. Different kind of behaviours are evolved in different runs, with better or worse checkerboard patterns, stable checkerboards, blinking ones or unstable ones. The way the colonies use information is also quite variable: some colonies communicate the whole way through a simulation, e.g. one type of cooperating cell communicated and the other did not so that the cells always “know” what state they have to be in; or cells communicate when they change state; while other colonies only use communication in the beginning of the simulation to set up a checkerboard pattern and then stop all communication (Fig. 6).

If we look at the fitness values reached by the experiments with \( \alpha = 0 \) (Fig. 3 and 4), we can also see that the evolved genomes achieve better results than the randomly generated ones. The evolved ones reach values of about 0.8 with the better ones above 0.9, while the randomly generated ones only reach values of around 0.6 up to 0.7.

For the second set of experiments, with the varying \( \alpha \), the first remark is that, whatever the value of alpha is, evolutionary runs which manage to reach cooperating states reach a similar quantitative level of cooperation (Fig. 4). This means...
Figure 4: Fitness vs. $\alpha$ for those simulations which evolved group cooperation. Red circles: mean fitness of all experiments with a given $\alpha$ that achieved multicellular cooperation; vertical bars: 95% confidence interval; horizontal maroon line and gray lines: mean fitness of random population and zone of 95% confidence.

Figure 5: Frequency of achieving cooperation vs. Weighting $\alpha$ in favour of individualistic behaviour. Circles: Measured proportion of evolutionary simulations which did converge to a multicellular state for a given level $\alpha$ of individualistic weighting; line: estimation of the probability function for achieving the transition to multicellularity.

Figure 6: Snapshots of some simulations. Left in each pair (a, b & c): the statuses of the cells, here the cells are either in cooperative “red” or “green” state, or in blue for the “individualistic” state. The whole multicellular organism is rewarded for building a checkerboard-pattern. Right in each pair (a, b & c): cells which are “communicating” with their neighbours are coloured in blue depending on the “amount” of communication.
that in this model there is a sort of clear switch between non-cooperative and cooperative colonial organization, rather than a gradual change in degree depending on the particular advantages of individualistic behaviour. This idea of a clear switch is also evident in Fig. 5; this figure represents an approximation (logistic regression, chi-squared value: $3.186 \times 10^{-19}$) of the probability of achieving a cooperative state (operationally defined by the criterion of “fitness higher than $\alpha$”, which is justified as we have seen in Fig. 4). Evolutionary runs which evolve cooperation all have similar fitnesses, so every evolution which reaches a fitness score higher than $\alpha$ at a point in its history will end with cooperative colonies sooner or later. The probability drops relatively fast when $\alpha$ gets higher (further experiments to be published elsewhere give evidence that the slope can get much steeper if the size of the population of the GA increases); this kind of graph is characteristic of phase transitions, which would indicate that there would appear to be a value of $\alpha$ around which the “choice” of evolution changes sharply from simple non-cooperative colonies, easy to achieve and to maintain, to more complicated cooperative, early multicellular higher level organisms.

Discussion
Research into the complexification of living systems has been for a long time in the realm of biologists (Arthur, 2000; Buss, 1987; Maynard Smith and Szathmáry, 1995; Bonner, 1988) and later on a field studied by some theoretical biologists using the tools of mathematics (Michod, 1999; Nehaniv and Rhodes, 2000). This paper describes a simple discrete developmental genetic regulatory network model based on computational paradigms which could be used to study certain transitions in evolution. This platform has the advantage of being very little parametrized: the GRN model does not require any parameter choices; the GAs and the environment need few and most of them are either easy to control and/or very robust. This makes the whole system relatively easy to analyze.

The problem we used this platform on, the transition of fitness between two different levels of selection, is a fairly untouched problem. We managed to show that in our platform we can observe a sharp transition around a certain value of a parameter weighting the disadvantage of being cooperative. Cooperating in the model is shown to be much more difficult than not (the probability that a randomly drawn genome builds a cooperating network is much smaller than drawing a genome encoding an individualistic GRN). We could say that in our model, even if cooperation is always better than acting individualistically, if the disadvantage due to individualistic behaviour is too small, then multicellular colonies will still evolve to produce less fit non-cooperative individualistic cells and evolution stops in local optima. If we take a broader viewpoint and do make a tentative suggestion, we could state that this might be one of the reasons that in the biological world some environments are populated mostly by primitive organisms: the fitness advantage of being complex simply can not outweigh the difficulties evolving and maintaining the complexity of differentiated multicellularity.

But further work is necessary to make this platform more versatile and credible. Different issues have still to be tackled so as to study the transition from the single cell to the multicellular organism. One issue is to establish that the results here are not model-dependent by investigating the effects of using different models of controller, different types of spaces (2D grid, 3D grid, hex grid, real space, general networked space...), different methods of selection, and of communication. The problem of intrinsic fitness is also linked to fundamental aspects of the model: how to get rid of the standard GA’s explicit fitness? The first step to that would be to get rid of the genetic algorithm itself and design more “natural” evolution for the model, and to get rid of the idea of fitness goals as in this model. And finally if all that would be done the problem would be how to analyse the data? The closer one comes to a model that has all the desired characteristics, the harder it may be to make out what actually happens in the simulations.

References