On the Dynamics of an Artificial Regulatory Network

W. Banzhaf

Department of Computer Science, University of Dortmund, D-44221 Dortmund
banzhaf@cs.uni-dortmund.de
http://ls11-www.informatik.uni-dortmund.de/people/banzhaf/
Tel: +49 231 9700-953 Fax: +49 231 9700-959

Abstract. We investigate a simple artificial regulatory networks (ARNs) able to reproduce phenomena found in natural genetic regulatory networks. Notably heterochrony, a variation in timing of expression, is easily achievable by simple mutations of bits in the genome. It is argued that ARNs are useful and important new genetic representations for artificial evolution.

1 Introduction

Regulatory networks are a fascinating new area of research in biology [4, 5]. With the advent of whole genome information and the realization that - in higher organisms - but a tiny fraction of DNA is translated into protein, the question what the rest of DNA is doing becomes all the more pressing. Regulation seems to be a very reasonable answer for a functional role of unexpressed DNA. Even in single-celled organisms, regulation takes up a substantial part of their highly compressed genomes. According to Neidthardt et al. [18], 88 % of the genome of the bacterium *E.Coli* is expressed, 11 % is suspected to contain regulatory information (see also Thomas [22]).

It is also recognized that the information on DNA-strings controlling the expression of genes is key to understanding the differences between species and thus to evolution [12]. How evolution actually managed to evolve different structures of multicellular organisms from more or less the same proteins seems to be a question of the sequence and the intensity of events during development. Even at the level of biochemical reactions necessary for, e.g. metabolism, regulation plays an important role, thus spanning reaction times from milliseconds (physiology) to Megayears (evolution).

There are three major genetic mechanisms, all tied to regulation [5], which allow such a variety of reactions of living organisms to the pressure for survival:

1. Interactions between the products of genes
2. Shifts in the timing of gene expression (heterochrony)
3. Shifts in the location of gene expression (spatial patterning)

Regulatory networks are used by Nature to set up and control mechanisms of evolution, development and physiology. Regulatory networks unfold the patterns and shapes of organism morphologies and of their behavior. In addition, they mediate between development and evolution, since many evolutionary effects can be followed through their regulatory causes.

It is therefore natural to ask whether we can learn from this type of controlling the organization of matter in the area of artificial evolution. Over the last decades, simple approaches to artificial evolution have proven useful in optimization problems of engineering and in combinatorial problems of computer science [19, 21, 11, 9, 15, 2, 8, 7]. Most of these approaches, however, use a primitive genotype-phenotype mapping without implementing any dynamics into this process of mapping information into behavior. Only recently it was realized that it may the dynamical (and possibly complex) mapping of genotypes to phenotypes, that allows natural systems to evolve with ease (see, for example, [13]).

How can we make use of these insights in artificial evolutionary systems? Previous work in the area is scattered. Eggenberger [6] has studied the patterning of artificial 3D-morphologies. Reil [20] has set up an artificial genome and studied some consequences for artificial ontogeny. Kennedy [14] examined a model of gene expression and regulation in an artificial cellular organism. Bongard and Pfeifer have considered the relation between evolving artificial organisms and behavior [3].

In this contribution we shall present a recently conceived model of a regulatory networks which should be useful for artificial evolutionary systems. The model is a simplification and abstraction of what the author perceives as key elements of the protein-genome interaction in regulatory networks. It is not yet connected to a semantics of structures or behavior, but shows a rich behavioral dynamics. Thus it seems most appropriate to study this artificial regulatory network (ARN) in the context of Artificial Life.

The paper is organized as follows: Section 2 explains the overall view of the artificial regulatory network model, section 3 views it from the static and dynamic perspective. Section 4 explains the concept of heterochronic control. Section 5 exemplifies the plasticity of such systems to evolutionary pressure, section 6 summarizes the discussion and outlines future steps.

2 The artificial regulatory network

Our ARN consists of a bit string with direction (like $5' \rightarrow 3'$ in DNA), the 'genome', and mobile information-carrying molecules, 'proteins', which are equipped with bit patterns for interaction with the genome. Together, they represent a theoretically closed world with a network of interactions between genome and proteins, and a dynamics of protein concentration development determined by this network.

More technically, a mechanism for reading off genes and for producing proteins of particular bit-patterns is given in the form of a 'genotype-phenotype mapping'. Proteins are able to wander about and to interact with any pattern

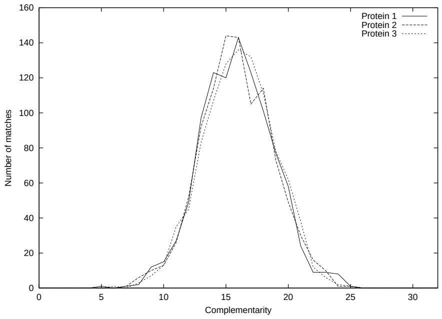

Fig. 1. Distribution of matches for sample proteins with their genome. The distribution is roughly Gaussian, as should be expected from random genomes.

on the genome, notably with 'regulatory sites' located upstream from genes. By attaching to these special sites, they can positively (by enhancing) or negatively (by inhibiting or silencing) influence the production of (other) proteins. We observe the production of proteins and the dynamics of their concentration changes which are a result of the interplay between all the interactions taking place simultaneously. At this time, there is no energy or raw material considered in the system.

The genome is implemented as a sequence of 32-bit (integer) numbers. The length of the sequence, l_G, determines the length of the genome and is frequently used as a parameter. A particular START pattern, the 'promoter', is used to signal the beginning of a gene on the bit string (analogous to an open reading frame (ORF) on DNA), starting at the next integer. The signal used is arbitrary and was chosen as '01010101', a one-byte pattern which in a genome generated by randomly choosing '0's and '1's will appear with a probability of $2^{-8} \approx 0.0039 = 0.39\%$. Genes have a fixed length of $l_g = 5$ 32-bit integers resulting in a bit pattern of 160 bits for each gene (this could be changed later into a STOP signal).

Upstream from the promoter site two special sites are located, one enhancer site and one inhibitor site, both of length 32 bits. Attachment of proteins to these sites will result in changes in protein production of the corresponding gene. It is assumed that an equally low production of proteins takes place if both sides are unoccupied. Usually, however, there will be proteins around to influence the expression rate of a particular gene, and we shall look at that in more detail later. In this simple model, we restrict ourselves to just one regulatory site for expression and suppression of proteins, a radical simplification with regard to natural genomes, where 5-10 regulatory sites are the rule that might even be occupied by complexes of proteins.

Proteins are produced from genes by feeding their bit patterns into a genotype-phenotype mapping function and producing mobile elements carrying other bit patterns, the proteins. In this model, therefore, we disregard the transcription process completely. Further, there are no introns, no RNAs and no translation procedure resulting in a different alphabet for proteins. Instead, proteins consist of bit patterns of a particular type: Each protein is a 32-bit number resulting from a many-to-one mapping of its gene: On each bit position in the gene's integers the majority rule is applied so as to arrive at one bit for the protein. In the case of a tie (not possible with an odd number for l_g), this is resolved by chance.

Proteins can now be examined as to how they 'match' the genome: Each bit pattern of a protein can be compared to the genome pattern with the overlap being the number of bits set in an XOR operation. Thus, complementarity between genome and protein bit patterns determines their match. In general, it can be expected that a Gaussian match distribution results when shifting proteins over all the sequence of a random genome. Notably, there are a few high-matching and a few low-matching positions and many average-matching ones on the genome.

3 Static and dynamic view

Let us first look at examples from the static perspective. Table 1 gives three examples of genomes with increasing size. We list the number of genes which roughly follows the 0.39% rule, the maximum match between resulting proteins and their genome at any location, and the number of times such a maximum match has been found.

Table 1. Sample genomes of increasing size. As can be seen the number of proteins with maximum match remains about the same, but their specificity increases.

Genome length l_G	Number of genes	Maximum match	Frequency of max. match
1000	3	25	3
10000	37	28	4
100000	409	30	3

If we look at the distribution of matches the picture according to Figure 1 emerges for a sample genome with 3 proteins: Roughly a Gaussian distribution of matches is found for each of these genes.

Let's change perspective and look from the genome's point of view, and more specifically, from the point of view of regulation sites. A number of proteins are produced and floating by, with some providing better matches to the site, other proteins providing worse matches. In principle, each protein has the potential to interact with each regulatory site, and the degree of matching will determine the probability of occupation of a certain site with a certain protein.

Because proteins are competing for attachment to regulatory sites, the probability of occupation with a particular protein is dependent on the degree of matching of all other proteins to this site.

Under the simplifying assumption that occupation of two regulatory sites per gene modulates the expression of corresponding protein, a network of interactions between genes and proteins can be deduced, which can be parametrized

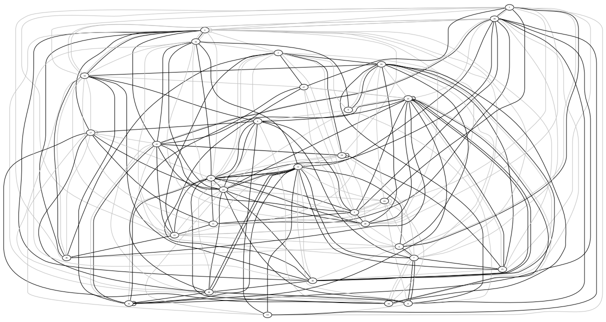

Fig. 2. Distribution of matches of proteins of a larger example. 9 out of 32 proteins are depicted and their matches with the enhancer (black) and inhibitor (gray) regulatory sites of all 32 genes of this example ($l_G = 10,000$).

by strength of match. Figure 2 shows a sample network, again taken from the example genome with 32 genes / proteins.

No evolution has taken place (recall these genomes are generated by randomly drawing bits), and the network of interactions shows a highly random view of the resulting interactions. These networks must be considered very complex (in terms of layers vs. participating nodes) and a deep hierarchy of interactions is visible[1].

In the rest of this section we shall concentrate on the dynamics of the interaction network. A match between a protein and regulatory site of a gene leads to activation or inhibition of protein production of the corresponding gene. Generally, the influence of a protein i with $i = 1, ..., n_p$ on an enhancer/inhibitor site is exponential in the number of matching bits, $exp(\beta(u_i - u_{max}))$ where u_{max} is the maximum match achievable.

The concentration of protein molecules c_j of protein j modulates this strength to produce the following excitatory / inhibitory signals for the production of protein i:

$$e_i = \frac{1}{N}\sum_j c_j e^{\beta(u_j^+ - \bar{u}_{max}^+)} \tag{1}$$

$$in_i = \frac{1}{N}\sum_j c_j e^{\beta(u_j^- - \bar{u}_{max}^-)} \tag{2}$$

where a scaling was done as to have a maximum match for the best matching protein, both in excitatory and inhibitory signals.

Given these signals, protein i is produced via the following differential rate equation

$$\frac{dc_i}{dt} = \delta(e_i - in_i)c_i - \Phi \tag{3}$$

[1] As has been observed in natural genome organization, shallow hierarchies, up to the point of modularity, are a hallmark of biological organisms. It is interesting to note that a simple process of duplication and divergence suffices to reach a similar state, even from a random genome [1].

A flow term assures that concentrations remain in the simplex, $\sum_i c_i = 1$, resulting in competition between sites for proteins.

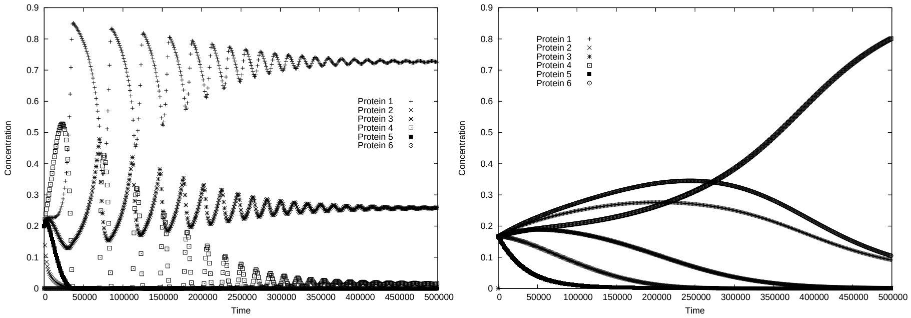

Fig. 3. Two different dynamical systems realized by two different genomes. Left: A dampened (nonlinear) oscillator type of dynamics. Right: Slow and smooth development of concentrations.

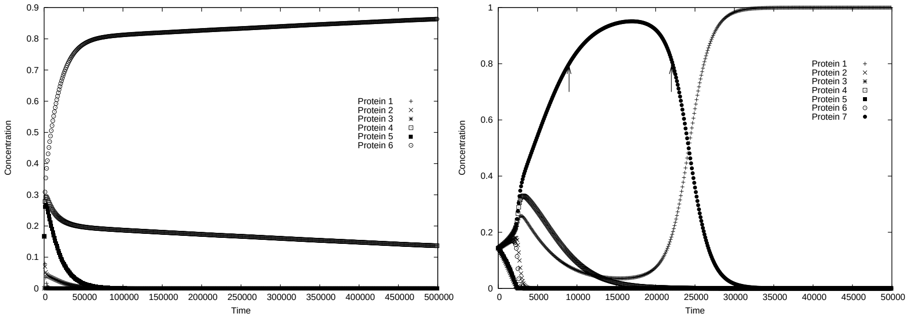

Fig. 4. Two other dynamics. Left: Settlement into a point attractor is well under way. Right: Extended transition phase with one protein achieving high values of concentration then switching the state to expression of another gene.

If we look at the dynamics of concentration changes of proteins, starting from a state of equal concentration that reflects the native low-level expression of all genes, we can observe that some proteins increase their level of concentration, then fall again, with usually one being left over. Thus, a typical dynamic system behavior can be seen, well known as 'point attractor' in dynamical systems theory [10].

For different random genomes (different number of genes, matching etc) the dynamics is remarkably different. There are cases of longer and shorter time scales, there are complicated and simple dynamics. Figures 3 - 4 show four different dynamics resulting form four different genomes.

It should be noted that this richness of dynamics is merely a result of different genomes of the same length, with different patterns for proteins resulting in different matching and regulation results. No development has yet been put in place. There are three types of competition effective simultaneously, (i) competition of proteins for binding sites (only regulatory sites considered), (ii) competition of binding sites for proteins, and (iii) competition of genes for raw material for production of proteins. This latter competition is implemented by normalizing the strength of the inhibition / enhancement signals through division by N.

4 Heterochronic control

If we look at this from the perspective of how many proteins are above or below a certain production threshold we can observe the turning on or turning off of genes (on/off could be set equal to x 2 or x 1/2 of initial production, or it could be based, as here, on an absolute concentration value, 0.8 here). This translates into a timing of onset/termination of protein production. Figure 4, right, for instance, shows the timing of onset and termination of concentrations above 0.8 for protein 7 (arrows), $t_{on} = 9,000, t_{off} = 22,000$.

Changing the degree of matching between regulatory sites and proteins by one or two bits can result in dramatic changes in the dynamics, but it must not. Sometimes there are no changes at all and we have a neutral variation. Sample changes that actually varied the expression are shown in Figure 5. It is interesting to note that variations in patterns are translated by the ARN into time variations, similar to what was observed in natural GRNs [5]

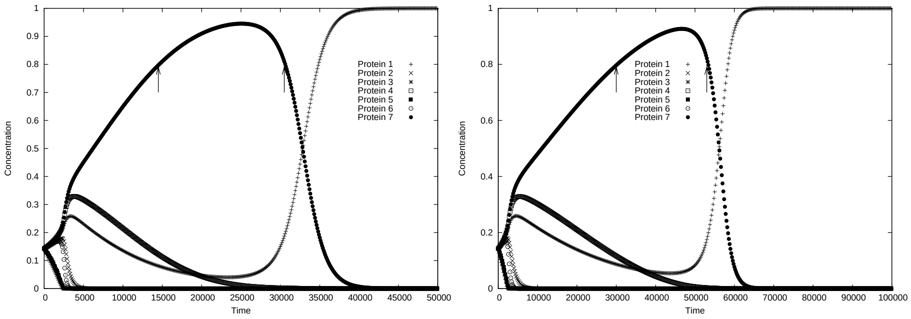

Fig. 5. Genome of Figure 4. Left: Degree of matching between protein 7 and inhibitory site to gene 4 changed by one bit. Timing of expression of protein 7 changes substantially: $t_{on} = 14,500, t_{off} = 30,500$. Right: Degree of matching between protein 7 and inhibitory site to gene 4 increased by another bit, timing changes even further $t_{on} = 30,000, t_{off} = 53,000$.

Heterochrony, i.e. a variation in the timing of onset or offset of certain genes are heavily used in development for generating particular structural effects [17, 16]. As we can see by comparing Figures 5, left and right, and 4, right, small changes cause small effects. The same principle could be also of use in physiological reactions, for instance under the control of external factors exceeding certain threshold values.

Interestingly, the range of possible changes is partitioned logarithmically, due to the change of occupation probability, that is depending on an exponentiated matching difference between proteins and DNA bit-patterns. This can be seen most easily, if we put all concentration curves of protein 1 and 7 into one plot, see Figure 6. We can clearly see the range of changes expanding with further additions of bit flips.

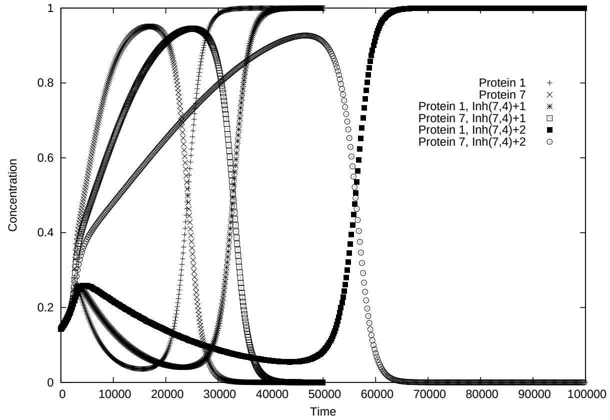

Fig. 6. Genome of Figure 4. Degree of matching between protein 7 and inhibitory site to gene 4 changed progressively. Timing of expression of protein 7 changes in increasing step sizes.

5 Evolution

The most important question to be addressed with such a model is whether it would be possible to define arbitrary target states and evolve the genome / protein network toward this target state. Our first results in a typical simulation are shown in Figure 7. It shows the progress of a network in approaching the target concentration of a particular protein, here protein 6. As we can see, the evolutionary process quickly converges towards this target state. It must be emphasized, that the very simplest way of doing evolution was used here, a $(1 + \lambda)$ evolution strategy, with $\lambda = 1$ [19]. Various experiments were performed with the same genome (not shown here), allowing evolution of other concentration levels for other proteins. We can see from the figure, that steep declines in the deviation (error) curve are followed by stagnation periods. These stagnation periods, are, however, accompanied by continued changes in the genome under evolution. It is merely the mapping of the genome that does not show many consequences of these variations. By construction we designed a system with many neutral pathways. During some periods it does not look as if there is evolutionary progress, but nevertheless changes happen in genomes due to neutral steps.

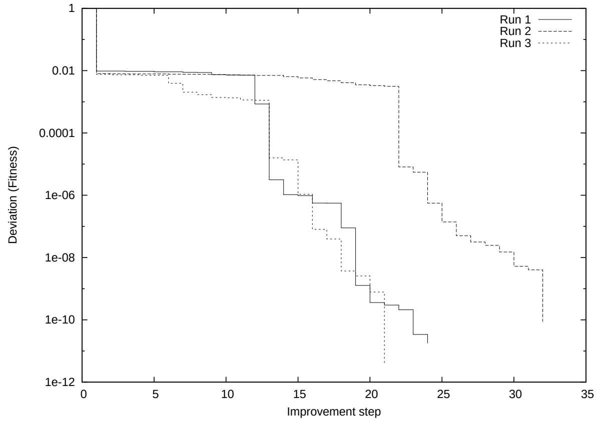

Fig. 7. Evolution at work: 3 different runs of a (1+λ) strategy to arrive at a prespecified concentration of one particular protein: $c_6 = 0.085$ at time $t = 100$.

This can be seen if we consider the changes in concentration levels of all proteins at $t = 100$ in Figure 8. Here we can discover that all protein concentrations change over time, with many stagnation periods for all proteins. Huge steps are sometimes shown by certain proteins, which are not reflected in the fitness of an individual, due to the focus on measuring only the deviation from $c_6 = 0.085$ for fitness.

6 Summary

In this contribution we have shown that a simple model for artificial regulatory networks can be formulated which captures essential features of natural genetic regulatory networks. Although our investigation is preliminary in that it is only qualitative in results, the different behavior of these networks from usual genetic represenations can be seen already from the few examples shown here.

With this contribution we have just started on a path that relates changes in time and intensity to tiny pattern changes on bit strings. As such, the network picture of a genome might be a very fruitful approach and could possibly provide the algorithmic "missing link" between genotypes under constant evolutionary changes and remarkably stable phenotypes that we find in the real world.

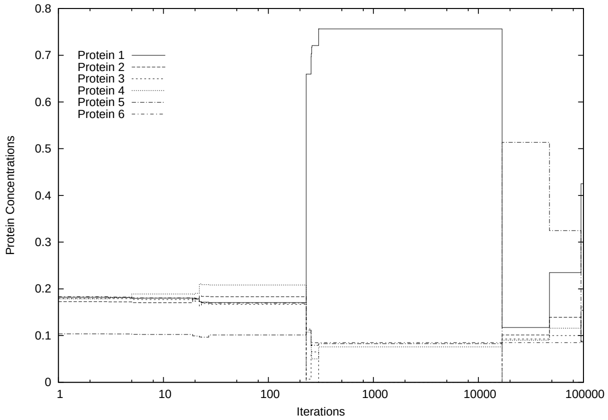

Fig. 8. Evolution at work: Same run as in Figure 7, with all protein conentrations protocolled at $t = 100$. As can be seen, protein concentration of selected protein 6 meanders towards goal state $c_6 = 0.085$, whereas other protein concentrations pass through huge swings.

Acknowledgement

The author gratefully acknowledges a sabbatical stay at the Institute for Genomics and Bioinformatics at UC Irvine, where part of the ideas that lead to this work were born. Especially he wants to acknowledge the hospitality of its director, Prof. Pierre Baldi, and of its manager, Mrs. Ann Marie Walker.

References

1. W. Banzhaf. Artificial Regulatory Networks and Genetic Programming. In R.L. Riolo et al., editor, *Genetic Programming – Theory and Applications*, 2003, to appear.
2. W. Banzhaf, P. Nordin, R. Keller, and F. Francone. *Genetic Programming - An Introduction*. Morgan Kaufmann, San Francisco, CA, 1998.
3. J. Bongard and R. Pfeifer. Behavioral selection pressure generates hierarchical genetic regulatory networks. In W. B. Langdon et al., editor, *Proceedings of the Genetic and Evolutionary Computation Conference*, page 132. Morgan Kaufmann, 2002.
4. J.M. Bower and H. Bolouri (Eds). *Computational Modeling of Genetic and Biochemical Networks*. MIT Press, Cambridge, MA, 2001.
5. E.H. Davidson. *Genomic Regulatory Systems*. Academic Press, San Diego, CA, 2001.
6. P. Eggenberger. Evolving morphologies of simulated 3d organisms based on differential gene expression. In Inman Harvey and Phil Husbands, editors, *Proceedings of the 4th European Conference on Artificial Life*, pages 205 – 213. Springer, 1997.
7. D.B. Fogel. *Evolutionary Computation*. IEEE Press, New York, 1995.
8. L.J. Fogel, A.J. Owens, and M.J. Walsh. Artificial intelligence through a simulation of evolution. In M. Maxfield, A. Callahan, and L.J. Fogel, editors, *Biophysics and Cybernetic Systems*, pages 131–155, 1965.

9. D.E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley, Reading/Mass., 1989.
10. M. Hirsch and S. Smale. *Differential Equations, Dynamical Systems and Linear Algebra.* Academic Press, Reading, MA, 1997.
11. J.H. Holland. *Adaptation in natural and artificial system.* MIT Press, Cambridge, MA, 1992.
12. L. Hood and D. Galas. The digital code of dna. *Nature*, 421:444 – 448, 2003.
13. H. Kargupta. Editorial: Special Issue on Computation in Gene Expression. *Genetic Programming and Evolvable Machines*, 3:111 – 112, 2002.
14. P.J. Kennedy and T.R. Osborn. A model of gene expression and regulation in an artificial cellular organism. *Complex Systems*, 13, 2001.
15. J.R. Koza. *Genetic Programming: On the Programming of Computers by Natural Selection.* MIT Press, Cambridge, MA, USA, 1992.
16. M. McKinney. Heterochrony: Beyond words. *Paleobiology*, 25:149 – 153, 1999.
17. M. McKinney and K. McNamara. *Heterochrony: The Evolution of Ontogeny.* Plenum Press, New York ,NY, 1991.
18. F.C. Neidhardt. *Escherichia Coli and Salmonella typhimurium.* ASM Press, Washington, DC, 1996.
19. I. Rechenberg. *Evolutionsstrategie "93.* Frommann Verlag, Stuttgart, 1994.
20. T. Reil. Dynamics of gene expression in an artificial genome - implications for biological and artificial ontogeny. In D. Floreano et al., editor, *Proceedings of the 5th European Conference on Artificial Life*, pages 457–466. Springer, 1999.
21. H.-P. Schwefel. *Evolution and Optimum Seeking.* Sixth-Generation Computer Technology Series. John Wiley & Sons, Inc., New York, 1995.
22. G.H. Thomas. Completing the e. coli proteome: a database of gene products characterised since completion of the genome sequence. *Bioinformatics*, 7:860 – 861, 1999.