# A Framework for the Co-evolution of Genes, Proteins and a Genetic Code Within an Artificial Chemistry Reaction Set

Ken Gardiner, James Harland, and Margaret Hamilton

RMIT University, School of Computer Science and Information Technology,
GPO Box 2476V, Melbourne, Victoria, 3001, Australia
{ken.gardiner,james.harland,margaret.hamilton}@rmit.edu.au

**Abstract.** We present an artificial chemistry model where genotypic and phenotypic strings react with each other. The model prevents the genome from directly coding for genotype-phenotype mappings or for gene-replication enzymes. Experiments demonstrate the genome can evolve to manipulate reactions of phenotypic strings in such a way as to alter the genotype-phenotype mapping, and produce gene-replication enzymes.

## 1 Introduction

Artificial chemistries [2] are abstract chemical models, where entities represented as 'molecules' undergo collision reactions or transformations according to a set of rules, inside some specified environment. Artificial chemistries have been applied to the investigation of many life-like phenomena including autocatalytic 'metabolisms' [1,3].

Some artificial chemistry models include both informational (gene) and functional (protein) molecules and permit translation of proteins from genes, as well as gene replication by proteins [4,8]. It has been shown that evolving a genotype-phenotype mapping provides the opportunity to transform a problem representation into a form that is easier to solve [5]. Artificial chemistries with genes and proteins have been designed to evolve a genotype-phenotype mapping, often implemented by permitting genes to code for genotype-phenotype mapping molecules [7,9].

We are interested in developing an artificial chemistry where 'services' required by the genes cannot be directly produced from molecules coded for by the genes. This feature is intended to force the genome to only manipulate *protein-protein* reactions (from an inflow of 'food' proteins) in order to construct molecules for providing those services. While models such as [6,10] have evolved protein-protein metabolisms by manipulating genomes external to the reaction set, we are interested in evolving protein-protein metabolisms under internally produced genetic control.

Our initial artificial chemistry is designed so that molecules for the gene 'services' of gene-duplication and genotype-phenotype mapping can only be implemented by *service* molecules produced from protein-protein reactions.

M. Randall, H.A. Abbass, and J. Wiles (Eds.): ACAL 2007, LNAI 4828, pp. 281–291, 2007.

The artificial chemistry imposes some general syntax restrictions on the structure of genes, proteins and service molecules, but does not specify a priori the effectiveness of any particular (legal) arrangement of atoms. Instead, the semantics of a molecule depends on the set of other molecules that it is in, forcing the genes and service molecules to co-evolve in order to be effective.

This paper presents the model and describes the current implementation. Preliminary results are presented. They demonstrate the model is capable of evolving from initially random molecules, to genetically manipulate protein-protein reactions and produce service molecules for genotype-phenotype mapping and potential gene-duplication.

## 2 The Model

An artificial chemistry can be described [2] as a triple *(S,R,A)* where $S$ is the set of all possible molecules, $R$ is the set of collision rules describing interactions among molecules, and $A$ is the control algorithm for describing the domain and how the rules are applied to the molecules. The main contribution of our model is the approach to the collision rules.

### 2.1 The Control Algorithm

The control algorithm determines how the collision rules will be applied to a collection of molecules, manages the 'reactor vessel' environment of the molecules, and implements an evolutionary algorithm based on a reaction-set fitness function.

Each reaction set operates in its own simulated well-stirred (i.e. dimensionless) vessel, which is initially seeded with a food stock of random protein molecules and a random stock of genes. The food stock does not include gene-replication molecules or genotype-phenotype mapping molecules. The food stock has its concentration increased at a steady rate, while the concentration of genes remains constant. The simulated vessel has a total atom-count limit, beyond which molecules are randomly selected for overflow. Future implementations will permit gene replication and gene overflow. Currently we prevent gene molecules from overflowing, and do not implement gene replication.

Once the concentration of a molecule exceeds a user specified threshold, it can potentially take part in chemical reactions. Such reactions may change the concentrations of the molecule and result in new molecular species, which may also take part in reactions if their concentrations exceed the concentration threshold.

The control algorithm implements an evolutionary algorithm. A generation consists of running each reaction set for a user defined number of cycles. A cycle involves adding food stock to the reaction set and stochastically performing molecular collisions until the concentrations of reaction inputs are insufficient to run any more reactions. The simulated vessel then overflows until the number of atoms in the vessel falls below a specified threshold value. The total number of gene-replication molecules produced by a reaction set over the cycles is used as that set's fitness score for that generation.

At the end of a generation, the evolutionary algorithm copies the two highest scoring reaction sets into the next generation. The rest of the generation is populated with copies of reaction sets with some of their genes mutated. The reaction sets are selected for copying by weighted roulette wheel selection based on their fitness score.

### 2.2 Molecules

We define molecules as consisting of character-based strings, drawn from the set of {F,G,C,P,1,2,3,4,x,y,z}.

Gene molecules can only be composed from {x,y,z}. Gene molecules are capable of undergoing mutation, while other molecules are not.

Protein molecules can be composed of any atoms as long as they contain at least one atom from {F,G,C,P,1,2,3,4}.

*Service* molecules are a subclass of protein, used to implement a genotype-phenotype mapping, or to perform gene replication. Genotype-phenotype service molecules consist of two atoms drawn from {F,G,C,P,1,2,3,4} followed by two or more atoms drawn from {x,y,z}. Gene-replication service molecules have a mirror-image syntax to genotype-phenotype service molecules: i.e. two or more atoms from {x,y,z} followed by two atoms drawn from {F,G,C,P,1,2,3,4}.

### 2.3 Collision Rules

The model supports two types molecular interaction: protein-protein interactions and interactions between service molecules (a subtype of protein) and genes.

**Protein-protein interactions.** If two or more protein molecules collide, one of the proteins is randomly chosen to act as a catalyst in a potential reaction, performing some arrangement of operations on the other molecule(s).

We chose a simple pattern matching operation to determine whether colliding protein molecules interact. Protein-protein reactions are only possible if the molecules bind. A subset of atoms, called *latch* atoms, drawn from {1,2,3,4}, determine if protein molecules will bind, and how they will be aligned if a binding does occur. Latch atoms on one molecule are attracted to latch atoms on another molecule, according to the attraction patterns 1:3, 3:1, 2:4 or 4:2. A protein will bind to a (catalyst) protein if more than a specified number of their latch atoms bind (in the current implementation, this threshold is two). The heterogeneous binding pattern was chosen to reduce the probability of a protein binding with, and possibly destroying, itself.

A protein acting as a catalyst may perform *cut* (ligation) or *paste* (polymerisation) operations on the bound molecule(s). The cut operation is specified by a *C* atom, and paste by a *P* atom.

Protein molecules may also contain two inert atoms, indicated by *F* and *G*.

Figure 1 illustrates the operation of protein binding, with cut and paste operations. Sites of ligation or polymerisation are shown in grey.

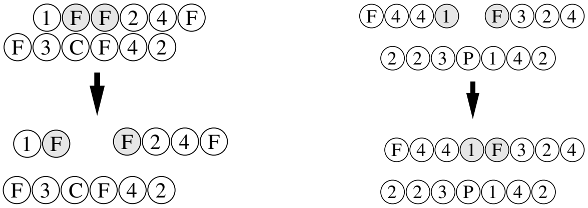

**Fig. 1.** The *cut* and *paste* operations

**Service molecule protein-gene interactions.** The model is intended to prevent genes from directly coding for molecules that provide services to genes such as gene replication or a genotype-phenotype mapping. However the model has to permit the production of such molecules from protein-protein reactions. In addition, the protein-protein reactions to produce genotype-phenotype mapping molecules must be simple enough for a reasonable probability that, given a random collection of proteins, genotype-phenotype mapping molecules could be produced by chance. Without this, the system would not have an initial genotype-phenotype mapping and genetic manipulation of the system would be impossible. The service molecule syntax, described in Sect. 2.1 was designed to meet these constraints.

A simple pattern matching operation was used to determine if and where service proteins could interact with genes. Genes are composed solely of *template* atoms, drawn from {x,y,z}. The template atoms of a service protein may bind with gene template atoms according to the patterns x:x, y:y or z:z. Each service protein template atom must bind with a gene template atom, otherwise a reaction will not occur.

To translate a gene into a protein, genotype-phenotype mapping molecules bind to the gene as illustrated in the left-hand side of Fig. 2. The first atom of each mapping molecule then polymerises to produce a new protein (in the example, protein *P2* is produced). The syntax of genotype-phenotype mapping molecules could have been designed to contain only a single atom from {F,G,C,P,1,2,3,4}, instead of two such atoms. The current syntax was chosen due to considerations for future modification of the model, which will not be presented here.

A gene may contain regions that no mapping molecule matches. These regions act to stop translation, permitting a gene to have several protein coding regions, each separated by 'stop' regions.

Gene replication could be implemented by a system similar to that for genotype-phenotype mapping. However in this case, it is the template atom regions of the gene-replication molecules that polymerise into a new gene molecule. The process is illustrated in the right-hand side of Fig. 2. Gene replication is currently not enabled.

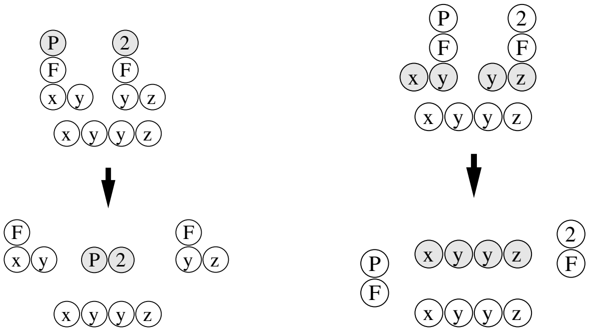

**Fig. 2.** Gene translation and replication

## 3 Experiments and Results

Recall that our model was designed to force genes to manipulate protein-protein reactions in order to produce service molecules required by the genes. In this section we describe our series of experiments to generate the expected behaviour of increasing production of potential gene-replication molecules.

The fitness score of a reaction set would increase if genes were able to manipulate protein-protein reactions to increase the production of gene-replication molecules. To test this, we initialised the artificial chemistry system with random molecules and ran it under various settings to see if fitness scores improved over time. None of these initial, unreported, experiments resulted in improvements in fitness score. Analysis showed that genes required more genotype-phenotype mapping molecules than could be produced from the provided concentration of food stock. Therefore genes could not be translated and could not influence protein-protein reactions. Increasing the concentration of each food stock species resulted in an unacceptable run time.

Based on these results, we altered our model to permit unlimited use of any genotype-phenotype mapping molecules produced from protein-protein reactions. The genes still had to evolve to control the protein-protein reactions leading to genotype-phenotype mapping molecule production. The total atom-count was kept constant by counting the number of atoms in gene-translated proteins and adding that number to the atom-overflow cycle of the simulator. The alteration to the model permitted improvements in the fitness score to evolve and various settings of the simulator were investigated (not reported here) resulting in the choice of running the simulator for 600 generations of 300 feed-react-overflow cycles each. Approximately 7% of a reaction set's non-gene contents were replaced with food stock in each cycle. The mutation rate was 0.2% of gene atoms. Figure 3 shows the best-of-generation fitness score for a run using these settings. The model was able to evolve increasingly successful reaction sets, increasing the system's fitness by 223%.

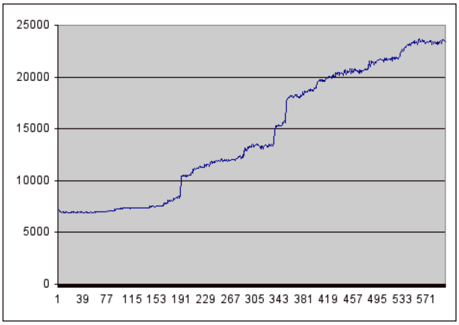

**Fig. 3.** Best-of-generation fitness verses generation

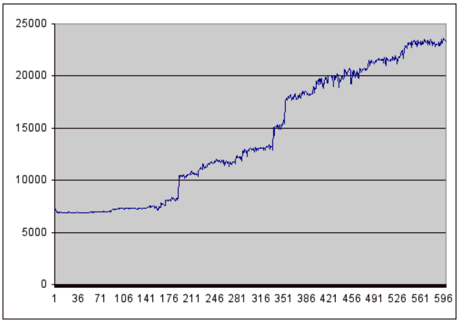

**Fig. 4.** Fitness verses generation of the ancestors of the highest scoring reaction set from generation 600

The evolutionary algorithm performed duplication and mutation of reaction sets, but not crossover. This meant each reaction set in generation 600 had a single ancestor in each of the previous generations. The ancestors of the winning reaction set from generation 600 were examined. Their scores are shown in Fig. 4. Analysis of these ancestors was undertaken to reveal how their genes evolved to increase the reaction set fitness.

### 3.1 Genetic Influence over Protein-Protein Reactions

There are two ways genes could influence the production of gene-replication molecules. Firstly, the genes could produce proteins that impede reactions detrimental to gene-replication molecule production. Such genes will be called *impeding* genes. Secondly, they could produce proteins that form part of the reaction path for the production of gene-replication molecules. These genes will be called *production* genes.

Production genes were identified by working backwards through the reactions from gene-replication molecule to food stock, and tagging any genes that produced molecules in the reactions.

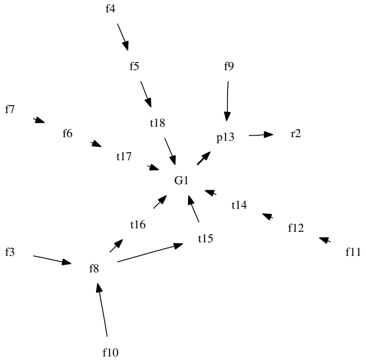

**Fig. 5.** Generation 59 production gene reaction subset

In order to determine if impeding genes existed, an additional experiment was performed where all genes except production genes in a reaction set were prevented from producing proteins. The results were then compared against running the reaction set with all genes turned off. It was found that turning off the production genes was equivalent to turning off all genes, indicating that the reaction set did not use impeding genes.

Improvements in reaction set scores (and thus gene-replication molecule production) began from generation 59. Prior to then, although genes were producing proteins, those proteins were not influencing reactions that produced gene-replication molecules.

At generation 59, the reaction set contained 9,526 molecular species and 4002 reactions. This was the first reaction set to contain a production gene, *G1*, producing a molecule that lead to the production of a gene-replication molecule *r2*. Figure 5 shows the reactions leading to the production of *r2*. Each molecular species has been given an integer identifier. The identifiers are prefaced as following: *f* indicates a food stock protein, *G* indicates a gene, non-food proteins are prefaced with *p*, *t* indicates a genotype-phenotype mapping molecule and *r* indicates a gene-replication molecule. An example reaction shown in the figure is: food protein *f11* reacts with food protein *f12* to produce genotype-phenotype molecule *t14*. The *t14* molecule is then used by gene *G1* (together with genotype-phenotype molecules *t15*, *t16*, *t17* and *t18*) to produce protein *p13*. Protein *p13* then reacts with food protein *f9* to produce gene-replication molecule *r2*.

Further evolution of the system produced additional production genes, resulting in the production of further species of gene-replication molecule. For example, the major increase in fitness at generation 191 was caused by the evolution of two additional production genes. By generation 600 the reaction set had 13 production genes, assisting the production of 9 species of gene-replication molecule. Figure 6 shows these reactions (a subset of the 21139 reactions occurring in the entire reaction set). In order to reduce the complexity of the figure, the only molecules labelled are genes and gene-replication molecules. Reactions between geneotype-phenotype molecules and genes are shown with dotted lines.

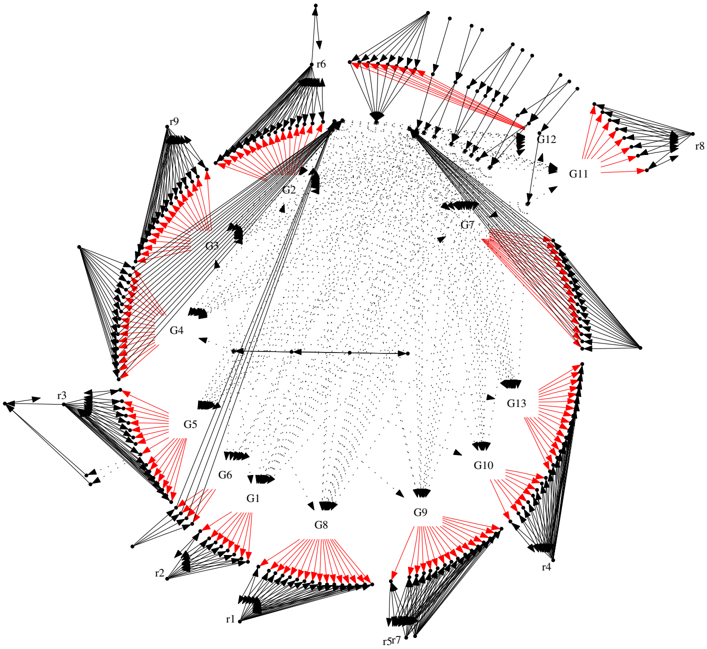

**Fig. 6.** Generation 600 production gene reaction subset

### 3.2 Genetic Influence over the Genotype-Phenotype Mapping

The genotype-phenotype mapping produced by the initial random reaction set was a many-to-many mapping between template-atom-pattern and protein-atom. This meant a single segment of gene could be translated multiple ways, resulting in the production of more than one protein. Table 1 compares the set of genotype-phenotype mapping molecules forming the initial genetic code and the set of mapping molecules produced in generation 600. The first atom (underlined) is coded for by the string of template atoms (italicised). This means some genotype-phenotype mapping molecules are functionally equivalent. For example $\underline{\mathrm{G}}\mathrm{P}zx$ and $\underline{\mathrm{G}}\mathrm{G}zx$ both map *zx* to *G*. Gene template atom sequences not included in the genetic code can be considered as stop instructions.

Table 1 shows that most of the genotype-phenotype mapping molecules produced from the initial random reaction set continued to be used in generation 600. However the evolving genes did modify the reaction set to expand the genetic code. The final genetic code included every two-atom template pattern except *zy* and *zz* (which therefore formed ‘stop translation’ codes) although a mapping for *zzx* was produced. Since genes could use an unlimited supply of any genotype-phenotype mapping molecules produced by protein-protein reactions, there was no incentive to prevent the waste of mapping molecules. This meant a single region of gene could be translated into multiple species of protein without penalty. Therefore there was no incentive to evolve a non-overlapping genetic

**Table 1.** Genotype-phenotype mapping molecules of generation 0 and generation 600

| Generation 0 | Generation 600 |
|---|---|
| $\underline{\mathrm{C}}\mathrm{P}xx$ | $\underline{\mathrm{C}}\mathrm{P}xx$ |
| | $\underline{3}\mathrm{F}xy$ |
| $\underline{\mathrm{G}}\mathrm{G}xz$ | $\underline{\mathrm{G}}\mathrm{G}xz$ |
| | $\underline{\mathrm{P}}\mathrm{C}xz$ |
| $\underline{\mathrm{F}}\mathrm{F}yx$ | $\underline{\mathrm{F}}\mathrm{F}yx$ |
| | $\underline{\mathrm{P}}1yx$ |
| $\underline{\mathrm{G}}4yy$ | $\underline{\mathrm{G}}4yy$ |
| $\underline{\mathrm{P}}\mathrm{P}yy$ | |
| $\underline{\mathrm{F}}4yy$ | $\underline{\mathrm{F}}4yy$ |
| $\underline{\mathrm{F}}\mathrm{F}yz$ | $\underline{\mathrm{F}}\mathrm{F}yz$ |
| $\underline{\mathrm{P}}\mathrm{F}yz$ | $\underline{\mathrm{P}}\mathrm{F}yz$ |
| $\underline{\mathrm{P}}\mathrm{P}yz$ | $\underline{\mathrm{P}}\mathrm{P}yz$ |
| $\underline{\mathrm{G}}\mathrm{P}zx$ | $\underline{\mathrm{G}}\mathrm{P}zx$ |
| | $\underline{\mathrm{G}}\mathrm{G}zx$ |
| $\underline{1}\mathrm{F}zzx$ | $\underline{1}\mathrm{F}zzx$ |

code, and every incentive to ensure almost every gene template pattern could be translated.

Analysis of the reaction sets showed genes modified the genetic code by evolving to code for proteins that modified the protein-protein reaction set to produce new classes of genotype-phenotype mapping molecule. For example, Fig. 7 shows part of the reaction set for generation 250. Protein *p67* was produced from translating gene *G20* (reactions producing the genotype-phenotype molecules used by *G20* are not shown). Protein *p67* then catalysed a reaction with food molecule *f59*, leading to the production of genotype-phenotype mapping molecule *t40* (which was then used by other genes). Such new species of mapping molecule could permit previously un-translatable gene template-atom sequences to be accepted and translated, effectively increasing the number of genes participating in the reaction set.

Analysis of the reaction sets also demonstrated a co-evolution of the genetic code, genes and other reaction set molecules resulting in autocatalytic reaction loops involving genes. For example, Fig. 8 shows part of another reaction subset from generation 250. It shows the genotype-phenotype molecules used by

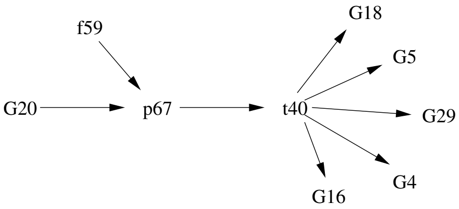

**Fig. 7.** Genetic control of reaction creating genotype-phenotype molecule

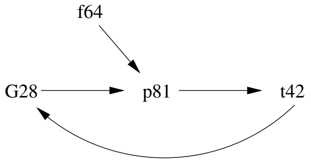

**Fig. 8.** Gene use of genotype-phenotype mapping molecule produced only from translation of that gene

gene *G28*, and the molecules required to create one of those genotype-phenotype molecules, *t42*. It can be seen that *t42* is produced from the reaction of protein *p81* with food protein molecule *f64*. However protein *p81*, used to create *t42*, is produced by gene *G28*, which requires *t42*. The full reaction set was examined, revealing that no other reaction produced *p81*. This leads to an apparent paradox in that gene *G28* couldn't be translated without genotype-phenotype mapping molecule *t42*, yet the translation of *G28* was required to produce *t42*. Analysis of the reaction set's ancestors showed that there used to be an alternative form of production of *t42* (thus enabling gene *G28* to use it), and later evolution caused the demise of the alternative pathway, leaving the autocatalytic loop between *G28* and *t42*.

## 4 Conclusions and Future Work

We have presented an artificial chemistry containing informational (gene) and functional (protein) molecules, designed so that gene 'services' of gene replication and a genotype-phenotype mapping can only be produced by genetic manipulation of protein-protein reactions. We presented preliminary results showing the system can and does evolve genetic control of protein-protein reactions in order to produce a genotype-phenotype mapping and increase production of potential gene-replicating molecules. Future work will include further experiments to examine the range of gene control over protein-protein reactions possible under our model.

Currently, a given genetic sequence can lead to the translation of more than one species of molecule due to overlap in the genetic code. If only one of the species from such translations was useful, then the genotype-phenotype mapping molecules used to produce the other translations would have been wasted. A future area of investigation will be to introduce a cost of wasting genotype-phenotype mapping molecules, providing a selection pressure to remove genetic code overlap.

Another area of investigation will be to place the replication of genes under genetic control, via the use of gene-replication molecules. Genes will be permitted to overflow the simulated vessel, providing pressure to copy genes before they are flushed out. The eventual intention is to divide each reaction set into two after a set time period, implementing a simple form of reproduction and further pressuring the reaction set to copy genes and other vital molecules before the division occurs.

## References

1. Bagley, R.J., Farmer, J.D.: Spontaneous Emergence of a Metabolism. In: Langton, et al. (eds.) Artificial Life II, SFI Studies in the Sciences of Complexity, vol. X, Addison-Wesley, Reading (1991)
2. Dittrich, P., Ziegler, J., Banzhaf, W.: Artificial Chemistries - A Review. Artificial Life 7, 225–275 (2001)
3. Farmer, J.D., Kauffmann, S.A., Packard, N.H.: Autocatalytic Replication of Polymers. Physica 22D, 50–67 (1986)
4. Hutton, T.J.: Evolvable Self-Reproducing Cells in a Two-Dimensional Artificial Chemistry. Artificial Life 13, 11–30 (2007)
5. Kargupta, H., Ghosh, S.: Toward Machine Learning Through Genetic Code-like Transformations. Genetic Programming and Evolvable Machines 3(3), 231–258 (2002)
6. Lohn, J.D., Colombano, S.P., Scargle, J., Stassinopoulos, D., Haith, G.L.: Evolving Catalytic Reaction Sets using Genetic Algorithms. In: Proc. IEEE Int. Conf. on Evolutionary Computation, New York, pp. 487–492 (1998)
7. Piaseczny, W., Suzuki, H., Sawai, H.: Chemical Genetic Programming - Evolutionary Optimization of the Translation from Genotype String to Phenotypic Trees. In: Sugisaka, M., Tanaka, H. (eds.) Proc. 9th Int. Symposium on Artificial Life and Robotics, vol. 2, pp. 571–574 (2004)
8. Suzuki, H.: Models for the Conservation of Genetic Information with String-Based Artificial Chemistry. In: Banzhaf, W., Ziegler, J., Christaller, T., Dittrich, P., Kim, J.T. (eds.) ECAL 2003. LNCS (LNAI), vol. 2801, pp. 78–88. Springer, Heidelberg (2003)
9. Suzuki, H., Sawai, H., Piaseczny, W.: Chemical Genetic Algorithms - Evolutionary Optimization of Binary-to-Real-Value Translation in Genetic Algorithms. Artificial Life 12, 89–115 (2006)
10. Ziegler, J., Banzhaf, W.: Evolving Control Metabolisms for a Robot. Artificial Life 7, 171–190 (2001)