Towards More Relevant Evolutionary Models: Integrating an Artificial Genome with a Developmental Phenotype

James Watson[1], Janet Wiles[1,2], and Jim Hanan[3]

[1] School of Information Technology and Electrical Engineering
{jwatson, j.wiles}@itee.uq.edu.au
[2] School of Psychology
[3] Advanced Computational Modelling Centre
jim@maths.uq.edu.au
The University of Queensland, Australia

Abstract. The relationship between the genotype and phenotype of organisms plays a key role in the evolutionary process. While Evolutionary Computation (EC) models have traditionally taken biological inspiration in the design of many key model components (e.g., genetic mutation and crossover, populations under natural selection, etc.), there is a need for more biological input in specifying how a genotype forms a phenotype. There are two powerful theoretical abstractions used in biology for explaining the evolutionary basis of phenotypic development. The first is that there is a sequence of hereditary information (the genotype) passed from one generation to the next. The second is that genes extracted from this sequence interact to form networks of regulation that, when coupled with environmental factors, control the development of an organism (the phenotype). An abstract model of gene regulation exists in the form of the Artificial Genome. This model provides a principled approach to extracting regulatory networks of genes from sequence-level information. L-systems provide a mature framework for modelling developmental phenotypes interacting within environments. This paper takes a step towards integrating these two models, providing a biologically-inspired modelling framework that bridges the chasm between processes occurring in evolutionary timescales, and those occurring within individual lifetimes.

1 Introduction

"*Currently many evolutionary studies are carried out at the molecular and genetic level and also at the population and community level. For making real progress in understanding evolutionary change it is a requisite to integrate these different levels. The current challenge is to link the genetic and functional/selective scenarios. This requires studies that link morphogenetic changes with, on the one hand, genetic changes and on the other hand changes of the form-function relationship of forms and structures.*" Frietson Galis ([1], page 238).

Artificial life simulations can be used to integrate models of processes that occur in distinct timescales, allowing greater understanding of the interaction between these processes. Evolutionary simulations require appropriate models at two disparate timescales; that of evolution (occurring over many generations), and that of individual development and interactions (occurring within a lifetime). Artificial life simulations of evolutionary processes involve models of arbitrarily complex organisms, specified by a genotype and its corresponding phenotype. Selective pressures are applied to the phenotype, which is the resulting artificial organism obtained through an interpretation of the genotype, often as a collection of features directly specified by "genes".

Models are necessarily abstractions of the biological phenomena they seek to emulate. One of the most fundamental abstractions in evolutionary models is the mapping from the genotype to phenotype. Although evolutionary operators that work at the genotypic level (such as mutation and crossover) have been largely inspired by known biological processes, the relationship between genotype and phenotype has been, by and large, entirely arbitrary, since the details of the relationship are not yet fully understood in biology.

In real organisms, there is more to specifying a phenotype than genetic information [2]. The genome interacts with the environment to control a developmental process, and selection acts on the whole phenotype throughout its development. *Lindenmayer* systems, or L-systems (described in Section 2) are a mature and biologically-inspired framework for modelling such developmental phenotypes. L-systems are parallel rewrite grammars that can be used to model organisms with environmental pressures acting on all aspects of the phenotype's development. Increasingly realistic entities can be simulated, including abstract physiologies [3], such as roots, shoots and signalling systems.

In order to evolve L-system models, previous work has used the L-system grammar itself as the genetic encoding, with the rendered plant the phenotype [4]. With this genetic encoding, the crossover and mutation operators were modified so that child grammars were valid L-systems (in particular, the number of opening brackets needed to be balanced against the number of closing brackets). Using the basic L-system string as genetic information meant that the standard crossover operator could swap unrelated parental information (e.g., a segment encoding leaf growth could be exchanged with a segment encoding root formation); poor offspring could often result from the combination of two fit parents. Typed, hierarchical data structures [5] and the use of timed, parametric deterministic, and non-deterministic stochastic L-systems [6] are examples of methods that successfully address this issue, by allowing sequence operators to alter portions of the L-system in a more controlled manner.

While such mappings between genotype and phenotype allow the evolution of artificial organisms for given selective pressures, they are limiting when studying the effects of sequence operators (such as crossover, mutation, etc.) within a biological context. For example, it may be the case that some biological gene duplications result in one functional gene, and one equivalent but unnecessary gene that is free from selective pressure (thus facilitating a random walk of

the phenotypic fitness landscape). Grammar-based genetic encodings cannot be easily used to model the effects of such a process, since every item in the genome encodes for a phenotypic trait. For the same reason, neutral mutations (where a change to the genotype causes no change in the phenotype) are problematic to model.

Biology provides two key abstract theories for use in understanding genetic systems. The first is that a sequence of information (DNA) is the manner by which hereditary information is passed between generations. The second is that regulatory interactions between genes extracted from this sequence control the growth of an organism. A biologically-inspired genetic model, the Artificial Genome, has been recently developed [7] (described in Section 3). While still a very simple model which abstracts away much of the complexity, it encapsulates the notion of regulatory networks of genes being extracted from sequence information.

In order to understand evolutionary processes that interact across distinct timescales, methods of integrating biologically-inspired models of genotypes and phenotypes must be considered. The aim of this study is to explore mappings between L-system phenotype models and the Artificial Genome, providing L-systems with a biologically-plausible genetic foundation and giving the Artificial Genome a functional role that can be placed under selective pressure. By using regulatory interactions between networks of genes extracted via the Artificial Genome model to control the development of an L-system phenotype, we present preliminary findings of the effect sequence-level mutations have on simple phenotypic development.

2 A Developmental Phenotype: L-Systems

L-systems are parallel rewrite grammars capable of specifying and visualizing the morphogenesis of organisms; primarily plants [8–10]. L-systems are comprised of a finite alphabet of values, an axiom (starting condition), production rules for the replacement of string components, and a derivation process which includes the parallel rewriting of strings. For example, a simple L-system might be based on the alphabet a,b. The rule $a \rightarrow ab$ means that at each derivation step, every a is to be replaced with ab, while the rule $b \rightarrow a$ means that every b should be replaced by an a at each step. If the axiom of this L-system is b, the following transformations would occur in five derivation steps (example from [8], page 3):

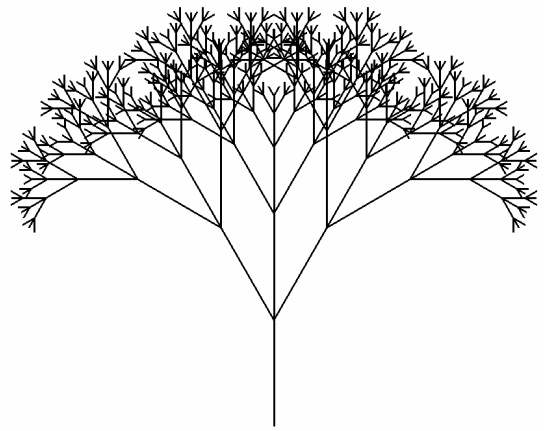
$$
\begin{array}{c}
b \\
\downarrow \\
a \\
\downarrow \\
ab \\
\downarrow \\
aba \\
\downarrow \\
abaab \\
\downarrow \\
abaababa
\end{array}
$$

This parallel development of grammars can be used to visualize the growth of more complex objects such as plants by using "turtle graphic" commands, where a line is drawn to follow the path travelled by the turtle (so named for historical reasons). Simple commands such as 'move forward' (F), 'turn left' (+), etc., are used to direct the turtle's path. The symbols `[` and `]` are included to allow branching structures. `[` pushes the state of the drawing turtle onto a stack, while `]` pops the most recently stacked state and makes it the current state (the reader is again referred to [8] (page 24), for background on bracketed L-systems). More complex language constructs, such as parameter definitions and conditional rewriting, have been built on top of this foundation by tools such as CPFG [11].

The CPFG model of simple tree growth shown in Figure 1 illustrates the development of branching structures using a small number of developmental rules. A distinct advantage of using L-system tools such as CPFG to model developing phenotypes is versatility; much more complex and realistic artificial organisms can be modelled this way, allowing a great variety of phenotypes to be simulated (see Figure 2).

3 A Biologically-Inspired Genotype: The Artificial Genome

The biologically plausible genetic system developed by Reil [7] extracts a Boolean regulatory network of interacting genes from a genotypic sequence string of four bases (implemented as 0, 1, 2, 3). Genes are found by searching the genetic string for the promoter sequence '0101'; once such a sequence is found, the following 6 digits are defined to be the gene's value. Regulation between genes is specified via a form of upstream regulation. The digits in the region between two given genes are defined to be the regulatory region for the latter gene. A gene is 'expressed' by incrementing each of its digits by 1, modulo 4 (the number of bases). The regulatory region of each gene is then searched for matches to any gene products; each match defines a regulatory link between the gene that produced the gene product and the gene whose regulatory region contained the match (see Figure 3).



```
#define STEPS 35
#define SIDEDELAY 4
#define MAINDELAY 6
#define GROWTHRATE 1.1
#define MAXIMUM 10
#define ANGLE 30


Lsystem: 1
derivation length: STEPS
Axiom: A(0)

A(delay) : delay <= 0
    --> F(1)[+(ANGLE)A(SIDEDELAY)][-(ANGLE)A(SIDEDELAY)]A(MAINDELAY)
A(delay) : delay > 0 --> A(delay-1)
F(len) : len < MAXIMUM --> F(len*GROWTHRATE)

endlsystem
```

Fig. 1. A simple tree-shape resulting from an L-System formalism implemented in CPFG. Note that transformations can be conditionally applied at each step via the use of parameters. By starting with the axiom A(0) (apex with no growth delay), a shoot of length 1 is drawn (F(1)), with three apices placed at its top (one ANGLE degrees to the left, one at ANGLE degrees to the right, and the third facing the same direction as the stem). The two apices placed at angles to the stem are given shorter delays to the centre stem. For STEPS iterations, the delay associated with each apex is decremented until it reaches 0, at which point a new shoot is drawn and three new apices are placed at its tip. Also at each step, the length of each shoot is extended proportionally according to GROWTHRATE until it reaches MAXIMUM length.

For evolutionary models, deriving regulatory networks via this approach presents advantages over the direct specification of network structure, such as greater freedom in network mutation and a principled approach to mapping genetic mutation operators to network-level change [12].

4 Mapping from Genotype to Phenotype

In this preliminary study, three parameters controlling phenotypic development (ANGLE, GROWTHRATE and SIDEDELAY) were specified by the Artificial Genome. An initial random sequence string of length 60,000 was used as the base genotype. On average, this genome length generates 225 genes, with an average gene connectivity of 12.475.

In a random genetic sequence, all possible gene values can occur with equal probability. In order to define and control phenotypic properties from the in-

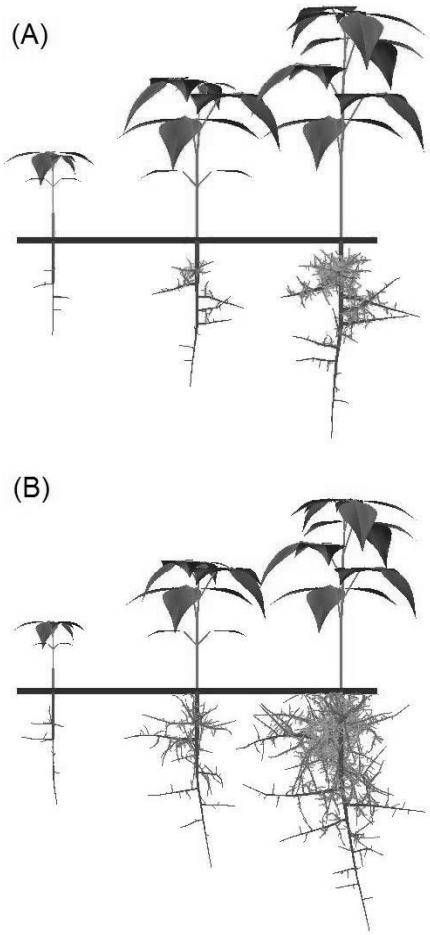


Fig. 2. Three snapshots of the growth of a 3D bean plant L-system model (A) illustrate the power of the formalism to capture the development of realistic phenotypes. Developmental modifications, such as favouring the growth of roots over shoots (B), result in a different path of development.

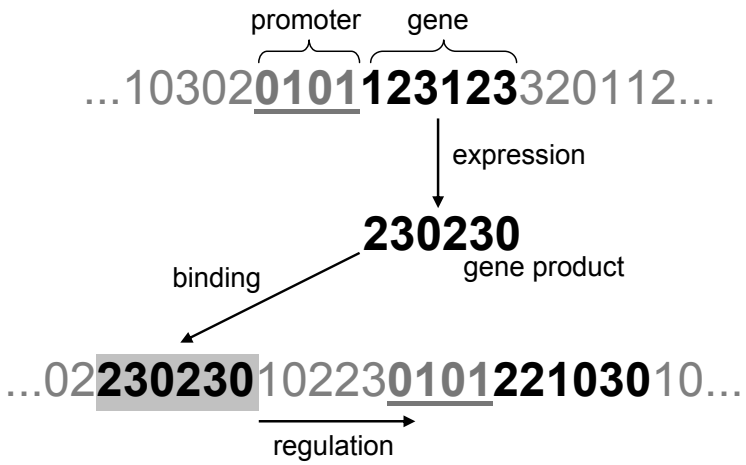


Fig. 3. The Artificial Genome model. Genes are defined as the digits following a promoter sequence '0101'. A gene is expressed to form a gene product by incrementing each gene element by 1, modulo 4 (since the genome only consists of the values 0, 1, 2 and 3). If the gene product of gene A matches a region in gene B's regulatory region (the region between gene B and its preceding gene), then gene A is defined as regulating gene B. In this illustration, the gene of value 123123 regulates the gene of value 221030.

teracting regulations of genes defined by the Artificial Genome, each gene was grouped into various 'classes' according to the sum of its constituent digits. For example, a gene consisting of the value 123123 would be assigned to group 12. Using regulation between such classes rather than between actual genes allows greater flexibility and predictability in the mapping process, since these classes occur at a predictable frequency for random sequences of a given length, and this frequency is unique to each class (see Figure 4).

The rules used to define the three key phenotype parameters are summarised in Table 1. These rules were devised to produce L-System parameter values in an appropriate range, given the random sequence described above.

Table 1. Rules used to define the three phenotypic traits under genetic control

Parameter	Derivation
ANGLE	(The number of times class 8 is regulated) / 6
GROWTHRATE	((The number of times class 9 is regulated) / 3400) + 1
SIDEDELAY	(The number of times class 10 is regulated) / 100

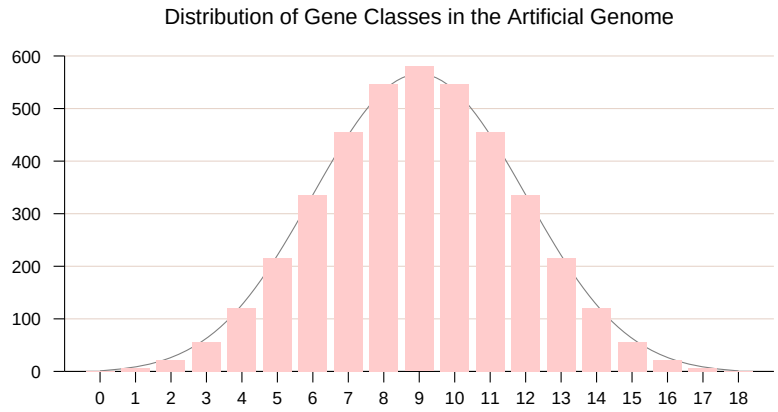


Fig. 4. Distribution of the occurrence of gene classes (x-axis) across all possible gene values in the Artificial Genome, where a gene is comprised of 6 digits.

5 Nature of the Mapping: The Effects of Sequence Mutation

By mutating the genome level of the model and analysing the resulting phenotypic changes, it is possible to study the dynamics of this biologically-inspired mapping. Three types of mutation were performed; single-point (where a single genotypic digit was changed), tandem duplication (where a 256-digit portion of the genotype was copied, and the copy inserted immediately after the original segment), and transposition (where a 256-digit segment was moved to a random location on the genome). The duplication and transposition sizes were chosen so that on average a single gene was affected (genome base of 4 (0,1,2,3) and promoter size of 4 ('0101')).

Any region of the genome that does not encode a promoter, a gene, or a binding site of gene regulation will not effect the regulatory network extracted from the sequence. In addition, any mutations not affecting the regulation of classes 8, 9 or 10 were also neutral. Consequently, a large number of single-point mutations had no effect on the developing phenotype. Any changes that did occur were minimal (see Figure 5). Transpositions had a greater effect on phenotype parameters, as a larger portion of the genome was altered by this operator. By far the biggest phenotypic alteration was caused by tandem duplication. Although the same sized segment of the genome was randomly selected for duplication as for transposition, duplications increased the size of the genome. Accumulated duplications continually added more genes, thus the rates of regulation for classes 8, 9, and 10 also, on average, increased.

Note that phenotypic development was not under selective pressure in this study. The competitive benefits of various sequence-level mutations could be

analysed by extending previous studies exploring the fitness landscapes of early land plants [13–15]. These studies by Niklas investigated the effects of simultaneous selective pressures on simple models of early land plants, but the local search for fit phenotypes was limited to immediately neighbouring morphologies. Consequently, searches for fit phenotypes were unable to jump over fitness valleys. Using the Artificial Genome model with an L-systems phenotype would remove this limitation, and would easily facilitate the study of more complex phenotypes.

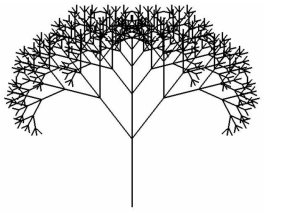
6 Conclusions

As understanding of the mapping from biological genotypes to phenotypes advances at an ever-increasing rate, appropriate computational frameworks that can be built upon and improved will be invaluable theoretical tools. The Artificial Genome provides sufficient flexibility to add additional steps to the network-extraction process. The arbitrary mapping from regulation network to phenotypic model can be easily updated, while the inherent flexibility of the L-systems formalism provides a very capable developmental phenotypic model for a wide variety of model organisms.

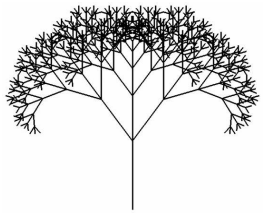
Artificial life has the potential to provide insights into biological processes by integrating across levels of abstraction; each level capturing the essence of one type of biological knowledge. At the level of the genome, the most powerful metaphors include DNA as a sequence of information, and genetic regulation as a network of interacting genes. At the level of the phenotype, it is constant selective forces acting on communities of developing organisms. As mentioned by Galis [1], a holistic approach is required for advancing our understanding of evolutionary processes. By linking a genetic model based on current biological metaphors, and a developmental phenotypic model that can be placed under simultaneous selective pressures, this work is a step in that direction.

7 Acknowledgements

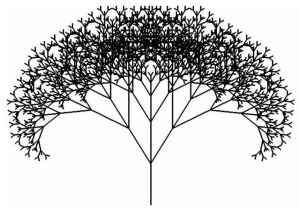
The models in this paper were implemented in the L-Studio front-end to CPFG [11], available at: http://www.cpsc.ucalgary.ca/Research/bmv
J. Watson was supported by a UQGSA.



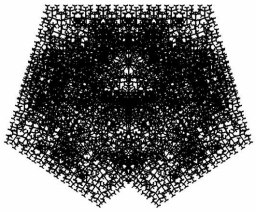
Original phenotype (side delay: **4**, growth rate: **1.07**, angle: **38)**



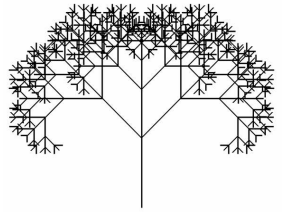
After 100 single-point mutations (side delay: **4**, growth rate: **1.07**, angle: **37)**



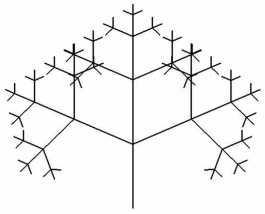
After 1000 single-point mutations (side delay: **3**, growth rate: **1.10**, angle: **32)**



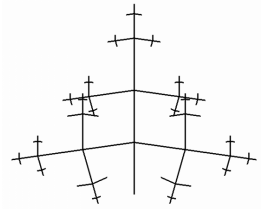
After 10,000 single-point mutations (side delay: **2**, growth rate: **1.11**, angle: **73)**



After 10 duplications (side delay: **4**, growth rate: **1.08**, angle: **45)**

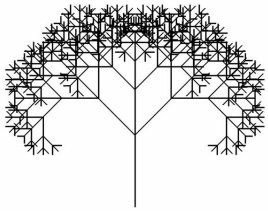


After 50 duplications (side delay: **6**, growth rate: **1.10**, angle: **67)**

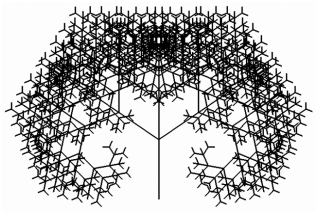


After 100 duplications (side delay: **8**, growth rate: **1.13**, angle: **98)**

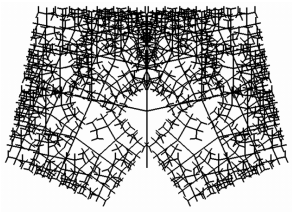
After 1000 duplications (side delay: **125**, growth rate: **3.05**, angle: **1421)**



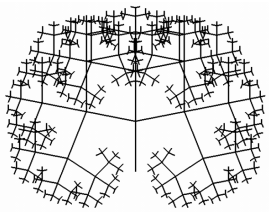
After 10 transpositions (side delay: **4**, growth rate: **1.07**, angle: **45)**



After 50 transpositions (side delay: **3**, growth rate: **1.07**, angle: **56)**



After 100 transpositions (side delay: 3, growth rate: **1.08**, angle: **78)**



After 1000 transpositions (side delay: **4**, growth rate: **1.11**, angle: **79)**

Fig. 5. Phenotypic effects of various sequence-level mutations. Neutral mutations can occur in regions of the genome that do not encode for promoters, genes, or binding sites of gene regulation, or for any portion of the genotype that encodes network components (links or nodes) unrelated to the regulation of classes 8, 9, or 10. Note the level of neutrality of the single-point mutations which was not found with the duplications and transpositions, since those operators affect much larger portions of the genome. One reason for the significant impact duplications had on phenotypic parameters is that the genome size increases with each duplication; greater numbers of genes mean greater rates of regulation as defined by the rules of Table 1.

References

1. Galis, F.: Book Review of Gerd B. Müller and Stuart A. Newman (Eds) (2003). Origination of Organismal Form. Beyond the Gene in Developmental and Evolutionary Biology. Acta Biotheoretica, Vol. 51(3) (2003) 237–238
2. Lemmens, P.: Book Review of Lenny Moss (2003). What Genes Can't do. Acta Biotheoretica, Vol. 51(2) (2003) 141–150
3. Renton, M., Hanan, J., Kaitaniemi ,P.: The Inside Story: Including Physiology in Structural Plant Models. In: Proceedings of Graphite 2003 (2003) 95–102
4. Mock, K.: Wildwood: The Evolution of L-System Plants for Virtual Environments. International Conference on Evolutionary Computation (1998)
5. Jacob, C.: Genetic L-System Programming. In: Davidor, Y., Schwefel, H.P., Reinhard, M. (eds.): Parallel Problem Solving from Nature III. Springer-Verlag, Jerusalem (1994) 334-343
6. McCormack, J.P.: Interactive Evolution of L-System Grammars for Computer Graphics Modelling. In: Green, D.G., Bossomaier, T. (eds.): Complex Systems: From Biology to Computation. ISO Press, Amsterdam (1993) 118–130
7. Reil, T.: Dynamics of Gene Expression in an Artificial Genome - Implications for Biological and Artificial Ontogeny. In: Floreano, D., Nicoud, J.-D., Mondada, F. (eds.): Advances in Artificial Life: 5th European Conference Springer-Verlag, Berlin (1999) 457–466
8. Prusinkiewicz, P., Lindenmayer, A.: The Algorithmic Beauty of Plants. Springer-Verlag, New York (1990)
9. Lindenmayer, A.: Mathematical Models for Cellular Interactions in Development, Parts I and II. Journal of Theoretical Biology, Vol. 18. (1968) 280–315
10. Prusinkiewicz P., Hanan J.S., Mech R.: An L-System-Based Plant Modeling Language. In: Nagl, M., Schürr, A., Münch, M. (eds.): Lecture Notes in Computer Science, Vol. 1779. Springer-Verlag, Berlin (2000) 395–410
11. Prusinkiewicz P., Hanan J.S., Mech R., Karwowski, R.: L-Studio/Cpfg: A Software System for Modeling Plants. In: Nagl, M., Schürr, A., Münch, M. (eds.): Lecture Notes in Computer Science, Vol. 1779. Springer-Verlag, Berlin (2000) 457–464
12. Watson, J., Geard, N., Wiles, J.: Towards More Biological Mutation Operators in Gene Regulation Studies. In Press; To appear in a special issue of BioSystems, Elsevier Science (2003)
13. Niklas, K.: Adaptive Walks Through Fitness Landscapes for Early Vascular Land Plants. American Journal of Botany, Vol. 84(1). (1997) 16–25
14. Niklas, K.: Evolutionary Walks Through a Land Plant Morphospace. Journal of Experimental Botany, Vol. 50(330). (1999) 39–52
15. Niklas, K.: Effects of Hypothetical Developmental Barriers and Abrupt Environmental Changes on Adaptive Walks in a Computer-Generated Domain for Early Vascular Land Plants. Paleobiology, Vol. 23(1). (1997) 63–76