Investigating Ontogenetic Space with Developmental Cell Lineages

Nicholas Geard and Janet Wiles
ARC Centre for Complex Systems
School of Information Technology and Electrical Engineering
The University of Queensland, Brisbane, 4072, Australia
{nic, j.wiles}@itee.uq.edu.au

Abstract

Development plays a significant role in biological evolution, and is likely to prove an effective route to overcoming the limitations of direct genotype-phenotype mappings in artificial evolution. Nonetheless, the relationship between development and evolution is complex and still poorly understood. One question of current interest concerns the possible role that developmental processes may play in orienting evolution. A first step towards exploring this issue from a theoretical perspective is understanding the structure of ontogenetic space: the space of possible genotype-phenotype mappings. Using a quantitative model of development that enables ontogenetic space to be characterised in terms of complexity, we show that ontogenetic landscapes have a characteristic structure that varies with genotypic properties.

Introduction

Development – and its relationship to evolution – is of increasing interest to researchers in both artificial life and biology. Developmental encodings in artificial evolution simulations have been shown to result in more compact, robust and evolvable representations (Stanley and Miikkulainen, 2003). Recent theories in evolutionary developmental biology suggest that reprogramming of developmental pathways is likely to be an important mechanism in morphological evolution (Arthur, 2004).

Development is the transformation of a genotype into a phenotype. Because evolutionary change operates on genotypes, while evolutionary selection operates on phenotypes, the nature of this mapping has the potential to influence evolution. A particular genotype-phenotype mapping can be considered as a point in *ontogenetic space*. Evolution via developmental reprogramming may therefore be regarded as movement through this ontogenetic space. An important implication of developmental reprogramming is that, by imposing structure on the variation available for natural selection to act on, development may bias the direction of evolution (Arthur, 2004).

Developmental systems are complex – while numerous models have been developed, few of these are amenable to quantitative measurement of robustness, variability, evolvability and complexity: all of which need to be investigated if we are to understand both how development interacts with evolution in biological systems, and how developmental encodings can be of use in artificial evolution.

In this paper we propose a general model for simulating developmental systems that enables questions about robustness, variability and evolvability to be addressed. Specifically, it features: a computationally efficient mapping from genotype to phenotype; an intuitive metric for describing the degree of phenotypic similarity, enabling robustness and variability to be quantified; a cell lineage representation of the developmental process, to which ontogenetic complexity metrics can be applied; and tools for the visualization of ontogenetic space. The aim of this study was to explore the behaviour of this model.

The next section of this paper provides background on biological development and the use of cell lineages to represent ontogeny. The developmental lineage modelling framework and ontogenetic complexity measure are then introduced. Ensemble studies are used to characterize the effect of genotype size and connectivity on the composition of ontogenetic space, measured in terms of complexity, and its local structure, measured in terms of robustness and variability. Finally, TreeView, a tool for visualizing ontogenetic space that supports the results of the quantitative studies, is introduced.

Development and Cell Lineages

In biological development, a single egg cell in a suitable environment transforms into a complex, multicellular organism. While contextual information from the environment plays a role in this process, the key locus of ontogenetic control is the information encoded in an organism's genome – specifically, in the network of regulatory interactions between its genes (Davidson et al., 2003). As development proceeds, these genes are expressed in a complex fashion, giving rise to temporal and spatial patterns of cell division and differentiation.

In several species, these patterns of division and differentiation are highly stereotyped and have been precisely measured, providing a detailed insight into the types of control

decision that occur during development. The ontogeny of an organism can be represented by a cell lineage diagram: a binary tree in which the root node represents the initial egg cell, terminal nodes represent the cells that constitute the final phenotype and non-terminal nodes represent intermediate developmental states (Stent, 1998; Geard and Wiles, 2005).

Cell lineages are useful representations of developmental mappings because they describe not only the identities of the cells constituting the final phenotype, but also their individual ontogenies. This record facilitates insight into both the nature of converging developmental pathways and the relationship between genotypic change and phenotypic variability.

The Developmental Lineage Model

Network genotype: A dynamic recurrent gene network (DRGN) model was used as a genotypic representation (Geard and Wiles, 2005). A DRGN is essentially a partially recurrent neural network containing three classes of nodes: input nodes, which receive information in the form of intercellular or environmental signals; regulatory nodes, which generate the complex dynamics from which control emerges; and output nodes, which trigger downstream developmental events such as division and differentiation, but have no direct effect on the regulatory core. The activation of each node is a real-value in the range $[0, 1]$. Connections between nodes also have a real-valued strength, with positive and negative values indicating excitatory and inhibitory relationships respectively and a zero value indicating no interaction. Each network update represents a single cell division cycle, after which the activation levels of each node are updated synchronously according to the logistic sigmoid function.

For the studies reported in this paper, two input nodes were used to specify differences in regulatory input received by each daughter cell due to varying cellular context and/or asymmetric cell division; the number of regulatory nodes (N) was varied between one and 32; five output nodes were used – one controlling whether a cell was dividing or quiescent, the remaining four competing to determine the a cell's fate upon differentiation. In all cases, connections existed between every input node and every regulatory node, and between every regulatory node and every output node (although it was possible for the strength of any of these connections to be zero). Connectivity between regulatory nodes (K) was varied from a single input per node to fully connected.

Development: The developmental process was initialized with a single cell containing a DRGN with all node activations set to zero; after a single network update, if the division output was below a given threshold, division occurred and two new cells were created, each containing an identical copy of the original network; the two input nodes of the left daughter cell were set to one and zero respectively, those of the right daughter cell were set to zero and one; the previous steps were repeated until either all cells were quiescent or a maximum of six cell divisions had occurred. The division threshold was increased exponentially after each division. Ontogeny can be represented by a cell lineage diagram, with each cell identified by the subset of fates to which itself or its dependents is restricted (Figure 1).

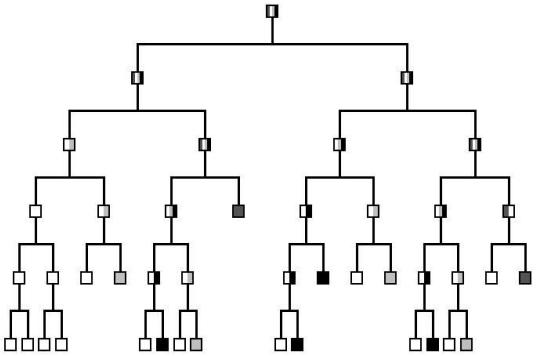

Figure 1: An example cell lineage generated by a DRGN with eight regulatory nodes. Shading of terminal nodes indicates different cell fates. Non-terminal nodes are colored to represent the subset of cell fates they will differentiate into. Color versions of this and other images are available online at `http://www.itee.uq.edu.au/~nic/_treeview`

Phenotype: A phenotype can be described in several different ways. The simplest description consists of the number of cells of each fate present at the end of development (i.e., the terminal nodes of the lineage). More complex descriptions are also possible, incorporating the spatial position of each fate, and the developmental stage at which it appeared (its depth in the lineage). For this study, the phenotype was defined as the ordered set of terminal cell fates. For example, the lineage shown in Figure 1 produces the phenotype AAAAABADABCADDABADABAC (where A = white, B = light gray, C = dark gray and D = black).

Ontogenetic complexity: Characterizing ontogenetic space required a metric for comparing different lineages. One possible metric for comparing developmental mappings is ontogenetic complexity. It has been suggested that the complexity of a developmental mapping may be under selective pressure both to increase and decrease in different evolutionary scenarios (Houthoofd et al., 2003). By characterizing an ontogenetic space in terms of complexity we obtain both insight into one of the gradients across which

a population may move during evolution and a platform from which to assess the balance between robustness and variability that is independent of a specific evolutionary target.

Numerous indicators have been proposed for measuring biological complexity. These measures generally focus either on phenotypic properties, such as organism size, number of cell types and hierarchical levels of organization (McShea, 1996) or ontogenetic properties, based on the notions of entropy and algorithmic complexity (Braun et al., 2003; Azevedo et al., 2005). No single definition or measure of complexity has achieved consensus; rather, different approaches have been applied in a pragmatic fashion.

For these studies we employed an algorithmic approach in which a lineage is transformed into the minimal set of production rules capable of describing it (Azevedo et al., 2005). These rules are of the form $X \rightarrow \{Y, Z\}$, where cell X is a non-terminal cell, and cells Y and Z may either be differentiated terminal cells or further non-terminals. Equivalent rules are then collapsed to produce a non-redundant set. Lineage complexity is defined as the size of this rule set expressed as a proportion of the total number of cell division events.

Using Complexity to Characterize Ontogenetic Space

Dynamic recurrent networks, as a class of computational devices, are known to be capable of a highly flexible range of behaviours. They also display characteristic dynamic properties, such as cyclic and chaotic behaviour, that may bias the distribution of cell lineages they generate. Therefore, it is likely that the probability distribution over the space of possible lineages will be different when the generating function is a recurrent network rather than, for example, a stochastic process.

To obtain an overview of the composition of ontogenetic space, we generated an ensemble of 50,000 DRGNs (eight fully connected regulatory nodes with connection strengths randomly drawn from a Gaussian distribution $G(0,2)$), allowed them to develop and computed their lineage complexity (Figure 2, circles).

Inspection of the lineages generated by this ensemble revealed that a large portion of ontogenetic space consists of two uninteresting cases: (1) the initial cell fails to divide and no development occurs; or (2) the system fails to stop dividing and no differentiation occurs. In between these two extremes a diverse range of systematic and quasi-systematic structures occur (e.g., Figure 1).

This observation revealed an issue with the lineage complexity measure described above – it does not necessarily accord with intuitive notions of what makes an "interesting" lineage. Specifically, the large region of the space inhabited by the uninteresting lineages described above is not reflected by the shape of the lower end of the distribution. For example, a cell that divides once and then stops has, by definition, maximal lineage complexity (i.e., it's one division event has a unique rule, giving a lineage complexity of 1.0) – an anomaly evidenced by the small spike in the upper end of the distribution (Figure 2, circles). As the lineage complexity measure was originally developed for comparing relatively large lineages of equivalent size, this unusual behavior was not previously an issue (Azevedo et al., 2005). However, we were interested in characterizing the entire class of lineages that could be produced by a given system, which includes individual lineages of widely varying sizes. In order to be able to make more principled comparisons, we defined *weighted complexity* as the product of lineage complexity and the number of differentiated cells (Figure 2, crosses).

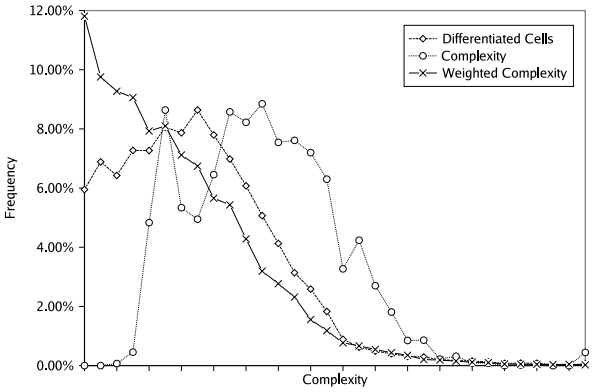

Figure 2: Distributions of differentiated cell number, Azevedo's lineage complexity, and weighted complexity. Note that the x-axis for each of these distributions is scaled as follows: differentiated cell number, [0,64]; lineage complexity, [0, 1]; and weighted complexity, [0,40].

The resulting distribution accords more closely with intuitions about complexity distribution: Uninteresting lineages with a weighted complexity close to zero occupy the lower end of the distribution, the bulk of the landscape consists of moderately complex lineages, and relatively rare high complexity lineages occupy the tail of the distribution.

Global Composition: The Effect of Regulatory Properties on Ontogenetic Space

In order to examine how genotypic properties – size, connectivity and connectivity distribution – affected the composition of ontogenetic space, three sets of random ensembles (Kauffman, 2004) were used:

- *Size:* the size of the regulatory layer was varied ($N = \{1,2,4,8,16,32\}$) for each ensemble, with all samples having full connectivity (each node had N inputs).
- *Connectivity:* the connectivity of the regulatory layer was varied ($K = \{1,2,4,8,16,32\}$) for each ensemble (each node had K inputs), with the size fixed ($N = 32$).

- *Connectivity Distribution:* regulatory size and mean input connectivity were fixed ($N = 32; K_{in} = 4$), but the distribution was varied for each ensemble. Four connectivity distributions were tested: flat – all nodes had *exactly* K inputs; Poisson – all nodes had *on average* K inputs; exponential – input connectivity was exponentially distributed with $\lambda = 2$; duplication-divergence (DD) – network structure was "grown" via successive node duplication and rewiring events (Solé et al., 2002).

Each ensemble consisted of 20,000 DRGNs with connection strengths drawn from a Gaussian distribution $G(0,2)$. The weighted complexity of each lineage was recorded. The number of unique phenotypic sequences generated by each ensemble was also recorded.

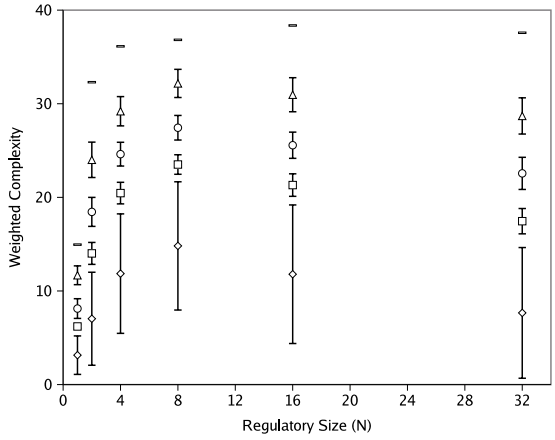

Figure 3: Complexity distributions for the size ensembles. Shown here are the maximum value (bar), and mean and standard deviation for the top 1% (triangle), next 5% (circle), next 10% (square) and whole distribution (diamond).

Due to the skewed nature of the distribution, mean landscape complexity was not the most suitable measure for comparison. Therefore, several additional measures were also calculated: the maximum complexity found and the mean and standard deviation of the top 1% of networks, the next 5% and the next 10%. These values provide a more descriptive indication of how the upper end of the distribution varies. The size ensembles (Figure 3) revealed a sharp increase in all measures as regulatory size increased up to $N = 8$, after which the maximum complexity continued to increase slightly, while the mean complexities began to decrease. In the connectivity ensembles (not shown), average complexities increased rapidly up to $K = 4$, after which they continued to increase, but at a reduced rate. In the distribution ensembles (not shown) little difference was observed between different ensembles, although average complexities for all three of the non-uniform distributions were slightly higher than for the flat case.

Variability – measured by the number of unique phenotypes found in each ensemble – followed a similar trend (Table 1). The number of unique phenotypes found peaked at $N = 8$ in the size ensembles and $K = 8$ in the connectivity ensembles (i.e., a regulatory connectivity density of approximately 0.25). Again, little difference was observed between the distribution ensembles (not shown).

Table 1: Phenotypic variation

Size (K=Full)	Unique Phens.	Conn. (N=32)	Unique Phens
1	1704	1	7639
2	9920	2	12458
4	18112	4	14877
8	19350	8	15235
16	16820	16	13736
32	11952	32	11952

The results of these ensemble studies support previous observations that, once a certain minimum level of regulatory complexity has been achieved, a wide variety of possible phenotypes becomes accessible (Solé et al., 2003). Furthermore, beyond this level, additional increases to regulatory complexity do not lead to corresponding increases in ontogenetic complexity or phenotypic variability and may actually reduce variability.

Local Structure: The Robustness and Variability of Developmental Lineages

The ensemble studies in the previous section provide a basic view of the *composition* of the ontogenetic and phenotypic spaces for different regulatory sizes and connectivities. What they don't provide is a picture of the *local structure* of these spaces – the scale at which much evolutionary change is likely to occur. In this section we use perturbation studies to characterize local structure in terms of *robustness* – the ability to buffer the negative effects of perturbation; and *variability* – the ability to generate variation in response to perturbation.

A developmental system may experience perturbation from two different sources: environmental and genetic (Wagner and Altenberg, 1996). We measured response to perturbation in terms of robustness and variability for the three sets of ensembles described above. Each ensembles consisted of 20,000 samples: these were created by generating 200 random DRGNs and developing 100 perturbed lineages, as described below, for each network. To focus effort on the more interesting regions of ontogenetic space, candidate networks were rejected if the weighted complexity of their unperturbed lineage was below eight (eliminating the uninteresting lineage cases described above). Perturbed lineages were generated in two different ways:

- *Environmental Stability:* Environmental perturbation during development was simulated by adding probabilistic noise to a subset of node activations. After each cell division, each node activation had a 10% chance of being perturbed by Gaussian noise with distribution $G(0, 0.05)$. Each DRGN was developed 100 times under these conditions and the phenotype generated by the perturbed system was compared to that of the original network.

- *Genetic Stability:* Robustness to mutation was tested by perturbing the connection strengths between nodes. Each mutant was generated by adding Gaussian noise with distribution $G(0, 0.1)$ to each connection in the DRGN. As above, 100 mutants were generated from each network and their phenotypes were compared.

The degree of similarity between two phenotypes was defined as the Levenshtein distance (Sankoff and Kruskal, 1983) between the unperturbed fate sequence U and the perturbed fate sequence P, divided by the length of U. Levenshtein distance is defined in terms of the minimum number of transformations required to change U into P, where possible transformations are the insertion, deletion and substitution of fates. A similarity of 1.0 indicated a perfect match. Two measurements of robustness were calculated: the percentage of the 20,000 perturbed systems in each ensemble with a degree of similarity of 1.0, indicating identical lineages (neutral variation); and the percentage of systems with a degree of similarity above 0.9 (nearly neutral variation). In addition, to give an indication of the potential variability that could be unlocked by environmental or genetic perturbation, the number of unique phenotypes generated from each starting point was recorded.

Table 2: Robustness to perturbation

	Environmental		Genetic	
Ensemble	1.0	> 0.9	1.0	> 0.9
$N = 4$	74.97%	99.90%	14.23%	59.45%
$N = 8$	66.12%	96.87%	7.74%	47.62%
$N = 16$	55.70%	90.75%	7.27%	37.77%
$N = 32$	48.52%	81.40%	6.55%	24.51%
$K = 2$	82.61%	98.04%	23.99%	47.97%
$K = 4$	78.27%	95.81%	23.50%	45.97%
$K = 8$	74.10%	93.71%	16.98%	39.73%
$K = 16$	65.33%	89.48%	13.58%	34.70%
$K = 32$	48.52%	81.40%	10.55%	24.51%
Flat	78.27%	95.81%	23.50%	45.97%
Pois	76.71%	96.03%	20.08%	44.94%
Exp	84.98%	98.86%	30.09%	55.07%
DD	76.63%	96.78%	20.17%	46.10%

Table 2 summarizes the results of the ensemble simulations. The three most noticeable trends in these results are: robustness increases as regulatory size decreases; robustness increases as as regulatory connectivity decreases; and robustness is greater for systems with exponentially distributed regulatory inputs. One possible explanation for the first observation is that the perturbations were performed on a "per gene" or "per connection" basis, therefore the total number of perturbations per network increased with the size or connectivity of the network. We ran additional simulations in which perturbations occurred on a "per genome" basis – that is, a fixed number of perturbations were applied across each of the ensembles. As the size and connectivity increased, the relative proportion of nodes and connections that were perturbed decreased, and robustness increased.

To explore the connection between robustness and lineage complexity, we calculated individual robustness values for each starting point in an ensemble and plotted them against the complexity value of the original lineage (Figure 4 shows the results for $N = 16$, full connectivity – other ensembles displayed similar trends). While it was expected that more complex lineages would be significantly less robust, the correlation was weaker than anticipated.

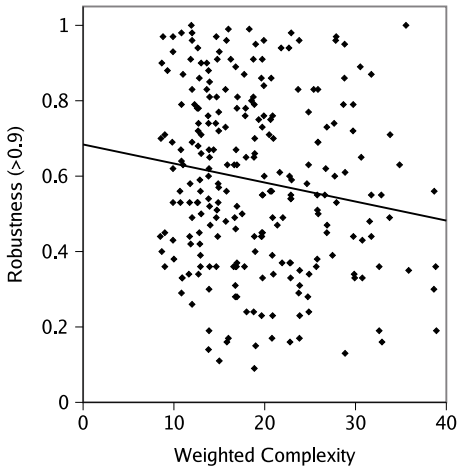

Figure 4: Robustness (proportion of mutants with similarity > 0.9) plotted against weighted complexity for the $N = 16$ ensemble. Each point represents a single network perturbed 100 times. The regression line indicates a very weak negative correlation.

In one sense, variability is the flip-side of robustness: any perturbation whose effect is not buffered by development produces phenotypic variation that can be selected for. The extent to which variability is useful can be estimated by the likelihood that a novel phenotype will confer some adaptive advantage on the evolving system. In the absence of an explicit measure of fitness, we considered that novel phenotypes were more likely to be adaptive if the phenotypic effects of perturbation were small in relation to the size of the system.

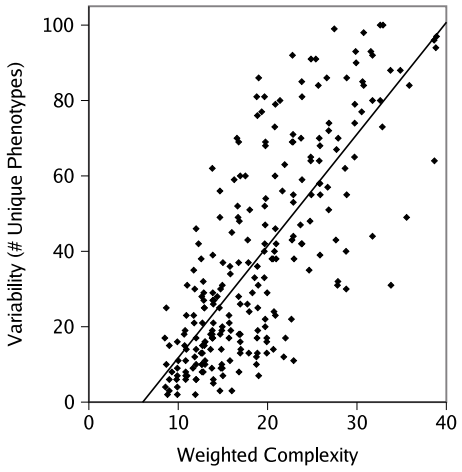

Figure 5: Variability (number of unique phenotypes) plotted against weighted complexity for the $N = 16$ ensemble. Each point represents a single unperturbed network. The regression line indicates a positive correlation.

Our initial expectations, based upon the global composition of the landscape, were that: (a) networks producing more complex lineages would be less robust and more likely to generate variation in response to mutation; and (b) networks producing less complex lineages would be more robust and hence less likely to generate variation in response to mutation. While these intuitions were borne out by the ensembles (Figures 4 and 5), we observed two additional classes of behaviour: (c) robust lineages that also displayed high variability – while seemingly counter-intuitive, because robustness was based on lineages being *very similar* rather than *identical*, it was possible for mutation to result in a wide variety of very similar lineages; and (d) less robust lineages with low variability, suggesting the existence of local peaks in otherwise flat regions of space, from which most mutations lead to a restricted set of possible variant lineages.

These final two characteristics have implications for ontogenetic evolution. High robustness with high variability suggests that in addition to complex peaks in the ontogenetic landscape, there are also complexity plateaus: regions in which a large number of complex lineages with a high degree of similarity are clustered together. Such regions could facilitate refinement and incremental modification. Conversely, regions of low robustness and low variability suggest the possibility of a complexity floor: large regions of simple but unstable lineages with little variation available for selection to act on.

Visualizing Ontogenetic Space

One difficulty with complex developmental systems is that the combinatorial explosion of parameters and variables results in a high-dimensional ontogenetic space that is rather difficult to visualize. To assist in understanding the structure of lineage space, we have developed a visualization tool, TreeView, that enables parameterized slices of space to be represented in a comprehensible format [1]. Figure 6 shows an example of a parameterized slice: to generate this image, a single DRGN (eight fully connected regulatory nodes) was randomly generated and the interaction strengths of this DRGN were modified by a constant scaling factor ranging between 0.1 and 3.5 (the x-axis); a series of lineages were then generated for each value of the scaling factor with the slope of the division threshold value being varied between zero and one (the y-axis).

This heatmap view of a slice of the ontogenetic landscape supports several of the intuitions developed via the quantitative studies: the large black region in the lower left corresponds to lineages which proliferate indefinitely and fail to differentiate (example (a) in Figure 6); lighter regions in the center of the map corresponds to clusters of moderate to high complexity lineages; darker regions in the top right of the map correspond to lineages that only divide a small number of times; the landscape consists of both broad regions of equivalent behaviour and transitional regions in which behaviour changes rapidly.

Conclusions

An evolutionary developmental system consists of a genotypic space, whose structure is determined by the mutation operators used; a phenotypic space, which can be associated with a selective fitness gradient; and an ontogenetic mapping, the properties of which determine how mutations in genotypic space are transformed into phenotypic change.

Our investigations using a DRGN genotype and a cell lineage phenotype suggest that genotypic properties such as regulatory size and connectivity can affect the complexity of the lineages that constitute ontogenetic space, as well as the robustness and variability of these lineages. Further simulations are required to ascertain how these properties influence behaviour in an evolutionary context – specifically, the extent to which the ontogenetic bias arising from the network nature of the genotype can act as an orienting force in evolution. These studies demonstrate that cell lineages, whilst being a minimalist description of ontogeny, can be an elegant and valuable contribution to investigations of development and evolution. They are biologically plausible, computationally simple, quantifiable by several metrics and amenable to visualization.

Acknowledgments

This research was funded by an APA and an ACCS scholarship to NG and an ARC grant to JW.

[1] Additional resources related to TreeView are available from `http://www.itee.uq.edu.au/~nic/_treeview`.

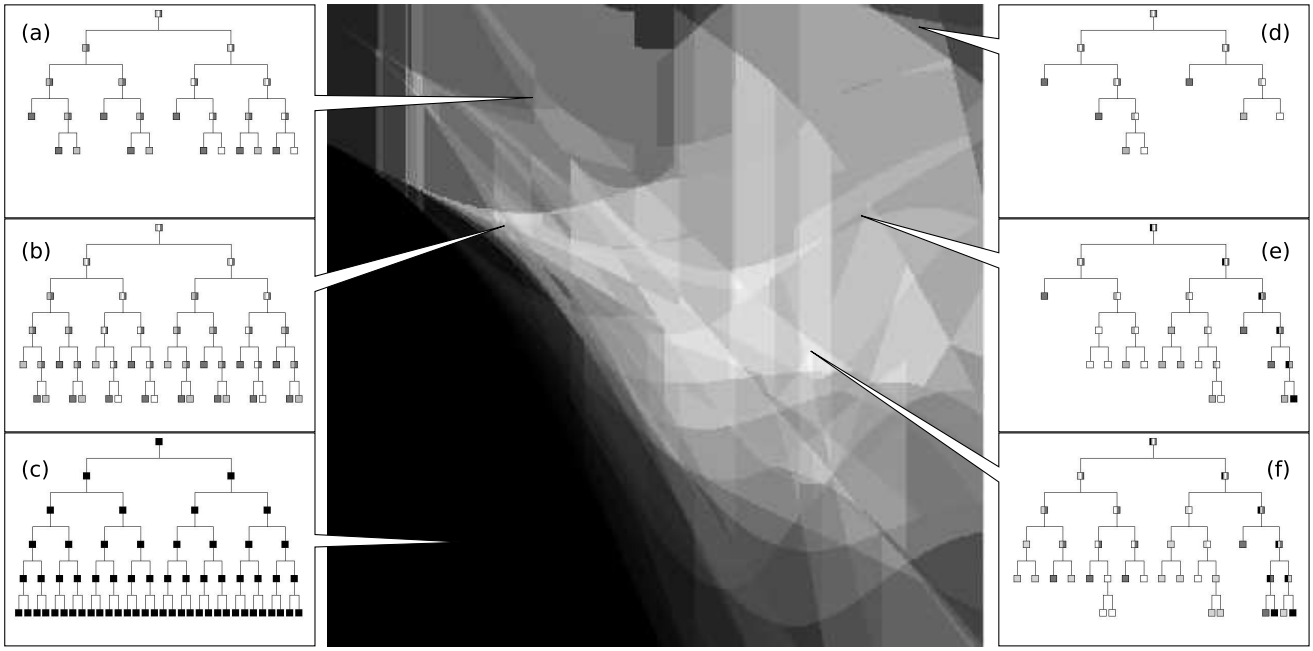

Figure 6: Heatmap representation of ontogenetic space. X-axis: scaling factor applied to interaction strengths; Y-axis: variance in the slope of the division threshold; Shading denotes weighted complexity (black = zero, white = 18). Six sample trees from various regions are shown, with weighted complexity values of (a) 8.47; (b) 11.48; (c) 0.0; (d) 4.67; (e) 12.0; and (f) 15.2.

References

Arthur, W. (2004). The effect of development on the direction of evolution: toward a twenty-first century consensus. *Evolution & Development*, 6(4):282–288.

Azevedo, R. B. R., Lohaus, R., Braun, V., Gumbel, M., Umamaheshwar, M., Agapow, P. M., Houthoofd, W., Platzer, U., Borgonie, G., Meinzer, H. P., and Leroi, A. M. (2005). The simplicity of metazoan cell lineages. *Nature*, 433:152–156.

Braun, V., Azevedo, R. B. R., Gumbel, M., Agapow, P. M., Leroi, A. M., and Meinzer, H. P. (2003). ALES: cell lineage analysis and mapping of developmental events. *Bioinformatics*, 19:851–858.

Davidson, E. H., McClay, D. R., and Hood, L. (2003). Regulatory gene networks and the properties of the developmental process. *Proceedings of the National Academy of Science, USA*, 100(4):1475–1480.

Geard, N. and Wiles, J. (2005). A gene network model for developing cell lineages. *Artificial Life*, 11(3):249–268.

Houthoofd, W., Jacobsen, K., Mertens, C., Vangestel, S., Coomans, A., and Borgonie, G. (2003). Embryonic cell lineage of the marine nematode *Pellioditis marina*. *Developmental Biology*, 258:57–69.

Kauffman, S. A. (2004). A proposal for using the ensemble approach to understand genetic regulatory networks. *Journal of Theoretical Biology*, 231(1):581–590.

McShea, D. W. (1996). Metazoan complexity and evolution: is there a trend? *Evolution*, 50(2):477–492.

Sankoff, D. and Kruskal, J., editors (1983). *Time warps, string edits, and macromolecules*. Addison-Wesley, Reading, MA.

Solé, R., Fernández, P., and Kauffman, S. (2003). Adaptive walks in a gene network model of morphogenesis: insights into the Cambrian explosion. *International Journal of Developmental Biology*, 47:693–701.

Solé, R. V., Pastor-Satorras, R., Smith, E., and Kepler, T. B. (2002). A model of large-scale proteome evolution. *Advances in Complex Systems*, 5(1):43–54.

Stanley, K. O. and Miikkulainen, R. (2003). A taxonomy for artificial embryogeny. *Artificial Life*, 9(2):93–130.

Stent, G. (1998). Developmental cell lineage. *International Journal of Developmental Biology*, 42:237–241.

Wagner, G. P. and Altenberg, L. (1996). Complex adaptation and the evolution of evolvability. *Evolution*, 50(3):967–976.