Directed evolution of an artificial cell lineage

Nicholas Geard[1,2] and Janet Wiles[2]

[1]School of Engineering and Computer Science, University of Southampton
Southampton SO17 1BJ, UK
[2]ARC Centre for Complex Systems, The University of Queensland
Brisbane, Queensland 4072, Australia
nlg@ecs.soton.ac.uk, j.wiles@itee.uq.edu.au

Abstract. Biological development is a complex process that mediates between genotypes, to which mutations occur, and phenotypes, on which selection acts. Properties of development can therefore have considerable impact on evolution. However, in many existing simulation models of development, the developmental process itself is difficult to recover and/or analyse. We have previously introduced a model of development in which the developmental process is represented as a cell lineage. Here we use this model to further explore the control of development, and the influence that development has on shaping an adaptive landscape.

1 Introduction

Novel phenotypic forms arise from gene mutations that reprogram developmental trajectories [1]. Evolution by natural selection occurs because certain individuals, by virtue of some heritable phenotypic trait, stand a better chance of surviving to pass on their genes to offspring than others. The specific phenotypic traits that increase an organism's chance of reproduction will depend on the nature of the ecological niche it inhabits. In a relatively stable environment, it is therefore possible to imagine an adaptive gradient mapped to phenotypic space.

The idea of an adaptive phenotypic space was introduced by Simpson [2], who described a two-dimensional landscape representing the possible combinations of two phenotypic characters in which elevation corresponded to fitness. The highest point in the landscape represents the phenotype that is most adapted to the current environment. Because environments are dynamic, the location of this optimum point will move over time. Simpson's adaptive phenotypic landscape is a descendant of the fitness landscape described by Wright [3] but differs in two respects. First, the axes of Wright's fitness landscape represent gene frequencies rather than phenotypic characters. Second, the structure of fitness landscapes is typically more complex due to epistatic interactions between genes.

There is an important relationship between genotypic and phenotypic landscapes. The adaptive phenotypic landscape specifies the direction of evolution favoured by selection. However, any movement from phenotype A to phenotype B in phenotypic space is contingent upon genotype B being mutationally accessible from genotype A in genotypic space (Figure 1). The mapping from a genotype

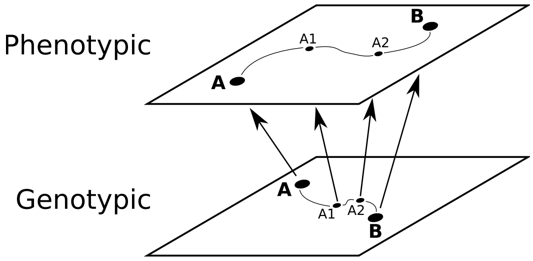

Fig. 1. Phenotypic adaptation depends on mutational accessibility. In order for phenotype adaptation to proceed from phenotype **A** to phenotype **B**, there must be a mutationally accessible path of genotypes between genotypes **A** and **B**. The mapping from genotypic to phenotypic space will be affected by the nature of development.

to a phenotype is defined by the developmental process; therefore, properties of the developmental process will affect adaptation. Determining the impact that development has on adaptive landscapes requires a better understanding of the mapping between genotypic and phenotypic space.

This study explores the effect on evolution of a developmental mapping based on the dynamics of a gene regulatory network. The following section describes the artificial cell lineage model. Two series of simulations are then used to explore the effects of different phenotypic constraints, and different target complexities, on adaptive search difficulty. Finally, the results of these simulations are analysed to provide insight into the characteristics of the adaptive landscape.

2 The artificial cell lineage model

The artificial cell lineage model consists of two components: a network component that generates the gene expression dynamics controlling development and a cell lineage component that defines how these dynamics are interpreted to define an ontogeny. The model is described briefly here; a more thorough description and justification can be found elsewhere ([4, 5]).

The genetic component of the model is defined by a network of interacting nodes, based on a standard recurrent network architecture. Three layers of nodes represent N_I input, N_R regulatory and N_O output genes respectively. All input nodes are connected to all regulatory nodes, all regulatory nodes are connected to all output nodes, and each regulatory node is connected to, on average, K other regulatory nodes (including self connections). The interactions between two network layers are represented by a weight matrix, in which the entry at row i, column j specifies the influence that gene j has on gene i. For the simulations, random networks were created by setting each weight to a value drawn at random from a Gaussian distribution with mean zero and standard deviation W. The state of the network was updated synchronously in discrete time steps, with the activation of node i at time $t+1$, $a_i(t+1)$, given by $a_i(t+1) = \sigma\Big(\sum_{j=1}^{N} w_{ij} a_j(t) -$

$\theta_i\Big)$ where w_{ij} is the level of the interaction from node j to node i, θ_i is the activation threshold of node i, and $\sigma(x)$ is the logistic sigmoid function.

A cell lineage is a record of a developmental trajectory in the form of a binary tree [6]: the root node represents the fertilised egg cell; the non-terminal nodes represent the transient states that cells pass through whilst differentiating; and the terminal nodes represent the final differentiated cells that exist at the end of the developmental process. Therefore, the terminal nodes of the cell lineage that represent an organism's phenotype, and the topology of the tree describes the relationship between all cells that existed at some point during development.

The network model described above is a general purpose computing device. In a developmental system, the computation performed is the transformation of a temporal sequence of contextual inputs into an ordered pattern of cell division and differentiation events. Two input nodes specified the relative position of a cell with respect to its sibling. After division, the activation of these nodes was set to $\{0, 1\}$ in the left daughter and $\{1, 0\}$ in the right daughter. The output nodes were used to determine cell division and differentiation. If the activation of the first output node was above a certain division threshold θ_d, that cell would divide, otherwise it would differentiate. In development, the likelihood of a cell continuing to divide decreases over time. To simulate this, the division threshold was scaled dynamically, according to $\theta_d = 1 - 0.01e^{\lambda d}$ where d was the depth of the current cell and λ was a scaling parameter. Once a cell stopped dividing, the remaining $N_O - 1$ output nodes were used to determine its differentiation type via a 'one-hot' or exclusive encoding scheme.

3 Evolving complex cell lineages

We have previously demonstrated that the artificial cell lineage model is capable of generating a diverse range of ontogenies of varying levels of complexity [7, 8, 4]. The aims of these simulations were to investigate the effect of different phenotypic distance metrics on the type of lineages located by adaptive search, and to investigate how the difficulty of adaptive search increased as phenotypic targets became more complex. The specific targets for the adaptive tasks used in this study are derived from the lineages of the organisms *C. elegans* and *H. roretzi*. The use of targets derived from real lineages is important because we know that they have been evolved once, and hence are of a biologically plausible level of complexity. Defining and measuring biological complexity are difficult issues: a full description of the metric employed here can be found in [5], and further exploration of the complexity of cell lineages can be found in [9].

We make a simplifying assumption that adaptation is occurring in a fixed environment, and the target phenotype is the most highly adapted to that environment. Fitness was calculated in terms the distance between the current and target phenotypes. In a real environment, ecological niches are highly dynamic, changing as environments change or according to fluctuations in co-evolutionary relationships. However, when environmental change is slower than adaptation, the assumption of a static fitness landscape is not implausible.

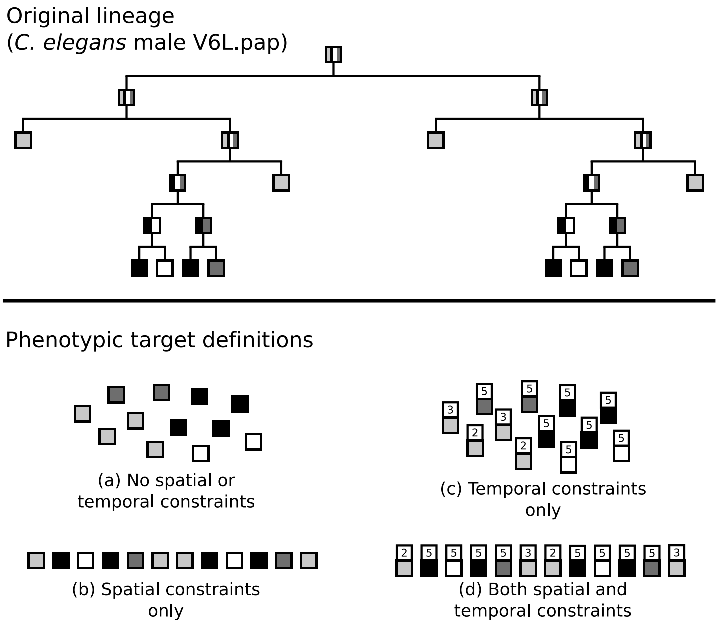
Original lineage
(*C. elegans* male V6L.pap)
Phenotypic target definitions
(a) No spatial or temporal constraints
(c) Temporal constraints only
(b) Spatial constraints only
(d) Both spatial and temporal constraints
3
5
5
5
5
3
2
5
5
2
5
5
2
5
5
5
5
3
2
5
5
5
5
3

Fig. 2. The four phenotypic distance metrics as applied to the *C. elegans* male V6L.pap lineage [10]. See text for a full description of each metric.

3.1 Measuring fitness

Cell lineages are an organisational, rather than morphological, description of a phenotype and can be quantified and compared in an automated fashion. We defined four metrics based on the phenotypic component of a cell lineage (*i.e.,* the terminal cells) in terms of the intersection between three types of constraint: on the set of cell identities, the relative spatial location of each cell, and the point in developmental time at which they appear. The first and most basic constraint is on the cell fate distribution: the requirement that a certain number of cells of each specific type are present at the end of development. The second and third constraints require that each terminal cell be correctly positioned in relation to the other cells in the phenotype, and appear at the correct time during development. We do not suggest that natural selection acts to explicitly satisfy these constraints, but rather that they may serve as surrogates for a broad range of selective criteria operating on development.

For each of the fitness metrics used in this study, the identity constraint was considered fundamental and always used, in addition to which temporal and spatial constraints could be applied either separately or together. The practical implication of each of these constraints and their intersection is illustrated in Figure 2. In each case, the fitness metric is applied to the terminal cells of the fully developed cell lineage.

Table 1. Performance of walks using different phenotypic distance metrics

Temporal Constraint	Spatial Constraint	Perfect Runs (of 500)	Unique Lineages
No	No	499	496
No	Yes	288	103
Yes	No	201	113
Yes	Yes	27	1

No temporal or spatial constraints. When there were no temporal or spatial constraints, a phenotype was considered as an unordered set of cell fates and the fitness $f(C,T)$ of the current cell fate set C with respect to the target cell fate set T was defined as: $f(C,T) = \big(|(C \cap T)| - |(C \ominus T)|\big)/|T|$ where $|T|$ is the size of set T, $C \cap T$ is the intersection of sets C and T and $C \ominus T$ is the symmetric difference of sets C and T.

Temporal constraints only. When temporal constraints were used, each cell fate was tagged with its depth in the lineage and preceding equation was used to calculate fitness.

Spatial constraints only. When spatial constraints were used, a phenotype was considered as an ordered sequence of cell fates and the fitness $f(C,T)$ of the current cell fate sequence C with respect to the target cell fate sequences T was defined as: $f(C,T) = \big(\textsc{Lev}(C,T)\big)/|T|$ where $\textsc{Lev}(C,T)$ was the Levenshtein distance between sequences C and T (see [5] for the algorithm used to calculate this metric) and $|T|$ was the length of sequence T.

Both temporal and spatial constraints. When both temporal and spatial constraints were used, each cell fate was tagged with its depth in the lineage and the preceding equation was used to calculate fitness.

3.2 Comparison of different phenotypic distance metrics

An initial set of adaptive walks compared the effect of using the four different phenotypic distance metrics described above as fitness measures. For each metric, an ensemble of 500 networks with eight fully connected regulatory nodes ($N = 8, K = 8, W = 2.0$) were created and adaptive walks were performed. Each adaptive walk consisted of 20,000 steps; at each step, a new network was created by replacing one weight at random with a new value drawn from a Gaussian distribution with mean zero and standard deviation W. The newly created network replaced the current network if its fitness was equal to or greater than that of the current lineage.

As anticipated, as phenotypic definition became more constrained, the difficulty of the search process increased (Table 1). With no spatial or temporal constraints, only one of 500 walks failed to find a perfect solution (*i.e.,* a cell lineage whose terminal nodes consisted of the correct quantity of each cell type). In contrast, with both spatial and temporal constraints, only 27 of 500 walks were able to find lineages that produced the target phenotype. When the phenotypic

Table 2. Target Lineage Details

Lineage	Number of	Number of	Maximum	Weighted
	Cells	Cell Types	Depth	Complexity
C. elegans maleV6Lpap	12	4	5	6.55
C. elegans C	48	4	6	11.23
C. elegans MSp	46	5	7	22.49
C. elegans MSa	48	5	7	26.55
H. roretzi (half)	55	7	6	31.57

definition incorporated either spatial or temporal constraints, around half of the runs found lineages that produced the target phenotype. Spatial constraints were moderately easier to satisfy than temporal constraints (288 compared to 201 perfect solutions).

The phenotypic definition had a significant effect on the variety of lineages that were found. Of the 499 solutions found with no spatial and temporal constraints, 496 of the lineages generating these phenotypes were unique. In contrast, the intersection of spatial and temporal constraints restricted the space of possible solutions to a single lineage, that of the original data set. One explanation for the lower rate of success under this phenotypic definition appears to be the structure of the adaptive landscape. Using the least constrained phenotypic definition means that a greater number of lineages map to the target phenotype, and hence a larger proportion of genotypic space maps, via ontogeny, to a perfect fitness value. When the most constrained phenotypic definition is used, only a single lineage maps to the target phenotype, and hence a much smaller proportion of genotypic space maps to a perfect fitness value.

3.3 Comparison of different phenotypic targets

The second series of adaptive walks compared the performance of adaptive walks on five target lineages derived from real data sets (Table 2). The first target lineage was the *C. elegans* male V6L.pap used above (shown in Figure 2). Three further target lineages from *C. elegans* were also used: the sublineage of the C founder cell, which produces the muscle and epidermis cells in the posterior region of the worm's body; and two sublineages, MSa and MSp, of the MS founder cell, which primarily produces the pharynx (a digestive organ), but also some muscle cells and the somatic gonad precursors [11]. The final target lineage was taken from the ascidian *H. roretzi* [12].

For each phenotypic target, an ensemble of 50 random networks ($N = 16, K = 16, W = 2.0$) was generated and adaptive walks were performed as above. The second phenotypic definition (spatial constraints only) was used to evaluate the fitness of each phenotype. Each adaptive walk consisted of 60,000 steps.

The results of these simulations demonstrate that adaptive search becomes more difficult as the complexity of the target lineage increases (Table 3). While almost half of the walks were able to locate the simplest lineage (*C. elegans*

Table 3. Performance of walks using targets of varying complexity

Target Lineage	Best Fitness	Remaining Errors	Avg. Fitness (Std. Dev.)	Perfect Runs (of 50)
C. elegans maleV6Lpap	1.0	-	0.938 (0.071)	24
C. elegans C	1.0	-	0.950 (0.038)	6
C. elegans MSp	0.956	3	0.852 (0.068)	-
C. elegans MSa	0.958	3	0.834 (0.076)	-
H. roretzi (half)	0.982	1	0.745 (0.074)	-

maleV6Lpap), the best performing walk on the most complex lineage (*H. roretzi*) contained a single incorrect cell after 60,000 steps. In order to demonstrate that the MSp, MSa and *H. roretzi* tasks were in fact achievable, the best performing networks on each of these targets were re-run with no limitations on the maximum length of the walk. At least one walk was able to locate each of the target lineages; however the search times required were on the order of 300,000 steps.

3.4 Analysis of an adaptive walk

The progress of an adaptive walk towards a target may be measured in several ways (Figure 3 shows the first 60,000 steps of a successful adaptive walk using the *H. roretzi* target, after which only one incorrect terminal cell remained).

Fitness followed a hyperbolic trajectory over the duration of the walk; a pattern commonly observed in evolution under both computational and *in vitro* conditions [13].

Complexity (as defined in [5]) tended to increase over the course of the adaptive walk, achieving the complexity of the target lineage after approximately 7,000 steps and thereafter fluctuating about that value. Comparing the fitness and complexity plots, it is evident that there is a degree of neutrality in the mapping from ontogenetic space (measured by complexity) to the fitness landscape. Clearly it is possible for multiple lineages to share an equal fitness value, and for an adaptive walk to move between these equivalent lineages via mutation.

Genotypic substitution rate measures the acceptance of newly created networks. Initially, around 60% of mutations are accepted (*i.e.,* are either beneficial or neutral). This probability decreases at a constant rate until around step 7,000. After this point, approximately 20% of mutations are accepted with a moderate decrease over the remainder of the walk. Should this statistic ever reach zero, it is possible that no further adaptation could occur as the network weights would be so finely tuned that any mutation would be detrimental. In practice, this phenomenon was never observed in any of the simulations reported here: there was sufficient neutrality in the gene network to lineage mapping to ensure that some change was possible.

Phenotypic substitution rate measures the acceptance of networks that generated a different phenotype to the previous network. Initially, around 10% of accepted networks generate different phenotypes. This probability decreases

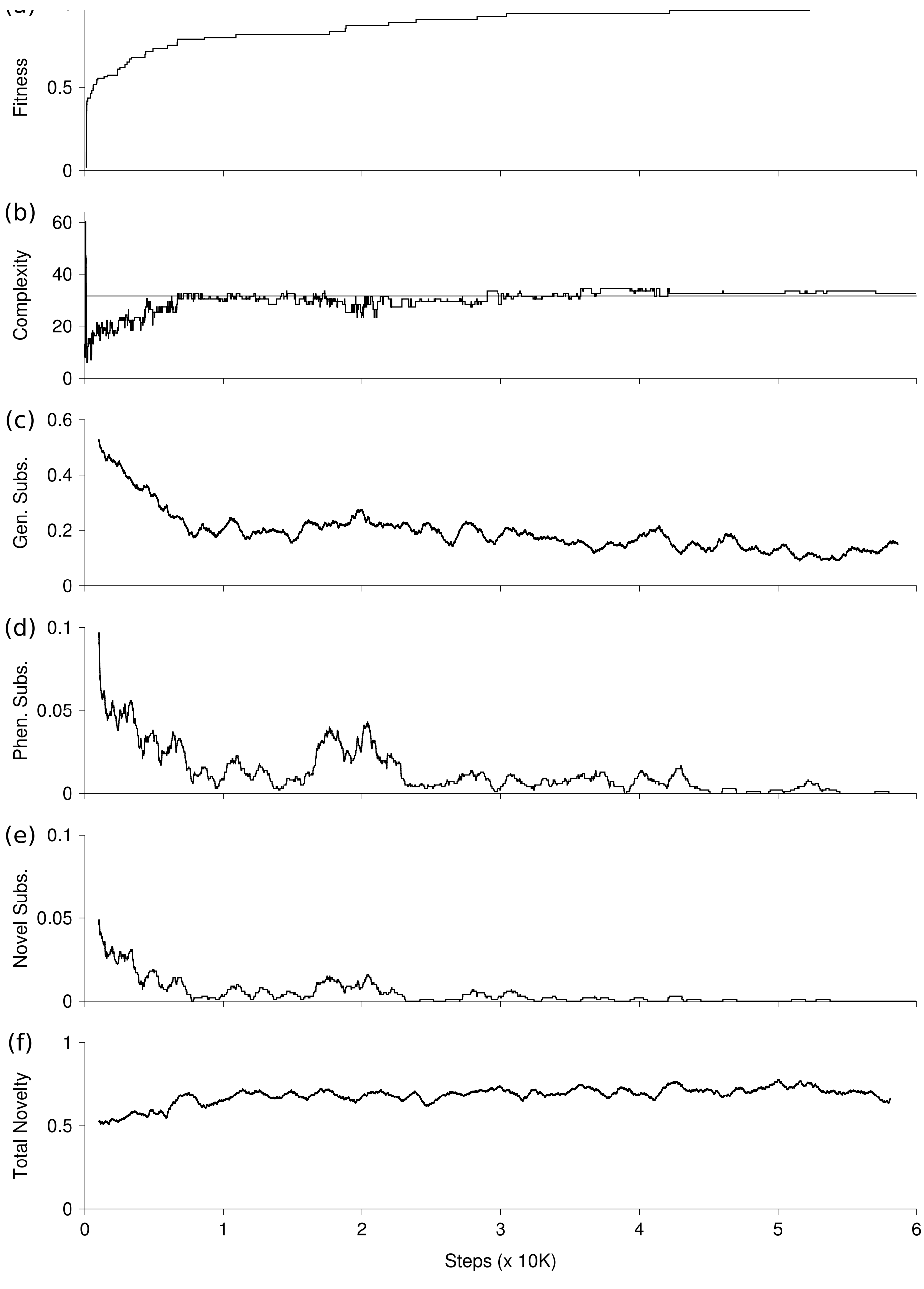

Fig. 3. Analysis of a single adaptive walk using the *H. roretzi* target lineage. From top to bottom, the plots show: (a) fitness; (b) complexity; (c) genotypic substitution rate; (d) phenotypic substitution rate; (e) accepted phenotype novelty rate; (f) generated phenotype novelty rate. See text for further details.

to almost zero after approximately 10,000 generations and thereafter fluctuates. Towards the end of the adaptive walk, the probability of phenotypic substitution falls to zero. The discrepancy between the probability of genotypic and phenotypic substitution can be explained by the degree of neutrality in the mapping from genotypic to phenotypic space: while a relatively constant number of mutations are accepted throughout the adaptive walk, the proportion of these that result in phenotypic change decreases.

Accepted phenotype novelty rate measures the acceptance of networks that generated a previously unseen phenotype. Again, a rapid initial decrease was followed by a gradual decrease to zero as the adaptive walk proceeded. Given the many-to-one mapping from genotypic to phenotypic space, it is possible that a previously seen phenotype could be rediscovered from an entirely different position in genotypic space. This rediscovery could therefore be advantageous if the new genotype responsible is located in a more promising region of genotypic space—one in which the mutationally accessible ontogenies result in more fit phenotypes.

Generated phenotype novelty rate: measures the generation of novel phenotypes by a newly created network, irrespective of whether its fitness is better than, equal to or worse than the current best. Phenotypic discovery remained high (above 50%) over the entire duration of the adaptive walk. This constant rate of discovery suggests that, while more accurate lineages do become harder to find, it is not due to the potential diversity of the system being exhausted. Novel phenotypes continue to be generated; however, the vast majority of these are less fit than the current best phenotype.

4 The relationship between evolution and development

The ontogenetic mapping results in multiple types of neutrality Two types of neutrality were observed to affect the adaptive exploration of genotypic space. The first is in the mapping from genotype and ontogeny. There are many different combinations of network weights that produce identical cell lineage trees. This neutrality accounts for the high rate of genotypic substitution observed in the adaptive walks (Figure 3(c)).

The second type of neutrality is in the mapping from phenotype to fitness. Considering for a moment just the spatially constrained phenotype definition: a mutation which swaps the identities of two incorrect terminal cells in such a way that they are still incorrect will produce a novel phenotype without any change in fitness. The adaptive walks revealed that phenotypes were frequently substituted while on a plateau of neutral fitness (Figure 3(a) and (d)). For example, during one long period of stasis (approximately steps 1200–1800) there was considerable neutral substitution until, around step 1800, a burst of novel phenotypic substitution resulted in further fitness increases. Two interpretations of this dynamic are possible. First, the neutrality may have been beneficial, as it allowed search to continue where it would otherwise have become trapped at a local optima. Second, the neutrality may have been a hindrance, introducing a

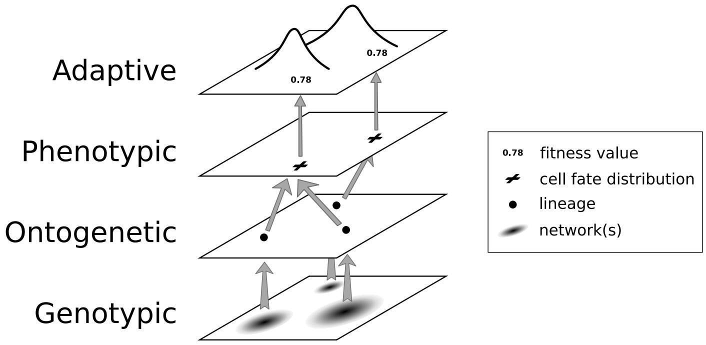

Fig. 4. Summary of different types of neutrality affecting adaptive search. Many different genotypes map to a single ontogeny. More than one ontogeny may map to a given phenotype. Finally, multiple phenotypes have equivalent fitness values.

long period of drift where a more rapid transition to a more fit phenotype could otherwise have been achieved. Distinguishing between these two possibilities is difficult, as it implies a comparison with a search landscape that lacks neutrality, but is otherwise identical.

A third type of neutrality—in the mapping from ontogeny to phenotype—is known to be possible: In the first series of adaptive walks, under all but the strictest set of phenotypic constraints, multiple lineages were located that mapped to the target phenotype. In practice, none of the adaptive walks in the second series were observed to exploit this form of neutrality. One possible explanation for this is that these neutral lineages are located at some distance from one another with respect to genotypic space, such that they are not mutationally accessible to one another. Figure 4 summarises the different types of neutrality that were observed or inferred from the adaptive walks.

Phenotypic improvements occur across a range of scales Analysis of the accepted mutations over the adaptive walk shown in Figure 3 revealed that mutations can cause phenotypic improvement across a wide range of scales. At the lower end of the spectrum were those frequently accepted mutations that modified the identity of a single terminal cell, and those that added or removed a single cell. At the upper end of the spectrum were those more rarely accepted mutations that introduced or removed a new cell type, and those that added or removed an entire branch of the cell lineage. The size of a phenotypic improvement was estimated by applying the fitness function using the pre-mutation lineage as the current solution and the post-mutation lineage as the target solution. The sizes of such changes follow a power law distribution (Figure 5).

The scale of evolutionary change is a subject of ongoing debate in evolutionary biology [14]. The essence of the debate concerns how to explain the evolution of species as inferred from the fossil record: is the selection of individual muta-

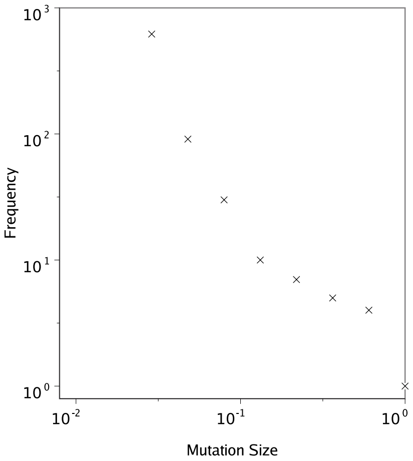

Fig. 5. The distribution of phenotypic improvements indicates that beneficial mutations occur across a range of scales. Mutation size was measured as the distance between the initial and mutant lineages for each accepted mutations and sorted into exponentially scaled bins.

tions a sufficient mechanism, or are higher-level evolutionary forces necessary? Fisher [15] argued that mutations of large effect would be far less likely to be beneficial, and hence only mutations of small effect were likely to be significant. Kimura [16] challenged this claim, pointing out that if very rare large beneficial mutations *did* occur, they would be more likely to be fixed, and hence the distribution of mutation sizes would be skewed. The distribution observed in Figure 5 supports the claim that mutations causing both large and small phenotypic changes can be accepted during an adaptive walk.

Figure 5 also highlights one of the benefits of the gene network approach to modelling ontogeny. If the cell lineage representation had been modified directly by the adaptive process, we would have needed to specify the sizes and types of mutations that were possible (*e.g.,* swapping sublineages, or adding and deleting terminals) As it is, we did not need to impose a preconceived step size on the adaptive process—it emerged naturally as a consequence of the dynamic mapping.

5 Conclusions

Our investigations demonstrate that adaptive search is capable of locating networks whose dynamics generate specific complex developmental patterns derived from real cell lineages. Search difficulty is affected both by the types of constraint (spatial and temporal) applied to the phenotypic targets, as well as their complexity. The dynamics of search suggest that the adaptive landscapes resulting from the proposed developmental mapping are dominated by the presence of several different levels of neutrality.

Acknowledgements The authors would like to thank R. B. R. Azevedo, R. Lohaus and K. Willadsen for useful discussions. This research was supported by an APA and an ACCS scholarship to NG and an ARC grant to JW.

References

1. Arthur, W.: The concept of developmental re-programming and the quest for an inclusive theory of evolutionary mechanism. Evolution & Development **2** (2000) 49–57
2. Simpson, G.G.: Tempo and Mode in Evolution. Columbia University Press, New York, NY (1944)
3. Wright, S.: The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proceedings of the 6th International Congress on Genetics **1** (1932) 356–366
4. Geard, N.: Artificial Ontogenies: A Computational Model of the Control and Evolution of Development. PhD thesis, School of Information Technology and Electrical Engineering, The University of Queensland (2006)
5. Geard, N., Wiles, J.: LinMap: Visualising complexity gradients in evolutionary landscapes. To appear in a special issue of Artificial Life (2007)
6. Stent, G.: Developmental cell lineage. International Journal of Developmental Biology **42** (1998) 237–241
7. Geard, N., Wiles, J.: A gene network model for developing cell lineages. Artificial Life **11**(3) (2005) 249–268
8. Geard, N., Wiles, J.: Investigating ontogenetic space with developmental cell lineages. In L. M. Rocha *et al.*, ed.: Artificial Life X, Cambridge, MA (2006) 56–62
9. Lohaus, R., Geard, N., Wiles, J., Azevedo, R.B.R.: A generative bias towards average complexity in artificial cell lineages. Proceedings of the Royal Society of London, Series B **274**(1619) (2007) 1741–1750
10. Braun, V., Azevedo, R.B.R., Gumbel, M., Agapow, P.M., Leroi, A.M., Meinzer, H.P.: ALES: cell lineage analysis and mapping of developmental events. Bioinformatics **19** (2003) 851–858
11. Sulston, J.E., Schierenberg, E., White, J.G., Thompson, J.N.: The embryonic cell lineage of the nematode *Caenorhabditis elegans*. Developmental Biology **100** (1983) 64–119
12. Nishida, H.: Cell lineage analysis in ascidian embryos by intracellular injection of a tracer enzyme. III. Up to the tissue restricted stage. Developmental Biology **121** (1987) 526–541
13. Lenski, R.E., Travisano, M.: Dynamics of adaptation and diversification: a 10,000 generation experiment with bacterial populations. Proceedings of the National Academy of Science, USA **91** (1994) 6808–6814
14. Leroi, A.M.: The scale independence of evolution. Evolution & Development **2**(2) (2000) 67–77
15. Fisher, R.A.: The Genetical Theory of Natural Selection. Clarendon Press, Oxford (1930)
16. Kimura, M.: The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge (1983)