Developmental motifs reveal complex structure in cell lineages

Nicholas Geard*, Seth Bullock

School of Electronics and Computer Science
University of Southampton
Southampton, SO17 1BJ, UK
{nlg,sgb}@ecs.soton.ac.uk

Rolf Lohaus, Ricardo B. R. Azevedo

Department of Biology and Biochemistry
University of Houston
Houston, Texas 77204-5001, USA
{rlohaus,razevedo}@uh.edu

Janet Wiles

School of Information Technology and Electrical Engineering
The University of Queensland
St Lucia, Queensland 4072, Australia
j.wiles@itee.uq.edu.au

*Corresponding author. Tel: +44 (0)23 8059 4470. Fax: +44 (0)23 8059 9179

Abstract

Many natural and technological systems are complex, with organisational structures that exhibit characteristic patterns, but defy concise description. One effective approach to analysing such systems is in terms of repeated topological motifs. Here, we extend the motif concept to characterise the dynamic behaviour of complex systems by introducing developmental motifs, which capture patterns of system growth. As a proof of concept, we use developmental motifs to analyse the developmental cell lineage of the nematode *Caenorhabditis elegans*, revealing a new perspective on its complex structure. We use a family of computational models to explore how biases arising from the dynamics of the developmental gene network, as well as spatial and temporal constraints acting on development, contribute to this complex organisation.

Keywords: cell lineages, development, gene regulatory networks,generative bias, motifs

1 Introduction

Many natural and technological systems—cellular networks, human language and societies, communication networks—exhibit structures and behaviours that are, in some fashion, organised [1–8]. While the precise details of each system's structure differ, certain topological features, such as feedback loops and hierarchies, appear in a variety of contexts [9]. Despite the presence of these recurring patterns, the organisation of these systems is not simple and their global structures resist concise description [10]. Evidence suggests that the structure of complex systems has implications for their functional properties, such as robustness and flexibility [11]. An important goal is therefore to untangle the relationship between a system's structure, dynamics and functional behaviour. A first step towards this goal involves characterising the complex structures that systems exhibit and identifying their origins.

Complex systems are typically not spontaneous assemblies of disjoint components; rather, they grow and unfold according to the dynamics of some generative process, operating in the context of the system's local environment. For example, an organism's morphology is a result of chemical and genetic processes taking place within cells, direct interactions between neighbouring cells, and chemical signalling between distant cells [12]; language competence arises as a product of learning mechanisms operating within a language community [13]; and the structure of the internet has evolved via a set of socio-technological growth mechanisms [1, 14]. The generative pro-

cesses responsible for producing complex systems can bias the range and type of structures that are observed [15–18].

Efforts to analyse the growth of complex systems are complicated by the fact that more recent structures tend to overwrite older structures, leaving little record of growth patterns. One class of complex biological system for which we do have rich data sets is nematode development. The developmental trajectories of several nematodes, such as *Caenorhabditis elegans*, are invariant and have been mapped in considerable detail in the form of cell lineages [19–22]. A cell lineage is a schematic representation of a developmental process that describes the ancestry of all cells generated during an organism's development in terms of patterns of division and differentiation events. Cells are positioned in a lineage according to their division orientation; by convention, cells dividing in the anterior, left or dorsal directions are positioned to the left of cells dividing in the posterior, right or ventral directions. Thus, while cell lineages omit precise details of developmental morphology, they retain a clear record of the genealogical relationship between cells.

A notable feature of the *C. elegans* cell lineages is its complex topology: cells of a particular type are distributed throughout the various sublineages, while any one sublineage can contain multiple cell types. Upon mapping the cell lineage, Sulston concluded that "the assignment of cell function follows certain broad rules to which there are numerous exceptions" [22]. The extent to which nematode cell lineages can be accounted for by a surprisingly small set of rules has since been revealed [23]. However, a clear understanding

of how to describe and account for the organisation of cell lineages remains elusive. Some features will surely be the result of selective pressures; however, others may emerge from the intersection of biases and constraints operating on the developmental system.

In this paper we introduce an analytic tool, *developmental motifs*, that provide a novel perspective on the relationship between generative processes and cell lineage topology. We build upon the concept of network motifs: the "recurring, significant, patterns of interconnections" observed in a variety of complex networks [9, 24]. Network motifs were introduced to reveal patterns of meso-level structure in complex networks. By analogy, developmental motifs are the repeated topological patterns that occur in lineages. In the context of cell lineages, motifs represent patterns of growth, rather than patterns of structure, and enable us to quantify the extent to which a lineage is regular or random across multiple organisational scales.

Developmental motifs, together with their application to cell lineages, are described in the following section. As a proof of concept, we use developmental motifs to analyse the cell lineages of *C. elegans* and related species, revealing the presence of a broad distribution of motif frequencies. We then use a suite of computational models to explore the role that generative biases and contextual constraints play in shaping the topology of cell lineages.

2 Developmental motifs

A cell lineage represents a developmental trajectory in the form of a binary tree. The root node of the tree represents the fertilized egg cell, the non-terminal nodes represent the transient states that cells pass through while differentiating, and the terminal nodes represent the final differentiated cells. The topology of the tree describes the genealogical relationship between all of the cells that existed at some point during development. We propose developmental motifs as a tool for describing this topology.

A developmental motif is a rooted binary tree of depth d, where d is typically small with respect to the depth of the entire cell lineage. Each leaf node of the motif is labelled as either terminal or non-terminal, corresponding to its status in the original lineage. The set of d-motifs consists of all possible motifs of depth d. For example, the set of 1-motifs contains only two members—a terminal node and a non-terminal node—while there are four possible 2-motifs and twenty-four possible 3-motifs (Figure 1). Each cell in a lineage that is at least $d-1$ cell divisions away from a terminal cell can be associated with a d-motif. The d-motif profile of a lineage is a frequency distribution over d-motifs appearing in that lineage (Figure 2). By extension, the d-motif profile of an ensemble of lineages is the frequency distribution of d-motifs appearing in all lineages in that ensemble. Taken as a whole, the distribution of profile sizes over motif depth (d) provides a signature of topological regularity of a lineage (or ensemble of lineages) across multiple

scales. For example, the profiles of very regular lineages would be expected to contain few distinct motifs, even at greater depths, while those of less regular lineages would display greater diversity.

While focusing here on cell lineages, we also recognise the presence of tree-like organisation in other complex systems, such as phylogenetic trees [25] and linguistic structure [26]. In other domains it may be appropriate to consider motifs that are n-ary, rather then binary, trees; however, the general principles of the approach remain valid.

3 Motif profile of *C. elegans* and other nematodes

What does the developmental motif profile of a real organism look like? The *C. elegans* hermaphrodite consists of 671 cells at hatching, and has a complex topology. Critical events during the first few cell divisions establish well-characterised sublineages that display modular and recursive patterns: cells of any one type are distributed throughout the various sublineages, while any one sublineage can contain cells of multiple types. [22, 23].

We computed the 3-motif profile of the *C. elegans* lineage, revealing a heavy-tailed distribution (Figure 3). Of the 24 possible motifs, 21 are present, but most occur infrequently, with the four most frequent 3-motifs accounting for 77.6% of the lineage. The qualitative features of this distribution—its breadth and long tail—are robust to several variations of the experimental

conditions, and are common to the lineages of related species. We computed additional profiles using deeper motifs (Figure 4A), motifs distinguished on the basis of cell type (*i.e.*, typological as well as topological patterns, Figure 4B – triangles), and motifs that ignore the orientation of cell division (such that isomorphic motifs were merged, Figure 4B – crosses). We also computed the motif profiles of two other nematode lineages, *Pellioditis marina* [19] and *Halicephalobus gingivalis* [20] (Figure 4C). In all cases, while minor differences were observed, the general shape of the motif profile is preserved.

4 Generative models of cell lineage development

4.1 A stochastic model of development

In what way are the motif profiles observed in the lineages of *C. elegans* and related species distinctive? Consider that *any* ensemble of randomly chosen 671-cell lineages will exhibit a motif profile with some characteristic distribution (a "null profile"). This null profile will not be uniform, as the occurrence of motifs is not independent, and some bias will arise from the constraint on cell number. For example, the proliferating 3-motif (labelled A in Figure 3) will be overrepresented in most large lineages. Furthermore, as motif depth increases, the number of possible motifs scales as a double

exponential (see Appendix A), making it increasingly unlikely that every possible motif will be represented.

The approach that we take to identifying a suitable null profile is to consider the profile resulting from a minimal generative process: a stochastic model in which each cell division is an independent random event [27]. The model is described fully in Appendix B and Figure 5 shows an example stochastic lineage. We used this model to create an ensemble of 1,000 lineages, each containing 671 terminal cells.

Lineages generated by the stochastic model contained a greater diversity of topological patterns than the *C. elegans* lineage: for motif depths greater than three ($d > 3$), each stochastic lineage required significantly more motifs to describe than the *C. elegans* lineage (Figure 6A). Given the rapid increase in number of possible motifs as motif depth increases, the probability of observing repeated motifs by chance decreases. The appearance of such repeated motifs in the *C. elegans* lineage therefore suggests a greater level of topological regularity compared to stochastic lineages of an equal size.

4.2 A dynamic regulatory network model of development

What is the source of this regularity in the *C. elegans* lineage? Research into morphogenetic pattern formation has shown that complex but regular patterns can result from relatively simple developmental mechanisms [12, 28, 29].

One important developmental control mechanism is the gene regulatory network in each cell [30]. To investigate the extent to which such developmental mechanisms can account for lineage regularity, we created a second ensemble of lineages using a generative model in which patterns of division were governed by the behaviour of a dynamic regulatory network.

The model gene network used to create developmental lineages was based on a dynamic recurrent network architecture [31, 32] that has been widely used to simulate the dynamics of gene expression [33–35] and the creation of artificial cell lineages [16, 36–38]. The dynamic regulatory network and developmental model are described fully in Appendix B and Figure 5 shows an example developmental lineage. An ensemble of 1,000 lineages was created using the developmental model with $N = 32; K = 8; \lambda = 0.225$. These parameters were chosen on the basis of initial trials to increase the likelihood of obtaining lineages containing approximately 671 cells.

The lineages produced by the developmental model were more regular than both the *C. elegans* lineage and those generated by the stochastic model: while 20 out of the 24 possible 3-motifs were represented across the entire ensemble of developmental lineages, each individual lineage contained only a small subset of these (five or six on average; Figure 6B). The *C. elegans* lineage therefore appears to share structural characteristics with both developmental and stochastic lineages: like a developmental lineage, much of it can be accounted for by a small number of motifs; like a stochastic lineage, it requires a much larger number of motifs to describe fully (Figure 7).

5 The influence of contextual constraints

Neither the stochastic nor the developmental model recover the broad distribution of motifs observed in the *C. elegans* lineage, nor the same multiscale regularity signature. What is missing? One likely explanation for this observation is that the gene network of *C. elegans*, with approximately 20,000 genes, is much more complex than the networks used in our simulations, and that different subnetworks might operate in different sublineages. Another possible explanation is that the gene network is not the only force shaping the cell lineage topology. Consider that the development of *C. elegans* is subject to specific spatial and temporal requirements, both globally—embryonic development must be completed inside the boundaries of the egg, before it hatches—and locally—all gut cells must be *co-located* within the embryo, for example [22]. In this section, we explore the possibility that some of the ways in which the *C. elegans* lineage departs from the stochastic and developmental models systematically reflect the influence of these spatio-temporal constraints.

5.1 A temporal constraint on the duration of development

A notable feature of nematodes is the speed of their embryonic development, possibly selected to reduce the duration of this vulnerable period, or to allow rapid colonisation of ecological niches [19]. The two most frequently observed

motifs in the *C. elegans* profile are the proliferating motif, in which none of the four terminal cells differentiate, and the terminating motif, in which all four terminal cells differentiate (motifs A and B in Figure 3). The high frequency of these particular motifs is a consequence of the inherently proliferative nature of early *C. elegans* development [22]. We therefore investigated the effect of a temporal constraint on the duration of development, as reflected by cell lineage depth.

We added a temporal constraint to the stochastic and developmental models described above by scaling the probability of cell division events to be inversely proportional to the depth of the cell (described in Appendix B; Figure 5 shows example scaled lineages). Again, two ensembles of 1,000 lineages, each containing 671 terminal cells were created (parameters for scaled developmental model: $N = 32; K = 8; \lambda = 0.425$). The resulting lineages proliferated earlier and were correspondingly less deep; model parameters were chosen to achieve a distribution of cell depths approximating that of the *C. elegans* lineage (Figure 8). In the case of the stochastic model, the temporal constraint reduced motif diversity, as the reduction in depth reduced the number of deep motifs contained in each lineage. However, the scaled stochastic profiles remained consistently more diverse than those of *C. elegans* for all motif depths (Figure 6A). In contrast, the temporal constraint increased motif diversity in the developmental model, as the deep but regular lineages produced by the standard variant were no longer possible, and the resulting lineages contained a wider variety of topological patterns

(Figure 6B).

5.2 A spatial constraint arising from the dimensionality of development

As noted above, cell divisions in *C. elegans* can be classified as occurring on either the anterior-posterior, dorsal-ventral or left-right axis [22], a distinction that has not thus far been incorporated in our analysis. The three-dimensional orientation of cell division plays an important role in *C. elegans* development by facilitating signalling events that establish and maintain the bilateral symmetry of an initially asymmetric embryo [19, 21, 22], as reflected in the complementary motifs exhibited by its lineage (*e.g.*, motifs C and D in Figure 3). By contrast, divisions in the lineages produced by the developmental and stochastic models are one-dimensional. In fact, examination of the developmental lineages revealed that all proliferation events were oriented identically (*i.e.*, individual developmental lineages were observed to contain either motif C or motif D in Figure 3, but not both), substantially reducing the number of potential motifs that each lineage could contain (Figure 9).

To investigate the effect of a spatial constraint arising from the dimensionality of development, we modified lineages created by the scaled stochastic and developmental models by reorienting a subset of cell divisions, such as may occur in a three-dimensional environment. Reversing the orientation of randomly chosen divisions left the scaled stochastic profiles unchanged, as

complementary motifs were already present (Figure 6A). However, for the scaled developmental profiles, reversing the orientation of as few as one in twenty divisions (a proportion comparable to the frequency of non anterior-posterior divisions in *C. elegans*) introduced complementary motifs at each depth that resulted in a profile signature comparable to that of *C. elegans* (Figure 6B). While the generated lineages are not identical to the *C. elegans* lineage, they share a common distribution of topological regularity across multiple scales.

6 Discussion

In this paper, we have demonstrated how the concept of motifs, originally used to analyse system structure, can also be applied to patterns of dynamic behaviour; here, the cell lineages arising from biological development. Analysing structures in terms of developmental motifs enables us to characterise the extent to which an system's organisation is regular or random. The motif profiles of *C. elegans* and related species are heavy tailed: Much of their structure follows a regular pattern; however, the exceptions to this pattern are not random and independent, but exhibit regularities of their own. We suggest that the distribution of motif profile sizes across motif depth (Figure 6) constitutes a signature of the topological regularity of a lineage across multiple scales. Comparing the *C. elegans* signature with those of artificial cell lineages generated by stochastic and dynamic regulatory network models

highlights those features that are distinctive to *C. elegans*.

For very shallow motifs ($d = 1, 2$), the regularity signatures converge because there are very few distinct motifs possible at this depth, and all of these tend to be represented in a single lineage. For very deep motifs ($d = 9, 10$), motif depth approaches the depth of the entire lineage, and the signatures converge as the total number of motifs (each of which tends to occur only once in a lineage) decreases. (The exception to this similarity is the unscaled stochastic lineages, which are much deeper than those produced by the other models.) In between these extremes ($3 \leq d \leq 8$), there is a larger disparity in profile size. As the number of possible motifs grows, stochastic lineages become increasingly diverse, with few instances of repeated motifs. In contrast, developmental lineages (those generated by the dynamic network models) exhibit only a small increase in motif diversity, reflecting the inherent regularity of a deterministic production system. The *C. elegans* lineage has greater topological diversity than the developmental lineages, but retains more repeated structure than the stochastic lineages, across a range of organisational scales.

We further demonstrated how relatively straightforward modifications to our basic models, reflecting the influence of spatial and temporal constraints, could lead to lineages sharing a similar topological signature to that of *C. elegans*. This similarity suggests that while some features of the *C. elegans* lineage are almost certainly the result of selection for adaptive morphologies or behaviours, others may be explicable in terms of the intersection between

generative bias and contextual constraints. Understanding the range of lineage topologies that occur in the absence of selection is important because it provides us with a sense of the raw material available for evolution to act on. Strong conclusions cannot be drawn on the basis of three samples, but they do provide a proof in principle of the approach and support our prediction that generative factors play a role in lineage topologies. Validating the significance of these regularity signatures will require comparison across the cell lineages of a wider range of species. Unfortunately, data for such a comparison is not currently available, although the development of new techniques for lineage mapping promises to extend the range of species for which it is possible to obtain cell lineage data [39–41]. In addition, recent technological advances in assaying patterns of gene expression in the *C. elegans* lineage raise the possibility of developing predictive gene network models that will further enhance our understanding of the relationship between developmental gene networks and lineage topology [42].

As mentioned above, the evolutionary relationship among species and grammatical structure in linguistics are also commonly represented as trees. Furthermore, phylogenies and languages are also systems whose structure is likely to have been shaped both by intrinsic dynamics and external forces. It is intriguing to consider what types of regularity may be revealed by the application of developmental motifs to other complex systems.

A Calculating the number of possible motifs

The number of possible binary motifs of depth d, for standard (oriented) and non-oriented motifs can be calculated as follows:

Standard (oriented) motifs The number of trees of height at most d, $a(d)$, is given by: $a(1) = 1$ (a single leaf node) and $a(d) = a(d-1)^2 + 1$. The number of motifs of depth d, $n(d)$, is then given by $a(d+1)-a(d-1)$, resulting in a series that scales as a double-exponential [43] (Table 1). The reasoning for this is as follows: The number of unlabelled motifs of depth $d+1$ is given by $a(d+1)$. If all nodes of depth $d+1$ in these motifs are removed, and their parents labelled as being non-terminal, we have the number of labelled motifs of depth d. We then remove all motifs of depth $d-1$ or less, giving $a(d+1) - a(d-1)$.

Non-oriented motifs The number of non-oriented motifs of depth d (for $d \geq 3$), $n'(d)$ is given by $\frac{(n'(d-1)+2)!}{2(n'(d-1)+1)!} - 1$, where $n'(1) = 2$ and $n'(2) = 3$ (Table 1). The reasoning for the non-oriented case is somewhat different from the standard case. A motif of depth d comprises a binary tree in which the left and right branches contain two motifs of depth $d-1$, with the possibility that at most one of those branches may be a terminal cell. Therefore we can calculate the number of possible motifs of depth d in terms of combinations with repetition, given by $\frac{(n+k-1)!}{k!(n-1)!}$, where $n = n'(d-1)+1$ and $k = 2$. We then subtract one to account for the case where both branches are terminal

cells, as this would result in a motif of depth less than d.

B Lineage models

B.1 Stochastic lineage models

A stochastic lineage with C terminal cells is created as follows:

1. Begin with a single cell c_0.
2. Choose a terminal cell c uniformly at random; with probability p_δ, append two child cells to c:

$$p_\delta = \begin{cases} 1.0, & \text{in the standard stochastic model} \\ 0.5^\delta, & \text{in the scaled stochastic model} \end{cases}$$

   where δ is the distance between c and c_0.
3. Repeat step 2 until the lineage contains C terminal cells.

B.2 Developmental lineage models

A network consists of two input nodes (providing contextual information to a cell), N regulatory nodes (each with K connections to other regulatory nodes), and one output node (used to control cell division). The activation

of node i at time t, $a_i(t)$ is given by

$$a_i(t) = \sigma\Big(\sum_{j=1}^{2} w_{ij}a_j(t-1) + \sum_{k=1}^{N} w_{ik}a_k(t-1) - \theta_i\Big)$$

where w_{ij} is the level of the interaction between input node j and regulatory node i, w_{ik} is the level of the interaction between regulatory nodes i and k, θ_i is the activation threshold of node i, and $\sigma(x)$ is the sigmoid function $\sigma(x) = (1 + e^{-x})^{-1}$. Weight values were initialised randomly from the Normal distribution $N(0.0, 2.0)$.

A developmental lineage is created as follows:

1. Initialise a single instance of the network, representing the initial cell c_0, by setting the activation of all of its nodes to 0.0.

2. For the current terminal cell c, update the activation of its network as described above.

3. A cell c divides if the activation of its division node is below p_δ:

$$p_\delta = \begin{cases} 1 - \lambda, & \text{in the standard developmental model} \\ 1 - 0.01e^{\lambda\delta}, & \text{in the scaled developmental model} \end{cases}$$

   where δ was the depth of the cell, and λ is a model parameter. If division is to occur, append two child cells to c, each containing a copy of the c's network with identical weights and node activations. Set the

activation of the two input nodes to $(0, 1)$ in the left child and $(1, 0)$ in the right child.

4. Otherwise, if the activation of the division output node is above p_δ, label c as being differentiated.

5. Repeat steps 2 to 4 until either all cells are labelled as differentiated, or some predefined limit on division depth had been reached.

Acknowledgements

We thank J. Noble and R. A. Watson for their helpful comments and discussion. N.G. and S.B. acknowledge financial support from the Engineering and Physical Sciences Research Council (EP/D00232X/1). R.A and R.L. acknowledge financial support from the National Science Foundation (EF-0742803). J.W. acknowledges financial support from the Australian Research Council.

References

[1] L. A. Adamic and B. A. Huberman. The web's hidden order. *Commun ACM*, 44:55–59, 2001.

[2] A.-L. Barabási and Z. N. Oltvai. Network biology: Understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113, 2004.

[3] M. Batty. The size, scale, and shape of cities. *Science*, 319:769–771, 2008.

[4] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439:462–465, 2005.

[5] R. Ferrer i Cancho and R. V. Sole. The small world of human language. *Proc R Soc London Ser B*, 268:2261–2265, 2001.

[6] D. C. Plaut, J. L. McClelland, M. S. Seidenberg, and K. Patterson. Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychol Rev*, 103(1):56–115, 1996.

[7] M. S. Seidenberg. Connectionist models of word reading. *Curr Dir Psychol Sci*, 14(5):238–242, 2005.

[8] G. K. Zipf. *Human Behaviour and the Principle of Least Effort. An Introduction to Human Ecology.* Addison-Wesley, Cambridge, MA, 1949.

[9] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and

U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298:824–827, 2002.

[10] R. Badii and A. Politi. *Complexity: Hierarchical Structures and Scaling in Physics*. Cambridge University Press, Cambridge, England, 1997.

[11] H. Kitano. Biological robustness. *Nat Rev Genet*, 5:826–837, 2004.

[12] I. Salazar Ciudad, J. Jernvall, and S. Newman. Mechanisms of pattern formation in development and evolution. *Development*, 130:2027–2037, 2003.

[13] S. Kirby. Spontaneous evolution of linguistic structure—An iterated learning model of the emergence of regularity and irregularity. *IEEE T Evolut Comput*, 5(2):102–110, 2001.

[14] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[15] W. Arthur. The effect of development on the direction of evolution: toward a twenty-first century consensus. *Evol Dev*, 6:282–288, 2004.

[16] R. Lohaus, N. L. Geard, J. Wiles, and R. B. R. Azevedo. A generative bias towards average complexity in artificial cell lineages. *Proc R Soc London Ser B*, 274:1741–1750, 2007.

[17] J. Maynard Smith, R. Burian, S. Kauffman, P. Alberch, J. Campbell,

B. Goodwin, R. Lande, D. Raup, and L. Wolpert. Developmental constraints and evolution. *Q Rev Biol*, 60:265–287, 1985.

[18] W. C. Wimsatt. Generativity, entrenchment, evolution, and innateness: Philosophy, evolutionary biology, and conceptual foundations of science. In V. G. Hardcastle, editor, *Where Biology Meets Psychology: Philosophical Essays*, pages 139–179. MIT Press, Cambridge, MA, 1999.

[19] W. Houthoofd, K. Jacobsen, C. Mertens, S. Vangestel, A. Coomans, and G. Borgonie. Embryonic cell lineage of the marine nematode *Pellioditis marina*. *Dev Biol*, 258:57–69, 2003.

[20] W. Houthoofd and G. Borgonie. The embryonic cell lineage of the nematode *Halicephalobus gingivalis*. *Nematology*, 9(4):573–584, 2007.

[21] W. Houthoofd, M. Willems, K. Jacobsen, A. Coomans, and G. Borgonie. The embryonic cell lineage of the nematode *Rhabditophanes* sp. *Int J Dev Biol*, 52:963–967, 2008.

[22] J. E. Sulston, E. Schierenberg, J. G. White, and J. N. Thompson. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev Biol*, 100:64–119, 1983.

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

[24] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet*, 31:64–68, 2002.

[25] P.-M. Agapow and A. Purvis. Power of eight tree shape statistics to detect non-random diversification: A comparison by simulation of two models of cladogenesis. *Syst Biol*, 51:866–872, 2002.

[26] Aravind K. Joshi and Yves Schabes. Tree-Adjoining Grammars. In Grzegorz Rozenberg and Arto Salomaa, editors, *Handbook of Formal Languages: Volume 3. Beyond Words*, chapter 2, pages 69–124. Springer, Berlin, 1997.

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

[28] A. M. Turing. The chemical basis of morphogenesis. *Philos Trans R Soc London Ser B*, 237:37–72, 1952.

[29] G. Webster and B. Goodwin. *Form and Transformation: Generative and Relational Principles in Biology*. Cambridge University Press, Cambridge, 1996.

[30] E. H. Davidson, J. P. Rast, P. Oliveri, A. Ransick, C. Calestani, C.-H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C. T. Brown, C. B. Livi, P. Y. Lee, R. Revilla, A. G. Rust, Z. Jun Pan,

M. J. Schilstra, P. J. C. Clarke, M. I. Arnone, L. Rowen, R. A. Cameron, D. R. McClay, L. Hood, and H. Bolouri. A genomic regulatory network for development. *Science*, 295:1669–1678, 2002.

[31] J. L. Elman. Finding structure in time. *Cognitive Sci*, 14:179–211, 1990.

[32] J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, CA, 1991.

[33] E. Mjolsness, D. H. Sharp, and J. Reinitz. A connectionist model of development. *J Theor Biol*, 152:429–53, 1991.

[34] J. Vohradský. Neural model of the genetic network. *J Biol Chem*, 276: 36168–36173, 2001.

[35] A. Wagner. Does evolutionary plasticity evolve? *Evolution*, 50(3):1008–1023, 1996.

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

[37] D. A. Winkler, F. R. Burden, and J. D. Halley. Predictive mesoscale network model of cell fate decisions during *C. elegans* embryogenesis. *Artif Life*, 15(4):411–421, 2009.

[38] H. Yoshida, C. Furusawa, and K. Kaneko. Selection of initial conditions for recursive production of multicellular organisms. *J Theor Biol*, 233: 501–514, 2005.

[39] D. Frumkin, A. Wasserstrom, S. Kaplan, U. Feige, and E. Shapiro. Genomic variability within an organism exposes its cell lineage tree. *PLoS Comp Biol*, 1(5):e50, 2005.

[40] S. J. Salipante and M. S. Horwitz. Phylogenetic fate mapping. *Proc Natl Acad Sci USA*, 103:5448–5453, 2006.

[41] Adam Wasserstrom, Rivka Adar, Gabi Shefer, Dan Frumkin, Shalev Itzkovitz, Tomer Stern, Irena Shur, Lior Zangi, Shau Kaplan, Alon Harmelin, Yair Reisner, Dafna Benayahu, Eldad Tzahor, Eran Segal, and Ehud Shapiro. Reconstruction of cell lineage trees in mice. *PLoS ONE*, 3(4):e1939, 2008.

[42] X. Liu, F. Long, H. Peng, S. J. Aerni, M. Jiang, Sánchez-Blanco, J. I. Murray, E. Preston, B. Mericle, S. Batzoglou, E. W. Myers, and S. K. Kim. Analysis of cell fate from single-cell gene expression profiles in *c. elegans*. *Cell*, 139:623–633, 2009.

[43] A. V. Aho and N. J. A. Sloane. Some doubly exponential sequences. *Fibonacci Quart*, 11:429–437, 1973.

[44] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev*, 51(4):661–703, 2009.

[45] J. E. Sulston and J. G. White. Parts list. In W. B. Wood, editor, *The Nematode Caenorhabditis elegans*, pages 415–431. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1988.

Figure Legends

Figure 1: All twenty-four possible 3-motifs. A circle indicates a terminal node; an arrowhead indicates a non-terminal node.

Figure 2: Computing a 3-motif profile for a cell lineage. A: The *C. elegans* sublineage AB.praa, where n=neural cell, e=epithelial cell and x=apoptosed cell). **B:** The corresponding 3-motif profile, showing the seven different motifs present (see Figure 1), and the number of times they each occur. **C:** The non-oriented 3-motif profile, containing five different motifs. Note that the fifth and sixth motifs in panel B are isomorphic to the third and fourth motifs, and have therefore been merged. In the typological 3-motif profile (not shown) all motifs except the first are unique, resulting in a profile of size eight.

Figure 3: The 3-motif profile for the embryonic cell lineage of *C. elegans* **hermaphrodite.** Motifs are ranked in order of decreasing frequency. The first four motifs, which account for 77.6% of the lineage, are shown. The inset shows the motif profile on a log-log scale, together with the fit to a power law ($\alpha = 1.73$). Note that the fit is illustrative only, as the sample size is too small to allow us to rule out other distributions [44].

Figure 4: Varying profile depth, motif definition and species. A:

d-motif profile comparison for *C. elegans* for $d = 3, 4, 5$. The slope of the distributions decrease as motif depth increases—a result of a decrease the total number of motifs observed and an increase the proportion of motifs that are represented by only one or two instances—however, the general shape of the distribution is maintained. **B:** Motif profile comparison for *C. elegans* for the original motif definition ($d = 3$), the typological motif variant ($d = 2$), and the non-oriented motif variant ($d = 4$). Values of d were chosen such that the size of the set of possible motifs for each variant was of the same order of magnitude (24 normal motifs, 81 typological motifs and 54 non-oriented motifs). For the typological motif profiles, cells in the *C. elegans* lineage were classified into nine categories according to their structure and function [45]: 39 blast, 113 deaths, 78 epithelial, 2 germ, 10 gland, 20 intestinal, 108 muscle, 79 neural structural and 222 neurons. **C:** 3-motif profiles for *C. elegans* (671 terminal cells), *P. marina* (638 terminal cells) [19] and *H. gingivalis* (536 terminal cells) [20]. The range and general shape of the distributions are similar; however, the terminal cells of the *P. marina* and *H. gingivalis* lineages have not yet been fully characterised, leading to an increased representation of the terminating 3-motif (*i.e.*, containing four terminal cells) in the second rank position.

Figure 5: Example model lineages. Example lineages created by the standard and scaled versions of the stochastic and developmental models.

Figure 6: Comparison of motif profile sizes between *C. elegans*, and lineages produced by stochastic (A) and developmental (B) models across a range of motif depths. *C. elegans* motif profiles sizes (bold line) are shown on both plots. Stochastic and developmental lineages were produced according to standard (circles; $N = 32; K = 8; \lambda = 0.225$) and scaled (squares; $N = 32; K = 8; \lambda = 0.425$) variants of each model, as well as a scaled variant in which 5% of lineage branches were reversed (diamonds). Error bars show standard deviation of the profile size over each ensemble of 1,000 lineages, when larger than symbol. The stochastic models consistently overestimate motif diversity (Note: branch reversal has minimal effect on the diversity of stochastic profiles, therefore some data points in plot A overlap). The standard developmental model underestimates motif diversity (B–circles); however, the recognition of temporal and spatial factors influencing development (represented by scaled division probabilities and branch reversal) results in lineages that share a similar level of motif diversity with *C. elegans* across multiple scales (B–diamonds).

Figure 7: Mean fraction of a lineage described by the first M 3-motifs. Data shown for the *C. elegans* lineage and each of the stochastic and developmental models (standard: $N = 32; K = 8; \lambda = 0.225$; scaled: $N = 32; K = 8; \lambda = 0.425$). Error bars show standard deviation of lineage fraction over each ensemble of 1,000 lineages, when larger than symbol. Only 9 motifs are required to capture even the least regular developmental lineage.

By comparison, whereas 9 motifs capture 92.2% of the *C. elegans* lineage, a further 12 motifs are required to capture the remaining 7.8%. The stochastic lineages are closest to the hypothetical uniform case in which all motifs are represented equally; however some bias exists due to the nature of the stochastic process.

Figure 8: Cell depth distributions for *C. elegans* and the stochastic and developmental models (standard: $N = 32; K = 8; \lambda = 0.225$; scaled: $N = 32; K = 8; \lambda = 0.425$). Plot shows mean number of terminal cells occurring at a given depth for each model. Note the horizontal axis has been truncated at depth 25; however, the full range of cell depths for the lineages created by the standard developmental model was much greater, reflecting the presence of deep lineages in which a large majority of cells were generated via a stem-cell mode of division (*i.e.*, where either the left or right child consistently continued to divide, while the other child differentiated).

Figure 9: Comparison between standard and non-oriented profile sizes for *C. elegans* and developmental lineages. Motif diversity in developmental lineages is restricted by an absence of isomorphic, or complementary, motifs (see Figure 5). This plot shows the reduction in profile size of non-oriented motif profiles compared to standard motif profiles for *C. elegans* and the two developmental models (standard: $N = 32; K = 8; \lambda = 0.225$; scaled: $N = 32; K = 8; \lambda = 0.425$). The size of the *C. elegans* profile

decreases by over 50% for motif depths 3 and 4, indicating that many isomorphic motifs have been 'collapsed' into their non-oriented equivalents. By contrast, the size of the developmental profiles remains relatively constant, indicating a very low incidence of isomorphic motifs in the standard motif profiles—any individual lineage tends to contain motifs oriented to (*i.e.*, dividing towards) either the left or the right, but not both.