Developmental Motifs Reveal Complex Structure in Cell Lineages

NICHOLAS GEARD,1 SETH BULLOCK,1 ROLF LOHAUS,2 RICARDO B. R. AZEVEDO,2 AND JANET WILES3
1School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, United Kingdom; 2Department of Biology and Biochemistry, University of Houston, Houston, Texas 77204-5001; and 3School of Information Technology and Electrical Engineering, The University of Queensland, St Lucia, Queensland 4072, Australia

Received March 8, 2010; revised May 27, 2010; accepted June 25, 2010

Many natural and technological systems are complex, with organizational structures that exhibit characteristic patterns but defy concise description. One effective approach to analyzing such systems is in terms of repeated topological motifs. Here, we extend the motif concept to characterize the dynamic behavior of complex systems by introducing developmental motifs, which capture patterns of system growth. As a proof of concept, we use developmental motifs to analyze 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 organization. © 2010 Wiley Periodicals, Inc. Complexity 16: 48–57, 2011

Key Words: 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 behaviors that are, in some fashion, organized [1–8]. Although 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 such recurring patterns, the organization 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 behavior. A first step toward this goal involves characterizing 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 inter-
actions between neighboring cells, and chemical signaling
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 processes responsible for producing complex
systems can bias the range and type of structures that are
observed [15–18].

Efforts to analyze the growth of complex systems are com-
licated by the fact that more recent structures tend to over-
write 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 developmen-
tal 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 organ-
ism’s development in terms of patterns of division and differ-
entiation 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 developmen-
tal 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 subline-
ge can contain multiple cell types. Upon mapping the cell
lineage, Sulston concluded that “the assignment of cell func-
tion 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 understand-
ing of how to describe and account for the organization 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 article, we introduce an analytic tool, developmen-
tal motifs, that provides a novel perspective on the relation-
ship between generative processes and cell lineage topology.
We build upon the concept of network motifs: the “recurring,
significant, patterns of interconnections” observed in a vari-
city of complex networks [9, 24]. Network motifs were intro-
duced 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
organizational 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 analyze the cell
lineages of C. elegans and related species, revealing the pres-
ence of a broad distribution of motif frequencies. We then use
a suite of computational models to explore the role that gen-
erative 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 nonterminal 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 relation-
ship 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 nonterminal, corresponding to its sta-
tus 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 nonter-
nominal node—while there are four possible two-motifs and 24
possible three-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 lin-
eage 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 dis-
tinct motifs, even at greater depths, while those of less regular
lineages would display greater diversity.

While focusing here on cell lineages, we also recognize the
presence of tree-like organization 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 NEMATOMES

What does the developmental motif profile of a real organ-
ism 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-characterized
sublineages that display modular and recursive patterns of
cell
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 three-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 three-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 4(A)), motifs distinguished on the basis of cell type [i.e., typological as well as topological patterns, Figure 4(B) – triangles], and motifs that ignore the orientation of cell division [such that isomorphic motifs were merged, Figure 4(B) – crosses]. We also computed the motif profiles of two other nematode lineages, *Pellioides marina* [19] and *Halicephalobus gingivalis* [20] (Figure 4(C)). In all cases, although 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 three-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...
The three-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 [27].

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

The approach that we take to identify 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 [29]. 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 1000 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 with 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, 30, 31]. One important developmental control mechanism is the gene regulatory network in each cell [32]. 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 behavior of a dynamic regulatory network.

The model gene network used to create developmental lineages was based on a dynamic recurrent network architecture [33, 34] that has been widely used to simulate the dynamics of gene expression [35–37] and the creation of artificial cell lineages [16, 38–40]. The dynamic regulatory network...
and developmental model are described fully in Appendix B and Figure 5 shows an example developmental lineage. An ensemble of 1000 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 three-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 $\sim 20,000$ genes, is much more complex than the networks used in our simulations, and that different subnetworks might operate in different lineages. 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...
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 1000 lineages, when larger than symbol. The stochastic models consistently overestimate motif diversity (Note that 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).

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 1000 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 6(A)]. 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 6(B)].

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 as few as one in twenty divisions (Figure 6(A)) introduced complementary 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 1000 lineages, when larger than symbol. Only nine motifs are required to capture even the least regular developmental lineage. By comparison, whereas nine 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.

6. DISCUSSION
In this article, we have demonstrated how the concept of motifs, originally used to analyze system structure, can also be applied to patterns of dynamic behavior; here, the cell lineages arising from biological development. Analyzing structures in terms of developmental motifs enables us to characterize the extent to which an system’s organization 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. Computing 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 remains more repeated structure than the stochastic lineages, across a range of organizational 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 behaviors, others may be explainable 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 [41–43]. In addition, recent technological advances in assessing 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 [44].

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.

**APPENDIX 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:

### A.1. Standard (Oriented) Motifs

The number of trees of height at most \( d \), \( m(d) \), is given by:

\[
m(d) = a(d - 1) - ad - 1.
\]

The number of motifs of depth \( d \), \( n(d) \), is then given by

\[
n(d) = a(d - 1) - ad - 1 + 1.
\]

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 labeled as being nonterminal, we have the number of labeled motifs of depth \( d \). We then remove all motifs of depth \( d - 1 \) or less, giving \( a(d + 1) - a(d - 1) \).

### Table 1: The Number of Standard \((m(d))\) and Non-oriented \((n(d))\) Motifs of Depth \(d\)

<table>
<thead>
<tr>
<th>(d)</th>
<th>(m(d))</th>
<th>(n(d))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>672</td>
<td>54</td>
</tr>
<tr>
<td>5</td>
<td>458,304</td>
<td>1,539</td>
</tr>
<tr>
<td>6</td>
<td>210,066,388,224</td>
<td>1,199,569</td>
</tr>
</tbody>
</table>
The number of non-oriented motifs of depth \( n \) as this would result in a motif of depth less than \( n \). Therefore, we can calculate the number of possible motifs of depth \( n \) in terms of combinations with repetition, given by \( \binom{n+d-1}{d} \), 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 \( n \).

APPENDIX 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 \( c_i \) uniformly at random, with probability \( p_i \), append two child cells to \( c_i \):
   \[ p_i = \begin{cases} 1:0, & \text{in the standard stochastic model} \\ 0.5^i, & \text{in the scaled stochastic model} \end{cases} \]
   where \( i \) is the distance between \( c_i \) 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 \( j \) at time \( t \), \( a_j(t) \), is given by

\[
a_j(t) = \sum_{i=1}^{2} w_j a_i(t) + \sum_{k=1}^{N} w_{jk} a_k(t - 1) - \theta_j.
\]

where \( w_{jk} \) is the level of the interaction between input node \( j \) and regulatory node \( k \), \( w_j \) is the level of the interaction between regulatory nodes \( i \) and \( j \), \( \theta_j \) is the activation threshold of node \( j \), and \( f(\cdot) \) is the sigmoid function \( f(x) = \frac{1}{1 + e^{-x}} \). Weight values were initialized randomly from the Normal distribution \( N(0,0.2) \).

A developmental lineage is created as follows:

1. Initialize 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 \( c_i \), update the activation of its network as described above.
3. A cell \( c_i \) divides if the activation of its division node is below \( P_d \):
   \[ P_d = \begin{cases} 1 - \lambda, & \text{in the standard developmental model} \\ 1 - 0.01\lambda^{-1}, & \text{in the scaled developmental model} \end{cases} \]
   where \( \lambda \) is the depth of the cell, and \( \lambda \) is a model parameter. If division is to occur, append two child cells to \( c_i \), each containing a copy of the \( c_i \)'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 \( P_d \), label \( c_i \) as being differentiated.
5. Repeat steps 2 to 4 until either all cells are labeled as differentiated, or some predefined limit on division depth has been reached.

ACKNOWLEDGMENTS

We thank J. Noble, R. A. Watson and J. Jordan 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