



**Manchester
Metropolitan
University**

Waites, W and Cavaliere, Matteo and Cachat, E and Danos, V and Davies, JA (2018) *An information-theoretic measure for patterning in epithelial tissues*. IEEE Access. ISSN 2169-3536

Downloaded from: <http://e-space.mmu.ac.uk/621242/>

Version: Accepted Version

Publisher: Institute of Electrical and Electronics Engineers (IEEE)

DOI: <https://doi.org/10.1109/ACCESS.2018.2853624>

Usage rights: ["licenses_typename_i" not defined]

Please cite the published version

<https://e-space.mmu.ac.uk>

An information-theoretic measure for patterning in epithelial tissues

William Waites^{*||}, Matteo Cavaliere[†], Élise Cachat[‡], Vincent Danos^{*§}, and Jamie A. Davies[¶]

^{*}School of Informatics, University of Edinburgh

[†]School of Computing, Mathematics & Digital Technology, Manchester Metropolitan University

[‡]School of Biological Sciences, University of Edinburgh

[§]École Normale Supérieure, Paris

[¶]Deanery of Biomedical Sciences, University of Edinburgh

^{||}Corresponding author: wwaites@ieee.org

Abstract—We present *path entropy*, an information-theoretic measure that captures the notion of patterning due to phase separation in organic tissues. Recent work has demonstrated, both *in silico* and *in vitro*, that phase separation in epithelia can arise simply from the forces at play between cells with differing mechanical properties. These qualitative results give rise to numerous questions about how the degree of patterning relates to model parameters or underlying biophysical properties. Answering these questions requires a consistent and meaningful way of quantifying degree of patterning that we observe. We define a resolution-independent measure that is better suited than image-processing techniques for comparing cellular structures. We show how this measure can be usefully applied in a selection of scenarios from biological experiment and computer simulation, and argue for the establishment of a tissue-graph library to assist with parameter estimation for synthetic morphology.

I. INTRODUCTION

One of the major mechanisms for understanding tissue development is adhesion-mediated sorting of cell mixtures into homotypic groups, which was discovered by Steinberg in the 1960s [1]. Interest in this phase separation mechanism has recently surged, partly because of its ability to create synthetic biological patterning mechanisms [2] and partly because it has been found to drive events critical to the formation of organoids from stem cells [3, 4], making the process relevant to biotechnology as well as to basic development.

These investigations in experimental and synthetic biology have been paralleled by the development of analytic and computational models to explain pattern development. The first class of these are reaction-diffusion systems, such as those of Turing [5] and Gierer [6], in which a slowly diffusing activator molecule activates its own synthesis and also the synthesis of a rapidly diffusing inhibitor molecule. In such a system, small random asymmetries lead to slightly elevated production of activator morphogens and become centres of activator production and inhibit nearby sites from doing the same. The result is a field with separated spots, patches or stripes of high activator expression, which can be modelled for a two-component fluid system by the Cahn-Hilliard equation [7].

The second class of model is discrete, and patterning emerges from the mechanical properties of the cells themselves: cell-cell adhesion, contractility, and the balance

between cell surface area and volume. In this class are the Cellular Potts model [8] and the model of Newman et al. [9], in which motion takes place on a mesh of a scale much smaller than a cell, and Vertex models [10, 11], in which the system is represented as a dynamic and irregular mesh where polygons correspond directly to cells. Recently, analytic results have become available [12, 13] that predict cell shapes produced by both numerical simulations and models in a homogeneous setting and they have been demonstrated [14] to produce phase separation in simulation in a heterogeneous setting.

Against this background, there is a dearth of tools for comparing data produced by each of these disparate methods. Qualitatively, snapshots of tissues undergoing phase separation in simulation [13, 14] look similar to those produced experimentally by engineering cells with different levels of cadherin molecules [2]. In both cases the mechanism is understood to be Steinbergian differential adhesion, but the commonly used methods for quantitative techniques on epithelial sheets are mainly concerned with polygon distributions [15] or structural motifs [16] and are not straightforwardly extended to a setting with multiple cell types.

Graph-based distance, or graph similarity measures are well known. Eschera and Fu [17] define a distance between attributed feature graphs extracted from images in terms of transformations required to derive one from the other. Others such as Bunke and Shearer [18] define a distance (in fact, a metric) in terms of the size of the maximal common subgraph. Measures of these types do not, however, contain any intrinsic notion of pattern or information, so do not adequately capture these higher-level concepts. They are, in a sense, overspecific.

Shannon's entropy [19] has proven difficult to extend to two or more dimensions in a meaningful way. The fundamental problem is that entropy depends fundamentally on the underlying probability distribution over *some* set of possibilities, but there is no unique way to decide *which* set is appropriate. Entropy is an extrinsic anthropomorphic concept, not an intrinsic property of the system [20] precisely because of this freedom to choose the appropriate distribution. Information-theoretic measures for images are known, but they are typically constructed on the probability distribution of pixel values in an image [21–23], essentially transforming a two-dimensional

problem into one dimension, sacrificing spatial structure in the process.

The Maximum Entropy technique [24], widely used in image reconstruction from partial data, treats an image as a two dimensional structure, but is necessarily sensitive to image resolution. Likewise, other measures such as by Rubner et al. [25] and the vast literature on distances between images for retrieval purposes do encode something of the information content, are also relative to the image resolution. For that reason, without some kind of pre-alignment such as with Cuturi & Doucet’s technique of fast computation of Wasserstein Barycentres [26], they are not directly applicable to the task of comparing tissue examples from vastly different sources — experimental imagery on the one hand and simulation data on the other. A similar criticism can be made of Larkin’s delentropy measure [27] (however, see section 2 of Larkin’s paper for an extended discussion of information-theoretic measures of images).

In this paper, we provide such a method by defining a family of resolution-independent entropy measures on graphs that captures the different patterns observed throughout the literature on phase separation in cellular tissues. We choose to frame the measure in terms of graphs not only because the cell-cell contacts of epithelial and other biological tissues are intrinsically graph-like [28], but because it is independent of the scaling or resolution of imagery. The property of resolution-independence is important because it allows comparison across different experiments, both *in vitro* and *in silico*. Using this measure, it is possible to answer such salient questions as how quickly a pattern forms when starting from a random tissue or to meaningfully compare the degree of patterning observed in different numerical or wet-lab experiments. This capability enables workflows in synthetic mammalian biology where the goal is to engineer cell lines that will produce these kinds of patterns. Whether and to what extent the desired pattern is achieved can be consistently measured, and this information fed back into the system as the genome, host environment or external stimulus is adjusted.

II. MATHEMATICAL PRELIMINARIES

We will use some concepts from graph theory and from information theory and probability. We assume a basic level familiarity with these on the part of the reader. Nevertheless, we review some key definitions and clarify the notation that we use throughout.

A set, X is a collection of elements. The number of elements in the set, its cardinality, is written as $|X|$. If another set Y is a subset of X , written $Y \subseteq X$, then the chance of choosing an element x of X uniformly at random and finding that it is also an element of Y is $Pr(x \in Y) = \frac{|Y|}{|X|}$.

A partition of a set, is a set of non-empty subsets of X called $\{Y_i\}$, such that each element in X is in exactly one of the Y_i . A partition gives rise to a probability distribution, which has the property that,

$$\sum_i Pr(x \in Y_i) = \sum_i \frac{|Y_i|}{|X|} = 1 \quad (1)$$

The Cartesian product of two sets, $X \times Y$ is the set of pairs $(x \in X, y \in Y)$. If both sets are the same, this is also written as X^2 and analogously for higher powers.

A directed graph, G , consists of a set of vertices, V , also called nodes, and a set of edges that connect the vertices, $E \subset V^2$. A path of length n on the graph is a sequence of vertices, (v_0, v_1, \dots, v_n) such that $(v_i, v_{i+1}) \in E$ for $0 \leq i < n$. We take the special case of zero-length paths to be simply the set of vertices itself. Let us write $S_n(G)$ for the set of all paths of length n from the graph, G .

A *graph invariant* is a quantity that depends only on the structure of the graph itself and not any representation or labelling. In particular it is a quantity that is invariant under graph isomorphism.

Let C be a set of colours and $\chi : V \rightarrow C$ be a function that maps vertices to colours. A coloured graph, (V, E, χ) is a graph together with such a function. Note that χ induces a partition on G when applied to each vertex. This partition map groups vertices together by colour.

III. PATH ENTROPY

To motivate our pattern complexity measure more concretely, let us consider some exemplar simulated tissues, shown in Figure 1. These consist of three kinds of cells, represented as different colours, in equal proportion. The qualitative difference between each of the images is intuitively clear, from no discernible pattern, to a kind of quasi-uniform distribution of white cells, in long, thin stripes and round patches reminiscent of the “dappling” calculated by hand by Turing. We seek a measurement that can be made on these that is able to distinguish them.

We choose to define this measure on the coloured adjacency graph of cells, as opposed to an image of the tissue as in the approach taken, for example, in [27]. The reason for this choice is that when calculated on the graph, the measure is resolution independent. It can be applied equally well to simulation data that has no intrinsic notion of image or resolution or to processed outputs from experimental imagery of cell colonies or epithelial sheets. As an important goal is to be able to compare data from different sources, this property is important.

Let us proceed as follows. The entities of interest are cells so let us say that V corresponds to the set of cells in a given tissue. Further, let E be the edges, the adjacencies between cells. The patterns of interest are meaningful in terms of different kinds of cells so let the colours, C , correspond to the kind. For the purposes of this paper we are concerned with the resulting coloured graph which we call the *adjacency graph of cells*.

Intuitively, a pattern is found in the sequence of colours extending out in one direction or another from a given point in the tissue. To capture this, we lift the colouring function from operating on vertices, to operating on sequences of vertices, or paths, $\chi_n : S_n(G) \rightarrow C^n$ for a given path length, n . As with χ , the χ_n induces a partition of $S_n(G)$: paths with the same colour sequence get into the same class.

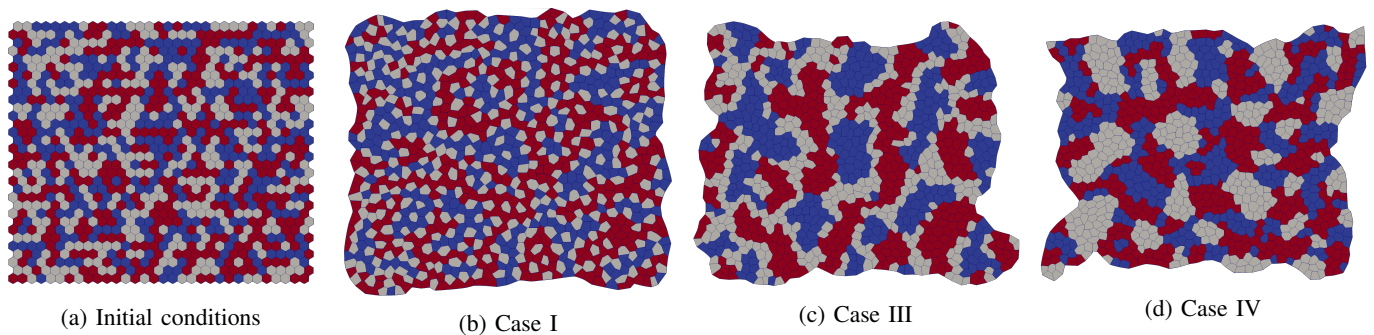


Figure 1: A selection of exemplar simulated tissue configurations for three cell types in equal proportion. At left, randomly distributed populations on a regular hexagonal lattice (typical initial conditions for simulation). The others are the resulting configurations after some elapsed time for different costs of heterotypic and homotypic edges between cells. In particular, the cost of a heterotypic edge with a white cell increases from left to right, and the cost of a homotypic edge between white cells decreases. The precise meaning of Case I through Case IV is explained in Section XI.

We use this partition to obtain a probability distribution over C^n ,

$$p_n(G)(s) = \frac{|\{\sigma \in S_n(G), \chi_n(\sigma) = s\}|}{|S_n(G)|} = \frac{|\chi_n^{-1}(s)|}{|S_n(G)|} \quad (2)$$

where $s \in C^n$. Where there is no risk of confusion or ambiguity, we will write $p_n(s)$ in place of $p_n(G)(s)$ from now on.

Definition. Given a coloured graph, $G = (V, E, \chi)$ and the probability distribution over colour sequences given by Equation 2, we define the n -th order Path Entropy on the graph to be the Shannon Entropy of this distribution:

$$E_n(G) = - \sum_{s \in C^{n+1}} p_n(s) \log(p_n(s)) \quad (3)$$

As with the probability distribution, we write simply E_n in place of $E_n(G)$ where there is no risk of confusion.

Note that though the motivation is a measure on planar graphs representing epithelial sheets, there is nothing in this formulation that presupposes such a restriction. The family of entropy measures is equally well defined on graphs that embed into three or higher dimensional spaces.

IV. GENERALISATION TO MOTIFS

The foregoing is concerned with paths only, one-dimensional sequences of vertices. There is evidence that it may be fruitful to consider two dimensional motifs, or graph fragments [16]. The approach given here can be straightforwardly applied to motifs. The general pattern for defining an entropy on a graph is to come up with a partition map and use the probability distribution that arises from that to get an entropy [29]. A set of motifs induces a partition on the graph: the set of sets of subgraphs matched by each motif. Indeed a path of length n is simply a special kind of motif.

In order to deal with coloured graphs, or heterogeneous tissues, the matching function is simply lifted to a form that distinguishes differently coloured motifs as opposed to the purely structural ones considered by Vincente et al. This is precisely analogous to the coloured paths that we have used

above. The corresponding notion of *Motif Entropy* follows directly.

V. COMPUTATIONAL COMPLEXITY OF PATH ENTROPY

The steps required to calculate E_n , following directly from the definition in Section III, are as follows.

- A. We begin by enumerating of paths of length n , $S_n(G)$. This can be accomplished with a depth-first search to depth n for each vertex. Fortunately, though the number of paths can be very large, $|V|^{n+1}$ for a complete graph, we do not need to store the paths themselves: we can simply proceed to the next step and count occurrences of each colour sequence. Naturally we need to produce each path, so we must have time complexity of $\mathcal{O}(k|S_n(G)|)$ where k is a factor describing the complexity of producing a single path. This method time complexity in the worst case of $\mathcal{O}(|V|^{n+1})$ [30] for a complete graph, and because paths can be produced incrementally during the search, k must be no more than a constant. For planar graphs of the kind considered here, where average degree $\langle d \rangle \approx 6$ [15], the situation is somewhat better, with time complexity of $\mathcal{O}(|V| \langle d \rangle^n)$. The depth-first search has space complexity of $\mathcal{O}(|V|)$ to keep track of each vertex visited.
- B. For each path, $\sigma \in S_n(G)$, we compute its colour sequence, $\chi_n(\sigma)$, and count the occurrences of each sequence. This requires visiting each vertex $v \in \sigma$ and computing $\chi(v)$. The time complexity is therefore $\mathcal{O}(|S_n(G)|)$, just as for the previous step. An upper bound on the space complexity can be obtained by supposing that all possible colour sequences occur. This is certainly the case for small numbers of colours and short paths such as we consider here. In this case, a count must be stored for each colour sequence, giving space complexity of $\mathcal{O}(|C|^n)$.
- C. We next compute the probability distribution, Equation 2. We must know $|S_n(G)|$, and for each colour sequence count from the previous step, $|\chi^{-1}(s)|$, to work out the ratio of paths with each sequence to the total number of paths. We can bound this as we have done with the

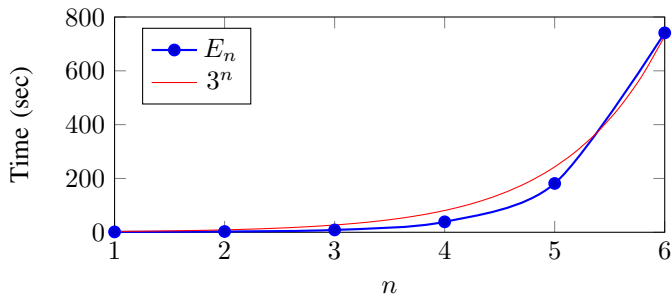


Figure 2: Empirical running time of calculation of E_n of the graph of Figure 1d for increasing values of n with an implementation in Python running on a 2.4GHz Intel Xeon E5645 CPU.

previous step at one division per sequence, and store a floating-point number for each, giving both space and time complexity of $\mathcal{O}(|C|^n)$.

- D. Finally, we calculate the entropy as in Equation 3. This entails iterating over each element, p_i , in the distribution and calculating $p_i \log(p_i)$, while keeping a running sum. This clearly has $\mathcal{O}(1)$ additional space complexity and the number of arithmetic operations is linear in the number of elements in the distribution, so time complexity is $\mathcal{O}(|C|^n)$.

In summary, the time complexity of calculating E_n is bounded by,

$$\begin{array}{ll} \mathcal{O}(n|V| \langle d \rangle^n + |C|^n) & \text{Average case} \\ \mathcal{O}(n|V|^{n+1} + |C|^n) & \text{Worst case} \end{array} \quad (4)$$

and the space complexity by,

$$\mathcal{O}(|V| + |C|^n) \quad (5)$$

Additionally we can verify empirically that the running time for the above procedure for calculating E_n increases comparably to an exponential function of n , as shown in Figure 2. As we see below, in practice it is unnecessary to calculate E_n directly for $n > 1$ so the exponential running time is not a serious handicap.

A. Linearity of Path Entropy

As discussed below in Section X, we find an empirical result that, for the graphs and colourings under consideration here, that the path entropy E_n is linear in n . That is,

$$E_n = (E_1 - E_0)n + E_0 \quad n > 0 \quad (6)$$

This observation is significant because as shown by Equation 4, the computational work to calculate E_n directly grows exponentially with n . Since it can be worked out simply from E_0 , E_1 and n , there is little benefit in the direct calculation.

It is important to note that this result does not hold in general. An easy way to find a counterexample is to construct a graph where the colour of the $(n + 1)$ th vertex in a path depends not only on the n th but also on previous vertices. Fortunately the paths in the coloured planar graphs that we consider here do not appear to have this property. An

interesting theoretical problem that we do not treat here is to precisely determine for which underlying coloured graphs this linear relation holds, and for graphs where it does not, what can be deduced about the path entropy for paths of lengths greater than two.

VI. RELATIVE ENTROPY

For completeness, and because it will be used later, we review the concept of *relative entropy* between two probability distributions. This is known in a more general setting as the Kullback-Leibler divergence [31] and is written,

$$D(p|q) = \sum_i p_i \log \left(\frac{p_i}{q_i} \right) \quad (7)$$

for two distributions, $p = \{p_i\}$ and $q = \{q_i\}$. For this to be well-defined, it is required that $p_i = 0$ if $q_i = 0$. Intuitively it gives a notion of distance between two distributions, however this intuition should be taken with a grain of salt: as formulated, in general it will violate the triangle inequality.

In the present context, we consider the distance from a reference graph containing paths R to a given graph G . The reference graph could be the initial conditions for a simulation or experiment or it could be an exemplar or “typical” pattern. This distance in this setting is simply,

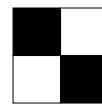
$$D_1(G|R) = \sum_{s \in C^2} p_1(G)(s) \log \left(\frac{p_1(G)(s)}{p_1(R)(s)} \right) \quad (8)$$

VII. EXAMPLES IN TWO COLOURS

To see how path entropy works in practice, and before considering real examples, let us consider a few simple cases. We first consider very simple patterns in two colours for which entropies can be calculated by hand on rectangular lattices, and then more complex but nevertheless artificial patterns in three colours on hexagonal lattices shown in Figure 5.

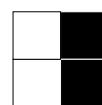
Starting with the simplest possible regular, symmetric two-colour, diagram will illustrate how the measure E_1 captures clustering. In what follows we do not impose periodic boundary conditions, although it would be perfectly natural to do so. Instead, we opt to consider, for clarity of presentation, the graphs exactly as they appear on the page.

Consider a 2x2 checkerboard,



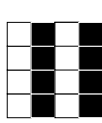
$$\begin{aligned} |S_1| &= 8 \\ |\chi_1^{-1}(wb)| &= |\chi_1^{-1}(bw)| = 4 \\ E_1 &= 1 \end{aligned}$$

This can obviously be extended to checkerboards of arbitrary size. Furthermore, larger checkerboards will, provided symmetry is preserved, give numerically the same value for E_1 because there are no like-colour adjacencies and every unlike-colour adjacency is reflexive.



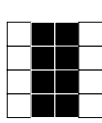
$$\begin{aligned} |S_1| &= 8 \\ |\chi_1^{-1}(ww)| &= |\chi_1^{-1}(bb)| = 2 \\ |\chi_1^{-1}(wb)| &= |\chi_1^{-1}(bw)| = 2 \\ E_1 &= 2 \end{aligned}$$

Rearranging the squares into stripes, we can see the measure E_1 distinguish between different kinds of regularity. With a little more work, we can see that this value for E_1 is characteristic of stripes one cell wide on a rectangular lattice,



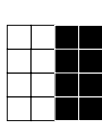
$$\begin{aligned} |S_1| &= 48 \\ |\chi_1^{-1}(ww)| &= |\chi_1^{-1}(bb)| = 12 \\ |\chi_1^{-1}(wb)| &= |\chi_1^{-1}(bw)| = 12 \\ E_1 &= 2 \end{aligned}$$

Rearranging the stripes into a thick and two thin, however, we see that the E_1 measure counts it as, in some sense, more regular. Or, more to the point, *more clustered*,



$$\begin{aligned} |\chi_1^{-1}(ww)| &= 12 \\ |\chi_1^{-1}(bb)| &= 20 \\ |\chi_1^{-1}(wb)| &= |\chi_1^{-1}(bw)| = 8 \\ E_1 &= 1.89 \end{aligned}$$

Finally, two thick stripes,



$$\begin{aligned} |\chi_1^{-1}(ww)| &= 20 \\ |\chi_1^{-1}(bb)| &= 20 \\ |\chi_1^{-1}(wb)| &= |\chi_1^{-1}(bw)| = 4 \\ E_1 &= 1.65 \end{aligned}$$

and this is maximally clustered and a local minimum of the E_1 entropy. It is a local minimum because any change would increase the number of heterotypic edges, and decrease the homotypic ones. Such a change to the distribution of paths can only increase the corresponding entropy.

These minima are interesting. In general, for the two-colour case, the entropy will have two minima: for a maximally clustered pattern and for maximally dispersed, checkerboard pattern. The latter is easily seen to be a global minimum as all adjacencies are of the same, heterotypic, type. For the clustered case, while as many edges as possible are homotypic, there still must be an interface between clusters of different colours so not all adjacencies can be the same. The outcome of choosing an arbitrary adjacency at random cannot then be certain, so the entropy must be greater than for the checkerboard.

VIII. TWO-SPECIES EPITHELIA

We are now in a position to apply these measures to some real-world cases. We start with some data from the same series as the phase separation study previously mentioned [2]. In that study, cells are genetically engineered to vary their level of production of cadherin molecules in response to external regulation using tetracycline. The cadherin molecules govern the adhesiveness of the cells to their neighbours. Two varieties of these cells, differing only in the nature of cadherin expressed, and therefore adhesiveness, upon tetracycline induction, were mixed randomly together in a 50:50 mixture and allowed to settle. Cell cultures from experiments with, and without tetracycline are shown in Figures 3a and 3b.

Some processing is needed to take this data into a form where the measures that we define here can be applied. The

procedure is relatively straightforward. First, positions and kinds of the nuclei are identified directly from the image. These provide the vertices for our graph. Next, neighbour relationships are derived from the Voronoi tessellation of these points. The results of this procedure on the confocal images are shown in Figures 3c and 3d.

The simulation method that we use to compare to this experimental data is similar to that of Osborne et al. [14] using the Chaste software package [32] and Farhadifar's potential [11]. In brief, the tissue is described by a potential,

$$U = \sum_{i \in V} \frac{K}{2} (A_i - A_0)^2 + \sum_{i,j \in V^2} \lambda_{ij} E_{ij} + \sum_{i \in V} \frac{\Gamma}{2} P_i^2 \quad (9)$$

where A_i and P_i are the area and perimeter of the i th cell, respectively, A_0 is the preferred area of a cell (assumed to be uniform in the population of interest in the present scenario), and E_{ij} is the length of an edge between cells i and j (defined to be zero if the cells are not adjacent). The first term represents compression or dilation of the cell away from its preferred area and the last, the contractility of the perimeter. The constants, K and Γ that trade off the relative importance of these effects are held fixed.

The entire coding for differential adhesion takes place in the middle term of Equation 9. λ_{ij} is cost per unit length of an edge between the two cells. For the two-species case, this matrix has entries that are either zero for cells that are not adjacent, or values that depend on the kind of each cell. Heterotypic edges have one value and homotypic another. In what follows, we abuse the notation slightly and interpret $\lambda_{\alpha\beta}$ to mean the cost per unit length of an edge between cells of type α and β , and we use Λ to refer to the matrix of these costs for different cell types.

The simulation proceeds from randomly coloured cells on a regular hexagonal lattice, and the tissue is allowed to relax, in a direction that minimises the potential, rearranging according to the standard topological transitions for foams [33, 34]. To avoid settling to a local minimum, at each step vertices are subjected to some noise, an additional small force in a random direction.

We model the effect of tetracycline indirectly, representing the induced adhesion effect as the cost of edges. For these simulations we used values for heterotypic edges approximately twice as costly as for homotypic¹ and the result is a time-series of tissue exemplars beginning with cells randomly distributed and gradually developing more structure, or clustering. The claim [14] is that this sequence is representative of the process that occurs *in vitro*. Figure 3e shows how the absolute entropy, E_1 , of these tissue exemplars changes over time. Clearly it is decreasing overall.

Finally, we can use Equation 8 to work out the extent to which tissue snapshots from the numerical simulation are similar to the experimental data. The relative entropies of the simulation to each of the experimental cases, with and without tetracycline are shown in Figure 3f. Each has a minimum, and the minimum for the case with tetracycline occurs later,

¹In particular, $\lambda_{pp} = \lambda_{ee} = 0.05$ and $\lambda_{pe} = \lambda_{ep} = 0.096$, and in all cases $\Gamma = 0.04$ and $K = 1$

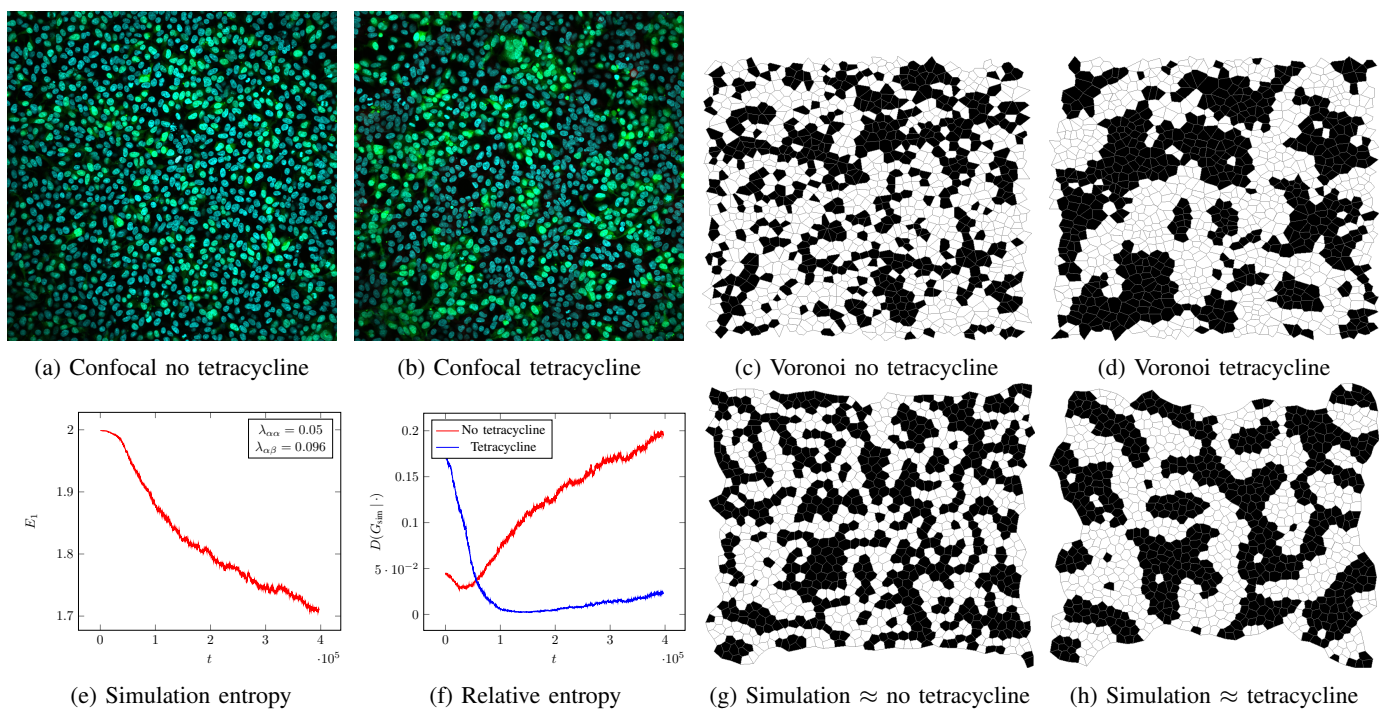


Figure 3: Top row experimental data, bottom row simulation data. Figures 3a and 3b show raw confocal images from Cachat et al.’s study of phase separation due to differential adhesion, after 24 hours. All cell nuclei are stained to appear blue, while only the nuclei of the E-cadherin variety appear green. Figures 3c and 3d show the graph derived from the voronoi tessellation of the cell centroids from the confocal images. Figure 3e shows the entropy trace of a typical simulation where heterotypic edges between cells are more costly than homotypic edges. Figure 3f shows the data from the same simulation, compared using relative entropy with the Voronoi tessellations of Figures 3c and 3d. Finally, Figures 3g and 3h show simulated tissues at the minimum of the relative entropy curves in Figure 3f, that is, those that correspond most closely to experiment by our measure. Both curves increase as the simulation becomes yet more clustered than the experiment.

at a stage where the simulation has become more ordered (lower absolute entropy) than without. The tissue exemplars corresponding to these two minima are shown in Figure 3g and 3h, and they correspond qualitatively well with their experimental counterparts. Our measure makes this impression quantitative.

Notice that the distance to the reference snapshot increases as the simulation progresses. This means that the simulated tissue, for these parameter values, becomes *more* ordered according to our measure than the experimental data. Given a suitably large library of simulation data with which to compare to experiment, one would naturally wish to find one where the distance measure converges to zero in order to make a well-supported claim that the simulation parameters are a good fit to the experiment.

IX. RATE OF PATTERN FORMATION

If a time-series of experimental data is available (unfortunately in this instance it is not) it is also possible to compare the rate of pattern formation. We can, however, show how path entropy can be used to quantify the rate of pattern formation with a set of numerical experiments. In these experiments we aim to understand more precisely how differential adhesion affects pattern formation. The salient model parameters are

the homotypic edge cost, $\lambda_{\alpha\alpha}$, which is held fixed, and the heterotypic edge cost, $\lambda_{\alpha\beta}$ which we allow to vary. Other parameters such as perimeter contractility, Γ , the area pressure constant, K , and the amount of noise, Z , we also hold fixed. The results are shown by plotting E_1 for various values of $\lambda_{\alpha\alpha}$ in Figure 4.

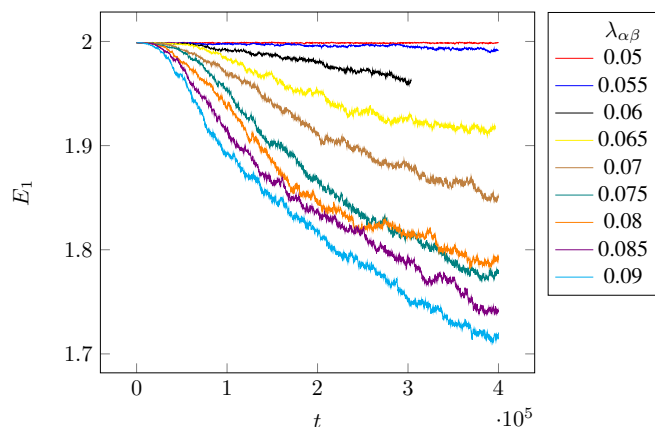


Figure 4: Path entropy time series for simulations with various values for the heterotypic edge cost, $\lambda_{\alpha\beta}$. In all cases the homotypic value is $\lambda_{\alpha\alpha} = 0.05$.

As suggested by the simple examples in Section VII and can be seen in Figure 3, lower entropy values correspond to more “dappling” patterns as they were described by Turing. The physical reason for the emergence of the pattern, in discrete models such as the Vertex or Cellular Potts models governed by a potential such as Equation 9 is quite simple. The heterotypic perimeter of a patch is expensive compared to the homotypic interior, so the dynamics simply arise from the process of minimising (up to the appropriate constants) the ratio of perimeter to surface area. Absent topological constraints, the shape that accomplishes this minimisation is a circle. In an equal mixture of cells constrained to be a planar graph, both kinds of cells cannot form circular patches simply because it is not possible to tile a plane with circles. Therefore competing but symmetric tendencies of each kind of cell to try to form circular patches results in the familiar pattern.

Given this understanding of the process, what we can read from the figure is, all else being equal, the greater the difference between the homotypic and heterotypic edge costs, the more rapidly the entropy of the tissue decreases. It takes about twice as long for the simulation with a heterotypic edge cost, $\lambda_{\alpha\beta} = 0.07$ as does the one for which $\lambda_{\alpha\beta} = 0.09$ to reach degree of pattern present that corresponds to $E_1 = 1.9$. When the heterotypic cost is only slightly larger than the homotypic cost, it may take much longer indeed to achieve that same degree of patterning.

Not shown are cases where the homotypic cost is allowed to vary, but the conclusions are straightforward and readily apparent from the time-series of our E_1 measure for them. Larger values of $\lambda_{\alpha\alpha}$ that are still smaller than $\lambda_{\alpha\beta}$ do result in patterning, but more slowly. This makes sense because these larger values are more rigid and as a result the entire system changes more slowly and the topological transitions that are necessary for pattern development due to cell migration less frequent. When $\lambda_{\alpha\alpha}$ is allowed to be greater than $\lambda_{\alpha\beta}$, the resulting pattern is very different because now rather than minimising the number of heterotypic edges they should be maximised. In this way we get patterns much like a checkerboard as predicted in Section VII. While these underlying mechanisms are well known, their effect is clearly exposed by studying the behaviour of E_1 .

X. EXAMPLES IN THREE COLOURS

The patterns in Figure 5 are all regular, except for the first, which is random. The random pattern is in fact representative of the initial conditions of the simulations which we will see later. They use three colours and a regular hexagonal lattice. This has some important consequences for the minimal entropy in a three-colour setting as we will see.

Figure 5 shows some example graphs in three colours and the corresponding path entropies. As usual, the number of cells of each colour is equal. Again, we include a randomly coloured graph and we include the generalisation of a checkerboard to a hexagonal lattice. We also include thin and thick stripes.

Some observations about the minima of the entropy can be made here and they are different from the two colour case. The example with thick stripes, or greater clustering, has lower

entropy than the others and it is a minimum by the same argument from Section VII, namely that any change can only increase the entropy by lessening the number of homogeneous edges.

In the three-colour case, however it has a lower entropy than the maximally dispersed, equivalent of the checkerboard. This is because it is not possible to colour a hexagonal lattice with only two colours while respecting the constraint that no two adjacent cells may have the same colour. Three colours are needed. This means that it is no longer true that for the maximally dispersed case all adjacencies are identical. The hexagonal checkerboard therefore no longer corresponds to a global minimum of E_1 . In fact the maximally clustered graph must now be the global minimum.

For these examples, the entropy for longer path lengths was also calculated directly. The results, shown in Figure 5e clearly illustrate that there is no benefit to the extra computational cost of calculating path entropy for paths of longer than 2 cells. This provides some further justification to our choice to confine our attention to E_1 . The reasoning about the minimum of E_1 for the three-colour case shows that this measure appropriately captures the degree of clustering or homogeneity.

XI. THREE-SPECIES EPITHELIA

Turning finally to the examples from Figure 1, we briefly study the patterning dynamics of epithelia consisting of three cells. We show that the E_1 metric can also be employed to evaluate whether one can distinguish the rate of pattern formation in systems with multiple cell types. As with the two-cell case, we consider interactions between cell types, but now form a 3x3 matrix,

$$\Lambda = \begin{bmatrix} \lambda_{rr} & \lambda_{rw} & \lambda_{rb} \\ \lambda_{wr} & \lambda_{ww} & \lambda_{wb} \\ \lambda_{br} & \lambda_{bw} & \lambda_{bb} \end{bmatrix} \quad (10)$$

accordingly as an edge is between red, r , white, w , or blue, b cells. We presume that this matrix is symmetric, and indeed it can always be symmetrised without changing the behaviour simply by taking, $\lambda'_{\alpha\beta} = \lambda'_{\beta\alpha} = \frac{1}{2}(\lambda_{\alpha\beta} + \lambda_{\beta\alpha})$.

We consider four cases, in an attempt to find a regime where the presence of a third kind of cell materially affects phase separation and pattern development. Namely,

- I. Homotypic red and blue edges are inexpensive, homotypic white edges are very expensive. Heterotypic edges with a white cell are very inexpensive and heterotypic red-blue edges are relatively expensive. In the absence of white cells, this behaves like the typical red-blue dappled pattern. Adding white cells should have them maximally dispersed.
- II. As with Case I, but the relationships to white cells inverted. Homotypic edges among white cells are now very inexpensive, and heterotypic ones are now very expensive. This is expected to form round patches of white cells.
- III. All homotypic edges have the same, low cost. Heterotypic edges with white cells are relatively inexpensive and red-blue edges are relatively expensive. The low cost

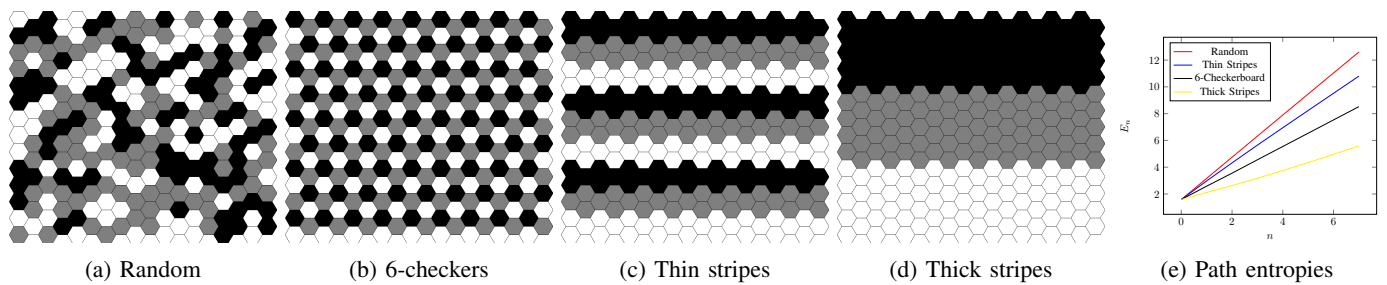


Figure 5: A selection of three-coloured planar graphs Figure 5e shows the path entropies, E_n , for these graphs, for path lengths n from 0 to 7.

of white-heterotypic edges produces long, thin, white borders between red and blue regions.

IV. As with Case III, but with the heterotypic costs inverted. Red-white and blue-white edges are now expensive and red-blue heterotypic edges are relatively inexpensive. This produces results very similar to Case II.

For these numerical experiments, in each case, the proportion of white cells was varied from 0 to 33%. The results of calculating time-series for E_1 are shown in Figure 6.

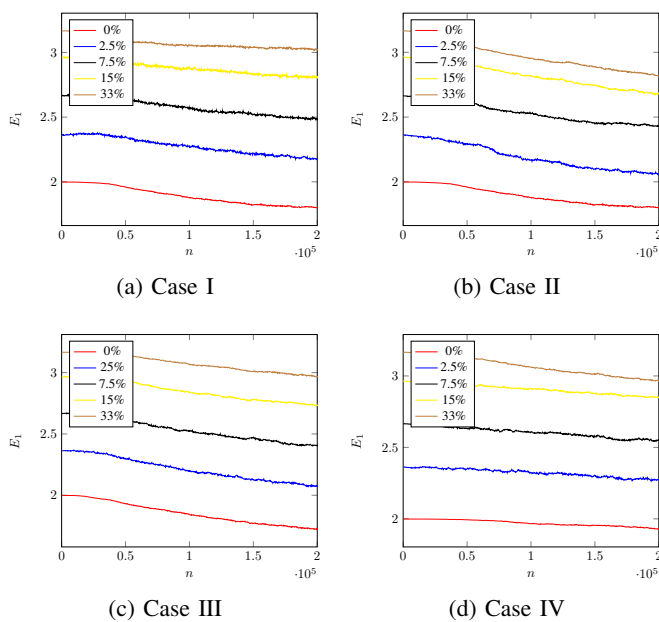


Figure 6: Entropy for various population fractions of white cells, for each of the cases above. The salient observation is that the slopes of the curves, the rate at which entropy changes, do not vary perceptibly with the amount of white cells.

This cursory search of four regions of the parameter space does not uncover a regime where a third kind of cell affects the rate or degree of pattern formation. This fact is made quite clear by the E_1 measure, whose rate of change is essentially the same for all of the cases. It remains an open area of research whether or not there is a regime where the presence of a third kind of cell accelerates or retards pattern formation by acting analogously to a lubricant or a glue.

XII. TISSUE LIBRARY FOR PARAMETER FITTING

A natural supposition, given this ability to measure how well patterns in simulated tissue graphs correspond to experimentally derived ones, is that we may be able to estimate the parameters in the Farhadifar potential, Equation 9, to the experimental data. This possibility is suggested by the observation that, not only is the degree of patterning measurable using our technique, so is the rate of pattern development. These two measures, E_1 and its time derivative could in principle be used for parameter estimation. Equally they could be used as predictors of experimental behaviour, for example estimating the concentration of a certain inducer required for a given rate of phase separation.

Up to normalisation, the parameters Λ and Γ do correspond to physical phenomena. This kind of fitting is indeed possible, with some limitations. The main limitation is that it is not possible to distinguish, within the region of interest between equally good pairs of parameters, (Λ, Γ) , along an iso-surface in the phase diagram [12, 13]. However, holding one fixed (Γ), it is indeed possible to derive an estimate of the corresponding value for Λ .

The procedure is simple but would require a large library of simulation data. For each parameter value, the time-series of E_1 can be calculated and stored, along with other statistics of interest (such as the degree distribution of cells). Data emanating from experimental imagery, processed into a coloured graph using the Voronoi tessellation or other techniques, can then be compared, and a best guess at the parameters arrived at. The time derivative of E_1 is important because often different adhesion values (Λ) that produce similar patterns can be distinguished by the rate at which the patterns appear.

Producing such a library is a very computationally intensive task. For the present work, we have simulated only a small subspace of possible parameter choices for two cells, and for three, and without necessarily reaching a steady state in all cases this has consumed several CPU-decades of processing time. Furthermore for accurate distributions, the tissue size should be as large as possible and at present tissues larger than about 5000 cells are prohibitive.

Despite the challenges, it is worthwhile to create and make available such a resource which the authors believe would be a valuable quantitative tool for synthetic morphology research.

XIII. CONCLUSIONS

Aside from application in synthetic morphology, the method presented can also be adapted to analyse samples of natural tissues and applied to the study of cancer progression. Recent successes using deep learning neural networks [35] to characterise cancer progression in tissue imagery samples are instructive. In that study accuracy rates with deep learning were comparable to trained pathologists but the technique does not permit inspection or reverse-engineering to identify the salient features being recognised. Successes using similar techniques have also been reported for identifying certain cardiovascular pathologies [36]. By contrast, our measure has much more stringent requirements on input data — we require input in the form of a coloured graph — but its principle of operation is straightforward to understand.

There is an important limitation when applying this technique to imagery from naturally occurring, as opposed to synthetic, tissue. Path Entropy is defined by cell types and their adjacencies. Synthetically engineered tissue designed to study mechanical interactions among cells is much more regular than its naturally occurring counterpart. This means it is correspondingly easier to extract the information needed to calculate path entropy from images of synthetic tissues. Accommodating structural heterogeneity in naturally occurring tissue likely requires segmentation techniques that consider actual cell boundaries and not a Voronoi tessellation derived from nuclei as we have done here. Advances in microscopy and optical technologies make possible high-throughput analysis and simultaneous measurements of proteins and other molecules (such as miRNA) in histological specimens and tissue micro-arrays. This allows the identification of subpopulations of genetically similar cells within tissue samples, using measurement of loci-specific fluorescence *in situ* Hybridization (FISH) spot signals for each nucleus [37, 38]. The use of neural networks to perform segmentation at the tissue level has been shown and remains a current topic of research [39–41]. These methodologies could facilitate the construction of the graph underlying an epithelial tissue and suggest an appropriate extension of the metric proposed in this work.

In this paper, we have defined a specialised class of entropy measures, path entropies, on adjacency graphs designed to quantify the degree of patterning present in cellular tissues and noted some of its interesting properties. We have demonstrated how this measure can be used on two dimensional epithelial tissues to establish a correspondence between experimental and simulation data that quantifies the impression of similarity between the patterns expressed. We have further demonstrated how the measure generalises to tissues consisting of three species and noted some differences from the two species case. Finally, we have proposed, for the specific application of synthetic morphology, the establishment of a library of tissue data upon which these measures can be calculated, to assist in parameter estimation, providing a useful quantitative tool for synthetic morphology.

XIV. ACKNOWLEDGEMENTS

W.W., M.C. acknowledge the support from the Engineering and Physical Sciences Research Council (EPSRC) grant

EP/J02175X/1 and from UK Research Councils' Synthetic Biology for Growth programme, the BBSRC, EPSRC and the MRC. W.W. also acknowledges support from the National Academies Keck Futures Initiative of the National Academy of Sciences award number NAKFI CB12. J.A.D. and E.C. acknowledge the support from the Biotechnology and Biological Sciences Research Council (BBSRC) grant BB/M018040/1 and the Leverhulme Trust grant RPG-2012-558.

Computational resources were supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) and by Microsoft's donation of Azure time to the Alan Turing Institute.

The authors wish to thank N. Behr, M. Black, P. Buneman, A. Clisham, R. Farhadifar, A. Fletcher and G. Plotkin and the anonymous reviewers for helpful discussion and criticism which greatly improved the quality of this paper.

Data processing was driven by the excellent GNU Parallel [42].

REFERENCES

- [1] M. S. Steinberg, "On the mechanism of tissue reconstruction by dissociated cells, III. Free energy relations and the reorganization of fused, heteronomic tissue fragments," *Proceedings of the National Academy of Sciences*, vol. 48, no. 10, pp. 1769–1776, 1962.
- [2] E. Cachat, W. Liu, K. C. Martin, X. Yuan, H. Yin, P. Hohenstein, and J. A. Davies, "2- and 3-dimensional synthetic large-scale de novo patterning by mammalian cells through phase separation," *Scientific Reports*, vol. 6, p. 20664, Feb. 2016. [Online]. Available: <http://www.nature.com/srep/2016/160209/srep20664/full/srep20664.html>
- [3] M. Unbekandt and J. A. Davies, "Dissociation of embryonic kidneys followed by reaggregation allows the formation of renal tissues," *Kidney International*, vol. 77, no. 5, pp. 407–416, Mar. 2010. [Online]. Available: [http://kidney-international.theisn.org/article/S0085-2538\(15\)54273-7/abstract](http://kidney-international.theisn.org/article/S0085-2538(15)54273-7/abstract)
- [4] J. G. Lefevre, H. S. Chiu, A. N. Combes, J. M. Vanslambrouck, A. Ju, N. A. Hamilton, and M. H. Little, "Self-organisation after embryonic kidney dissociation is driven via selective adhesion of ureteric epithelial cells," *Development*, vol. 144, no. 6, pp. 1087–1096, Mar. 2017. [Online]. Available: <http://dev.biologists.org/content/144/6/1087>
- [5] A. M. Turing, "The chemical basis of morphogenesis," *Phil. Trans. R. Soc. Lond. B*, vol. 237, pp. 37–72, 1952.
- [6] A. Gierer and H. Meinhardt, "A theory of biological pattern formation," *Biological Cybernetics*, vol. 12, no. 1, pp. 30–39, 1972.
- [7] J. W. Cahn and J. E. Hilliard, "Free energy of a nonuniform system. i. interfacial free energy," *The Journal of chemical physics*, vol. 28, no. 2, pp. 258–267, 1958.
- [8] F. Graner and J. A. Glazier, "Simulation of biological cell sorting using a two-dimensional extended Potts model," *Physical Review Letters*, vol. 69, no. 13, pp. 2013–2016, Sep. 1992. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevLett.69.2013>
- [9] S. Sandersius, C. J. Weijer, and T. J. Newman, "Emergent cell and tissue dynamics from subcellular modeling of active biomechanical processes," *Physical biology*, vol. 8, no. 4, p. 045007, 2011.
- [10] T. Nagai and H. Honda, "A dynamic cell model for the formation of epithelial tissues," *Philosophical Magazine Part B*, vol. 81, no. 7, pp. 699–719, Jul. 2001. [Online]. Available: <http://dx.doi.org/10.1080/13642810108205772>

- [11] R. Farhadifar, J.-C. Röper, B. Aigouy, S. Eaton, and F. Jülicher, "The Influence of Cell Mechanics, Cell-Cell Interactions, and Proliferation on Epithelial Packing," *Current Biology*, vol. 17, no. 24, pp. 2095–2104, Dec. 2007. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0960982207023342>
- [12] D. B. Staple, R. Farhadifar, J.-C. Röper, B. Aigouy, S. Eaton, and F. Jülicher, "Mechanics and remodelling of cell packings in epithelia," *The European Physical Journal E*, vol. 33, no. 2, pp. 117–127, Nov. 2010. [Online]. Available: <http://link.springer.com/article/10.1140/epje/i2010-10677-0>
- [13] R. Magno, V. A. Grieneisen, and A. F. M. Marée, "The biophysical nature of cells: potential cell behaviours revealed by analytical and computational studies of cell surface mechanics," *BMC Biophysics*, vol. 8, no. 1, p. 8, May 2015. [Online]. Available: <https://doi.org/10.1186/s13628-015-0022-x>
- [14] J. M. Osborne, A. G. Fletcher, J. M. Pitt-Francis, P. K. Maini, and D. J. Gavaghan, "Comparing individual-based approaches to modelling the self-organization of multicellular tissues," *PLOS Computational Biology*, vol. 13, no. 2, p. e1005387, Feb. 2017. [Online]. Available: <https://doi.org/10.1371/journal.pcbi.1005387>
- [15] D. Sánchez-Gutiérrez, M. Tozluoglu, J. D. Barry, A. Pascual, Y. Mao, and L. M. Escudero, "Fundamental physical cellular constraints drive self-organization of tissues," *The EMBO Journal*, vol. 35, no. 1, pp. 77–88, 2015.
- [16] P. Vicente-Munuera, P. Gomez-Galvez, A. Tagua, M. Letran, Y. Mao, and L. M. Escudero, "EpiGraph: an open-source platform to quantify epithelial organization," *bioRxiv*, p. 217521, 2017. [Online]. Available: <http://biorxiv.org/content/early/2017/11/13/217521>
- [17] M. A. Eshera and K.-S. Fu, "A graph distance measure for image analysis," *IEEE transactions on systems, man, and cybernetics*, pp. 398–408, 1984.
- [18] H. Bunke and K. Shearer, "A graph distance metric based on the maximal common subgraph," *Pattern recognition letters*, vol. 19, no. 3, pp. 255–259, 1998.
- [19] C. E. Shannon, "A Mathematical Theory of Communication," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, Jan. 2001. [Online]. Available: <http://doi.acm.org/10.1145/584091.584093>
- [20] E. T. Jaynes, "Gibbs vs boltzmann entropies," *American Journal of Physics*, vol. 33, no. 5, pp. 391–398, 1965.
- [21] D.-Y. Tsai, Y. Lee, and E. Matsuyama, "Information entropy measure for evaluation of image quality," *Journal of digital imaging*, vol. 21, no. 3, pp. 338–347, 2008.
- [22] R. C. Gonzalez, S. L. Eddins, and R. E. Woods, *Digital Image Publishing Using MATLAB*. Prentice Hall, 2004.
- [23] J.-F. Mangin, "Entropy minimization for automatic correction of intensity nonuniformity," in *Mathematical methods in biomedical image analysis, 2000. proceedings. ieee workshop on*. IEEE, 2000, pp. 162–169.
- [24] J. Skilling and R. Bryan, "Maximum entropy image reconstruction: general algorithm," *Monthly notices of the royal astronomical society*, vol. 211, no. 1, pp. 111–124, 1984.
- [25] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International journal of computer vision*, vol. 40, no. 2, pp. 99–121, 2000.
- [26] M. Cuturi and A. Doucet, "Fast computation of wasserstein barycenters," in *International Conference on Machine Learning*, 2014, pp. 685–693.
- [27] K. G. Larkin, "Reflections on shannon information: In search of a natural information-entropy for images," *CoRR*, vol. abs/1609.01117, 2016. [Online]. Available: <http://arxiv.org/abs/1609.01117>
- [28] L. M. Escudero, L. d. F. Costa, A. Kicheva, J. Briscoe, M. Freeman, and M. M. Babu, "Epithelial organisation revealed by a network of cellular contacts," *Nature communications*, vol. 2, p. 526, 2011.
- [29] M. E. Stickel and M. Tyson, "An analysis of consecutively bounded depth-first search with applications in automated deduction," in *IJCAI*, 1985, pp. 1073–1075.
- [30] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951. [Online]. Available: <http://www.jstor.org/stable/2236703>
- [31] G. R. Mirams, C. J. Arthurs, M. O. Bernabeu, R. Bordas, J. Cooper, A. Corrias, Y. Davit, S.-J. Dunn, A. G. Fletcher, D. G. Harvey, and others, "Chaste: an open source C++ library for computational physiology and biology," *PLoS Comput Biol*, vol. 9, no. 3, p. e1002970, 2013.
- [32] C. S. Smith, "Grain Shapes and Other Metallurgical Applications of Topology," *Metal Interfaces*, 1952.
- [33] D. Weaire and N. Rivier, "Soap, cells and statistics – random patterns in two dimensions," *Contemporary Physics*, vol. 50, no. 1, pp. 199–239, Jan. 2009. [Online]. Available: <http://dx.doi.org/10.1080/00107510902734680>
- [34] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, p. 115, 2017.
- [35] C. Xu, L. Xu, Z. Gao, S. Zhao, H. Zhang, Y. Zhang, X. Du, S. Zhao, D. Ghista, and S. Li, "Direct detection of pixel-level myocardial infarction areas via a deep-learning algorithm," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 240–249.
- [36] C. M. Croce, "Oncogenes and cancer," *New England Journal of Medicine*, vol. 358, no. 5, pp. 502–511, 2008.
- [37] M. Guillaud, C. Clem, and C. MacAulay, "An in silico platform for the study of epithelial pre-invasive neoplastic development," *Biosystems*, vol. 102, no. 1, pp. 22–31, 2010.
- [38] W. E. Reddick, J. O. Glass, E. N. Cook, T. D. Elkin, and R. J. Deaton, "Automated segmentation and classification of multispectral magnetic resonance images of brain using artificial neural networks," *IEEE Transactions on medical imaging*, vol. 16, no. 6, pp. 911–918, 1997.
- [39] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen, "Deep convolutional neural networks for multi-modality iso-intense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, 2015.
- [40] S. Su, Z. Gao, H. Zhang, Q. Lin, W. K. Hau, and S. Li, "Detection of lumen and media-adventitia borders in ivus images using sparse auto-encoder neural network," in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. IEEE, 2017, pp. 1120–1124.
- [41] O. Tange, "GNU parallel - the command-line power tool," *login: The USENIX Magazine*, vol. 36, no. 1, pp. 42–47, Feb 2011. [Online]. Available: <http://www.gnu.org/s/parallel>