Estimating cellular redundancy in networks of genetic expression

Raffaella Mulas[1,2,3] and Michael J. Casey[2,3]

[1] *The Alan Turing Institute, London, UK*
[2] *Mathematical Sciences, University of Southampton, UK*
[3] *Institute of Life Sciences, University of Southampton, UK*

Abstract

Networks of genetic expression can be modelled by hypergraphs with the additional structure that real coefficients are given to each vertex-edge incidence. The spectra, i.e. the multiset of the eigenvalues, of such hypergraphs, are known to encode structural information of the data. We show how these spectra can be used, in particular, in order to give an estimation of cellular redundancy of the network. We analyze some simulated and real data sets of gene expression for illustrating the new method proposed here.

1 Introduction

Single-cell RNA-sequencing experiments are ubiquitous in cell biology, measuring the expression of each gene (number of mRNA molecules) in single cells [14]. The resulting data are complex. Cells can form discrete clusters — cell types — or continuous relations — lineage trajectories. Genes inter-regulate each other's expression, forming non-linear gene expression programmes [7, 27, 39]. We introduce a measure for directly quantifying cellular complexity based on spectral hypergraph theory: an emerging field of mathematics that has been recently applied to physics [2–4, 6, 8, 23], and that can be applied to biochemical data analysis [15].

Graph theory forms a significant component of single-cell analysis (single-cell analysis referring to the analysis of single-cell data, not of single cells), particularly in dimension reduction and unsupervised clustering [5, 22, 38, 40]. By encoding data as graphs, the properties of graphs become the properties of the data, gaining us access to many powerful quantitative summaries of the data. But graphs can only partially represent single-cell data (typically encoding cell-to-cell distances): hypergraphs, a generalization of graphs, allow for a complete representation of the data. We therefore encode single-cell data as hypergraphs, and we show how recent advances in spectral hypergraph theory enable us to access the succinct structural properties of these data.

While the edges of a graph are pairwise connections between the vertices, with hypergraphs, links of any cardinality between the nodes are allowed. In other words, in a hypergraph, edges are *sets* of nodes. Here we consider an even more general class of hypergraphs, namely the *hypergraphs with real coefficients* that were introduced in [15], in which real coefficients can be assigned to the vertex-edge incidences (Figure 1).

Given a data set of gene expression with N cells and M genes, we model it as a hypergraph on N nodes and M edges in which each vertex i represents a cell, each edge
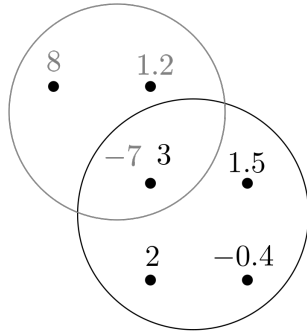
Figure 1: A hypergraph with real coefficients.

k represents a gene, and each coefficient $c(i, k)$ represents the proportion of transcripts of k assigned to i. We assume that the coefficients are normalized with respect to cells, so that

$$\sum_k c(i, k) = 1 \quad \text{for each cell } i. \tag{1}$$

Such cell-wise normalization is the norm in RNA-seq analysis (e.g. counts per million): the number of counts per cell can vary by orders of magnitude solely due to technical effects, distorting comparisons of absolute gene expression between cells [9, 17, 37].

We then consider the spectrum of the normalized Laplacian of the obtained hypergraph. As shown in [15], such spectrum is given by N real, non-negative eigenvalues, and it encodes many important structural properties of the hypergraph, therefore also of the data. If we denote by λ_N the largest eigenvalue, then the quantity

$$\mathcal{R} := \frac{\lambda_N}{N}$$

estimates the *cellular redundancy* of the network, and it is an inverse measure of cellular complexity. In fact, as we shall see in Section 2, we have that

$$\frac{1}{N} \leq \mathcal{R} \leq 1,$$

and $\mathcal{R} = 1/N$ if and only if and only if each gene is concentrated in one single cell, that is, we have no cellular redundancy at all. Moreover, we have that $\mathcal{R} = 1$ (equivalently, λ_N achieves its largest possible value N) if and only if each gene is distributed among all cells and

$$c(i, k) = c(j, k)$$

for each gene k and for all cells $i \neq j$, i.e., we have cellular redundancy.

We illustrate the meaning of this redundancy with two toy examples that involve two cells and two genes each. If genes are distributed among cells as in Table 1, we clearly expect a high cellular redundancy and in fact, by the above consideration, $\mathcal{R} = 1$, hence $\mathcal{R}$ achieves its maximum possible value. Conversely, if genes are distributed among cells as in Table 2, we then have low cellular redundancy and we therefore expect a small value of $\mathcal{R}$. As we shall see in details in Example 2.6 below, this is exactly the case.

Structure of the paper

In Section 2 we present the mathematical details of our model and we prove, in particular, our main theoretical results. In order to illustrate the method, in Section 3 we analyze

	Gene 1	Gene 2
Cell 1	0.3	0.7
Cell 2	0.3	0.7

Table 1: An example of gene expression with maximal cellular redundancy.

	Gene 1	Gene 2
Cell 1	0.1	0.9
Cell 2	1	0

Table 2: An example of gene expression with low cellular redundancy.

simulated data and in Section 4 we analyze real data sets of gene expressions. Finally, in Section 5 we discuss the methods and in Section 6 we draw some conclusions.

2 Theoretical results

We recall some definitions from [15] before proving our main theoretical results.

Definition 2.1. A *hypergraph with real coefficients* is a triple $\Gamma = (V, E, \mathcal{C})$ such that:

- $V = \{1, \ldots, N\}$ is a finite set of *nodes* or *vertices*;

- $E = \{k_1, \ldots, k_M\}$ is a multiset of elements $k_l \in \mathcal{P}(V) \setminus \emptyset$ called *edges*;

- $\mathcal{C} = \{c(i,k) : i \in V \text{ and } k \in E\}$ is a set of *coefficients* $c(i,k) \in \mathbb{R}$ and it is such that

$$c(i,k) = 0 \iff i \notin k. \tag{2}$$

From here on in this section we fix a hypergraph $\Gamma = (V, E, \mathcal{C})$ on N nodes and M edges. We assume that Γ has no isolated vertices, i.e., vertices that are not contained in any edge.

Definition 2.2. Given $k \in E$, its *cardinality* is the number of vertices that are contained in k. Given $i \in V$, its *degree* is

$$\deg i := \sum_{k \in E} c(i,k)^2. \tag{3}$$

The $N \times N$ diagonal *degree matrix* of Γ is

$$D := \text{diag}\big(\deg i\big)_{i=1,\ldots,N}.$$

The $N \times N$ *adjacency matrix* of Γ is $A := (A_{ij})_{ij}$, where $A_{ii} := 0$ for all $i = 1, \ldots, N$ and

$$A_{ij} := -\sum_{k \in E} c(i,k) \cdot c(j,k) \quad \text{for all } i \neq j.$$

The $N \times M$ *incidence matrix* of Γ is $\mathcal{I} := (\mathcal{I}_{il})_{il}$, where

$$\mathcal{I}_{il} := c(i, k_l).$$

Note that in our context, in both models that we are considering, each row of $\mathcal{I}$ represents a cell and each column of $\mathcal{I}$ represents a gene.

Definition 2.3. The *normalized Laplacian* of Γ is the $N \times N$ matrix

$$L := \mathrm{Id} - D^{-1}A = D^{-1}\mathcal{I}\mathcal{I}^\top.$$

The *dual normalized Laplacian* of Γ is the $M \times M$ matrix

$$L^* := \mathcal{I}^\top D^{-1}\mathcal{I}.$$

As shown in [15], L has N real, non-negative eigenvalues, denoted $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_N$ and called the *spectrum of* Γ, while L^* has M real, non-negative eigenvalues, counted with multiplicity. The non-zero eigenvalues of L and L^* coincide, with the same multiplicity, therefore, in particular, also the largest eigenvalue of L and L^* is the same. This allows us to simplify the computations depending on the size of the data set: If $M \geq N$ (respectively, $N \geq M$), it is convenient to compute the largest eigenvalues and, more generally, the entire non-zero spectrum of Γ, using L (respectively, L^*).

We now prove a general result on the largest eigenvalue of Γ, before focusing on the case when the coefficients are normalized as in (1).

Theorem 2.4. *The largest eigenvalue of Γ is such that $1 \leq \lambda_N \leq N$. Moreover, if all coefficients are non-negative,*

- *$\lambda_N = 1$ if and only if each edge has cardinality 1;*

- *$\lambda_N = N$ if and only if all edges have cardinality N and, for all vertices $i \neq j$ and edges $k \neq h$,*

$$\frac{c(i,k)}{c(j,k)} = \frac{c(i,h)}{c(j,h)}. \tag{4}$$

Proof. As shown in [15, Corollary 3.2], $\sum_{i=1}^N \lambda_i = N$. Since λ_N is the largest eigenvalue, this implies that $1 \leq \lambda_N \leq N$, and $\lambda_N = 1$ if and only if $\lambda_i = 1$ for all $i = 1, \ldots, N$. Now, since $L = \mathrm{Id} - D^{-1}A$, 1 is the only eigenvalue of Γ if and only if $A = 0$. Since all coefficients are non-negative, this happens if and only if two distinct vertices are not contained in common edges, which can be reformulated by saying that each edge has cardinality 1. This proves the first claim.

Now, by Theorem 6.3 in [15], since in our case the coefficients are non-negative, $\lambda_N = N$ if and only if all edges have cardinality N and there exists a function $f : V \to \mathbb{R}$ such that

$$g(k) := c(i,k)f(i)$$

does not depend on i, for all edges k and for all vertices i. This is equivalent to saying that

$$c(i,k)f(i) = c(j,k)f(j)$$

for all $i \neq j$, or equivalently

$$f(j) = \frac{c(i,k)}{c(j,k)}f(i).$$

Such f exists if and only if, for all $i \neq j$, $c(i,k)/c(j,k)$ does not depend on k, that is,

$$\frac{c(i,k)}{c(j,k)} = \frac{c(i,h)}{c(j,h)}$$

for all $k \neq h$ and for all $i \neq j$. $\qquad \square$

We now consider the case when the coefficients satisfy the normalization in (1).

Corollary 2.5. *If the coefficients satisfy* (1)*, then the largest eigenvalue is such that* $1 \leq \lambda_N \leq N$ *and, moreover,*

- $\lambda_N = 1$ *if and only if each edge has cardinality 1;*

- $\lambda_N = N$ *if and only if all edges have cardinality N and, for each edge k and for all vertices $i \neq j$,*

$$c(i, k) = c(j, k). \tag{5}$$

Proof. By Theorem 2.4 it is enough to prove that, if the coefficients satisfy (1), then (4) is equivalent to (5). Clearly, (5) implies (4). Vice versa, (4) together with (1) implies that

$$1 = \sum_h c(i, h) = \left(\sum_h c(j, h) \right) \cdot \frac{c(i, k)}{c(j, k)} = \frac{c(i, k)}{c(j, k)}.$$

Hence, $c(i, k) = c(j, k)$ for each edge k and for all vertices $i \neq j$. $\qquad\square$

In our context of networks of genetic expression, we can interpret Corollary 2.5 as follows. The largest eigenvalue λ_N achieves its smallest possible value 1 if and only if each gene is concentrated in one single cell. Moreover, if the coefficients satisfy the normalization in (1), then $\lambda_N = N$ if and only if each gene is uniformly distributed among all cells or, equivalently, we have cellular redundancy. In general, the bigger λ_N is, the more cellular redundancy there is. We choose to consider the normalized quantity $\mathcal{R} = \lambda_N / N$ as a measure of redundancy, so that is independent of the number of cells.

We conclude this section by discussing the example from the Introduction.

Example 2.6. In the example shown by Table (2),

$$\mathcal{I} = \begin{pmatrix} 0.1 & 0.9 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad D = \begin{pmatrix} 0.82 & 0 \\ 0 & 1 \end{pmatrix}.$$

The normalized Laplacian $L = D^{-1} \mathcal{I} \mathcal{I}^\top$ has eigenvalues

$$\lambda_1 \sim 0.89 \quad \text{and} \quad \lambda_2 \sim 1.11.$$

Hence, while for a general hypergraph on 2 nodes the redundancy $\mathcal{R} = \lambda_2 / 2$ is such that $0.5 \leq \mathcal{R} \leq 1$, in this case we have $\mathcal{R} \sim 0.55$, i.e., there is a very low cellular redundancy, as expected.

3 Simulations

We first demonstrate our hypergraph approach on simulated data. Single-cell RNA-sequencing is a counting process, measuring the number of transcripts expressed by each gene in each cell; accordingly, we simulate each gene as a Poisson process [28]. In practice, single-cell sequencing counts are over-dispersed compared with the Poisson, with additional variability arising from biological heterogeneity. However, for the simulation, we assume there is no excess biological variability, so parameterize each gene by a single mean parameter, μ [28, 34, 37]. We randomly generate μ for each gene by sampling from a gamma distribution (rate = 0.3; shape = 0.6) [41].

We repeatedly simulate 1800 genes over a population of 300 cells (following the median ratio of genes to cells found in the real data sets shown later). Any differences between cells arise solely from stochastic variation in sampling, so we expect cells to be highly redundant: over the course of 100 simulations, we find a mean cellular redundancy of 0.842 (see Table 3).

Data	$\mathcal{R}$
Simulated ($C = 1$)	0.842
Simulated ($C = 2$)	0.447
Simulated ($C = 3$)	0.316

Table 3: Cellular redundancy of simulations.

The principal biological source of heterogeneity in gene expression is a multiplicity of distinct cell types. Different cells types express distinct sets of genes (known as differential gene expression), introducing multimodality in the observed count distributions [30].

We recreate the effect of cell types in silico by splitting cells into C discrete subpopulations. Each subpopulation is defined by a tranche of highly-expressed genes (of size $1800/C$, rate = 0.3; shape = 0.6); the highly-expressing tranches of genes do not overlap between differing subpopulations, and genes not included in the high-expression tranche have on average 10x less expression (of size 1800- $1800/C$, rate = 3; shape = 0.6).

Visualizing $C = 3$ via principal component analysis (projection of high dimensional data set onto two dimensions, maximizing the variance retained), we observe that the subpopulations are clearly defined, with no overlap; moreover, the heterogeneity within each subpopulation is far less than the heterogeneity between subpopulations [12]. We, therefore, expect a sharp decline in cellular redundancy with the increasing number of differentially expressing subpopulations.

Simulating each $C \in \{1, 2, 3\}$ 100 times, we find a strong, significant negative correlation between the number of clusters and cellular redundancy (Spearman's rho = -0.94, p-value < 2.2e-16, two-sided, N = 100): for the simulated data, cellular redundancy captures the increase in population complexity.

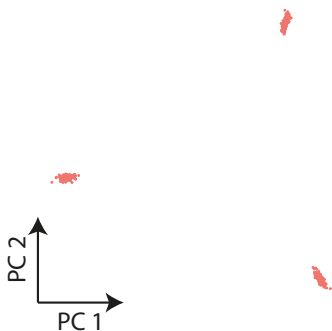
Figure 2: Simulation of three distinct clusters. Cells are projected onto the first two dimensions of a principal components analysis of genes.

4 Analysis of real data sets

We further demonstrate our hypergraph approach on three real data sets: 1) Svensson, a technical control data set (4,000 cells-equivalents, 24,116 gene-equivalents); 2) Tian, a mixture of three cancerous cell lines, representing a simple, discrete cluster structure (902 cells, 16,468 genes), and 3) Stumpf, a sampling from mouse bone marrow, representing a complex, continuous cluster structure (5,504 cells, 16,519 genes) [33, 35, 36].

Data	$\mathcal{R}$
Svensson	0.953
Tian	0.825
Stumpf	0.398

Table 4: Cellular redundancy of technical (Svensson) and biological data.

The technical control data set has been generated from a mixed solution of RNA (specifically endogenous RNA from human brain and External RNA Control Consortium spike-ins): variation between the cell-equivalents (the 'cells' are not strictly cells, rather technical equivalents) arises solely from technical measurement effects [34]. The data consists of two experimental batches, each of which forms a distinct cluster of cells in the first two principal components, albeit with some overlap (Figure 3a).

Tian separates into three clear clusters in the first two principal components (Figure 3b). Unlike with the simulated data, not all genes will be involved in the demarcation of cellular groupings — there will be substantially more overlap in gene expression between cells and so greater cellular redundancy. Moreover, even among those genes that define clusters, they may define multiple clusters or not be as differentially expressed (10-fold difference) as in the simulated data: we expect real data to display higher levels of cellular redundancy than the corresponding simulations.

This overlap in gene expression is emphasized in the Stumpf data set: cell types clusters are not discrete; instead, they form part of continuous cell differentiation trajectories
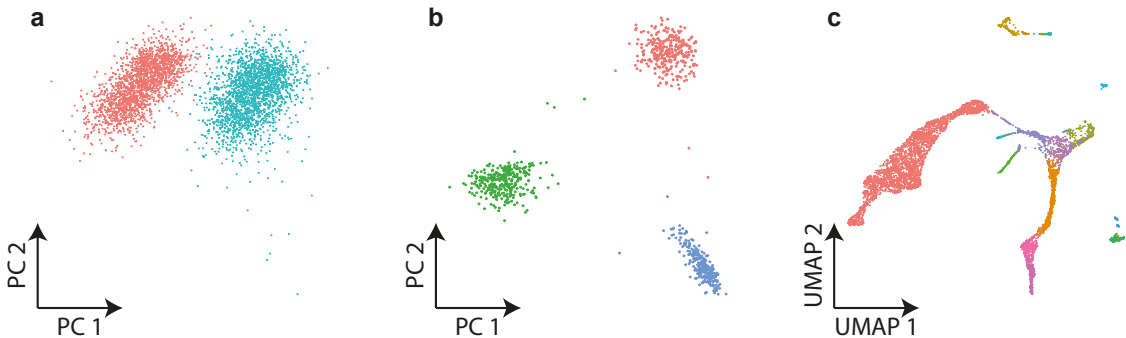

Figure 3: Simulation of three data sets: **a)** Svensson, **b)** Tian, and **c)** Stumpf. In each, cells are colored by cell type (recovered from data metadata) except for c) which is colored by experimental batch. a) and b) project cells onto principal components, and c) onto UMAP dimensions (Uniform Manifold Approximation and Projection), a nonlinear dimension reduction technique [22].
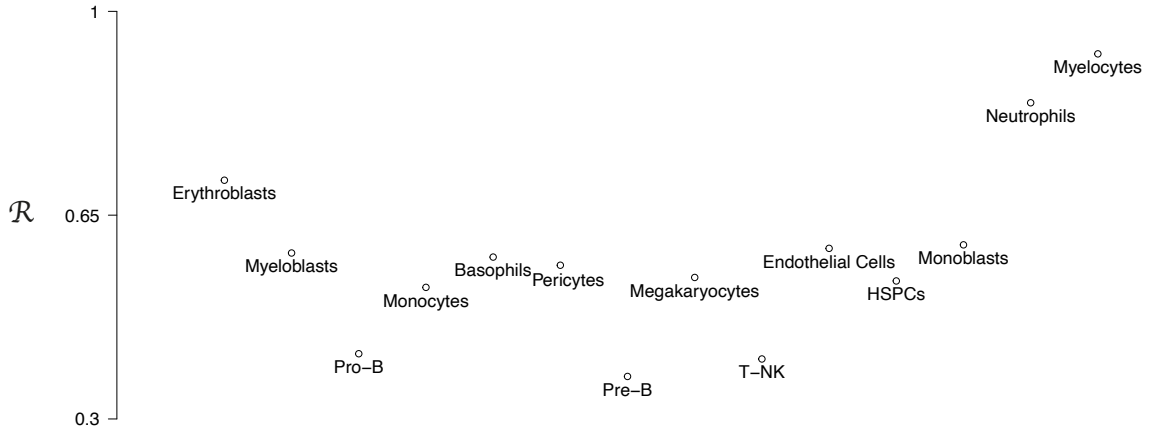
Figure 4: Cellular redundancy by cell type for Stumpf data set.

(Figure 3c). Thus, despite having substantially more clusters than any other data set, we expect the decline in redundancy to be muted by the continuous nature of the clusters.

We find that both Svensson and Tian have high levels of cellular redundancy. Tian in particular has a cellular redundancy on par with our simulation of a single subpopulation: overlap in gene expression substantially increases cellular redundancy. Moreover, Stumpf has a cellular redundancy midway between the simulation of two or three subpopulations: continuous cellular identity increases the measure cellular redundancy.

Calculating the cellular redundancy of each cell type within the Stumpf data set, we find that most types share a similar level of redundancy, with the notable exceptions of the Pro-B, Pre-B and T-NK (Natural Killer) cell types, and the Neutrophil and Myelocyte cell types (Figure 4). Both sets of cell types form coherent developmental lineages, the lymphocyte and neutrophil lineages respectively [33]. Given the consistency otherwise of cell type's cellular redundancy, we suggest that redundancy may detect over-/under-clustering of cells: the lymphocyte-lineage 'cell types' may capture a broader continuum of cellular development, so have reduced redundancy; conversely, the neutrophile-lineage 'cell types' may represent cell sub-types, with greater cellular redundancy.

5 Methods

5.1 Simulation

We simulated each gene as a Poisson process, with μ randomly generated from a gamma distribution. We used a pair of gamma distributions: one with a rate of 0.3 for highly-expressed genes and one with a rate of 3 for lowly-expressed genes; both parametrizations used a shape parameter of 0.6.

For each simulation, we simulated a total of 300 cells and 1800 genes. For C of 1, i.e., one subpopulation, all genes for all cells were highly expressing; for C of 2 or 3, half or a third of genes were highly expressed for half or a third of genes, respectively.

We calculated the mean redundancies for 100 repeats of each simulation. Spearman's rho who was calculated using the R function *cor.test*.

We used the R package Seurat for principle component analyses and UMAP visualizations [32]. Code for simulations and calculation of redundancy measures is available on

GitHub (`https://github.com/mjcasy/scHyperGraph`).

5.2 Technical and Biological Data collection

The Svensson data was downloaded as file "svensson_chromium_control.h5a" from `https://data.caltech.edu/records/1264`

The Tian data was downloaded as file "sincell_with_class.RDat" from `https://github.com/LuyiTian/sc_mixology`

The Stumpf data was downloaded as file "RData.zip" from `https://data.mendeley.com/datasets/csvm3kpkxd/1?fbclid=IwAR3j_hvq5Zt0cdxBBM72Fqr8zXwVC6XRpkY2JtVwbcj1gzyNLMVsdoFYWgQ`

6 Conclusions

We have introduced hypergraphs to the analysis of networks of genetic expression. We have identified, in particular, a measure that summarizes cellular redundancy and which can be obtained as follows. Given a network of genetic expression with N cells, in which the proportions of transcripts of genes assigned to each cell sum to one, we model it as a hypergraph with real coefficients in which each vertex i models a cell, each edge k models a gene, and each vertex-edge coefficients $c(i, k)$ is the proportion of transcripts of k assigned to i. The assumption on the sum of the transcripts, following existing cell-wise normalization strategies [9, 17, 37], can then be reformulated as $\sum_k c(i, k) = 1$, for each cell/vertex i. As shown in [15], the normalized Laplacian matrix associated to this hypergraph has N real, non-negative eigenvalues $\lambda_1 \leq \ldots \leq \lambda_N$ whose sum is N. These can be easily computed and encode qualitative properties of the hypergraph. The properties of the hypergraph can be translated into properties of the data and, therefore, the eigenvalues represent a signature of the data that can be computed with little computational effort.

We have shown, in particular, that the largest eigenvalue λ_N measures how much cellular redundancy is present in the network. Since such measure depends on the number of cells N, we considered $\mathcal{R} = \lambda_N/N$ as a normalized measure of cellular redundancy of the network. The choice of this measure is motivated by the theoretical results that we presented and, in order to illustrate our method, we have also analyzed both simulated and real data sets of gene expression. In both types of data, we find cellular redundancy recapitulates the expected complexity of the data: redundancy reduces with an increasing number of subpopulations in the simulated data and with increasing biological complexity in the real. Moreover, when broken down by cell type, cellular redundancy identified two distinct developmental lineages with increased and decreased redundancy respectively, suggesting cellular redundancy is informative on the nature of cell type.

In conclusion, in this work we have taken a step further in the study of spectral hypergraph theory and its applications to biology. It is worth mentioning that, while there is a vast literature on spectral graph theory applied to biochemical networks [1, 13, 18–21, 25, 31], as well as a growing literature on hypergraph modelling in biology [10, 11, 16, 26, 29], very little has been done, so far, in the study of spectral hypergraph theory applied to biology [15, 24]. Therefore, for future works, we aim to build upon the work presented here practically, applying the measure proposed here to a greater diversity of biological data, and theoretically, developing further novel measures and methods for biological data analysis from the spectra of hypergraphs.

References

[1] A. Banerjee and J. Jost. Graph spectra as a systematic tool in computational biology. *Discrete Applied Mathematics*, 157(10):2425–2431, 2009.

[2] A. Banerjee and S. Parui. On synchronization in coupled dynamical systems on hypergraphs. arXiv:2008.00469.

[3] F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, and G. Petri. Networks beyond pairwise interactions: Structure and dynamics. *Physics Reports*, 874:1 – 92, 2020.

[4] T. Böhle, C. Kuehn, R. Mulas, and J. Jost. Coupled Hypergraph Maps and Chaotic Cluster Synchronization. arXiv:2102.02272.

[5] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[6] T. Carletti, D. Fanelli, and S. Nicoletti. Dynamical systems on hypergraphs. arXiv:2006.01243.

[7] M.J. Casey, P.S. Stumpf, and B.D. MacArthur. Theory of cell fate. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 12(2):e1471, 2020.

[8] G.F. De Arruda, M. Tizzani, and Y. Moreno. Phase transitions and stability of dynamical processes on hypergraphs. arXiv:2005.10891.

[9] M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, et al. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14(6):671–683, 2013.

[10] E. Estrada and J.A. Rodríguez-Velázquez. Subgraph centrality and clustering in complex hyper-networks. *Physica A: Statistical Mechanics and its Applications*, 364:581–594, 2006.

[11] C. Flamm, B.M.R. Stadler, and P.F. Stadler. Chapter 13—generalized topologies: Hypergraphs, chemical reactions, and biological evolution. In S.C. Basak, G. Restrepo, and J.L. Villaveces, editors, *Advances in Mathematical Chemistry and Applications*, pages 300–328. Bentham Science Publishers, 2015.

[12] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[13] C.-H. Huang, J.J.P. Tsai, N. Kurubanjerdjit, and K.-L. Ng. Computational analysis of molecular networks using spectral graph theory, complexity measures and information theory. *bioRxiv*, page 536318, 2019.

[14] B. Hwang, J.H. Lee, and D. Bang. Single-cell rna sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, 50(8):1–14, 2018.

[15] J. Jost and R. Mulas. Normalized Laplace Operators for Hypergraphs with Real Coefficients. *Journal of Complex Networks*, 9(1):cnab009, 2021.

[16] S. Klamt, U.U. Haus, and F. Theis. Hypergraphs and cellular networks. *PLoS Comp. Biol.*, 5(5):e1000385, 2009.

[17] J. Lause, P. Berens, and D. Kobak. Analytic pearson residuals for normalization of single-cell rna-seq umi data. *bioRxiv*, 2020.

[18] A. Lesne. Complex networks: from graph theory to biology. *Letters in Mathematical Physics*, 78(3):235–262, 2006.

[19] B.D. MacArthur and R.J. Sánchez-García. Spectral characteristics of network redundancy. *Physical Review E*, 80(2), 2009.

[20] B.D. MacArthur, R.J. Sánchez-García, and J.W. Anderson. Symmetry in complex networks. *Discrete Applied Mathematics*, 156(18):3525–3531, 2008.

[21] O. Mason and M. Verwoerd. Graph theory and networks in biology. *IET systems biology*, 1(2):89–119, 2007.

[22] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426.

[23] R. Mulas, C. Kuehn, and J. Jost. Coupled dynamics on hypergraphs: Master stability of steady states and synchronization. *Physical Review E*, 101:062313, 2020.

[24] R. Mulas, R.J. Sánchez-García, and B.D. MacArthur. Geometry and symmetry in biochemical reaction systems. arXiv:2010.01049.

[25] A.D. Perkins and M.A. Langston. Threshold selection in gene co-expression networks using spectral graph theory techniques. In *BMC bioinformatics*, volume 10, pages 1–11. BioMed Central, 2009.

[26] A. Ritz, A.N. Tegge, H. Kim, C.L. Poirel, and T.M. Murali. Signaling hypergraphs. *Trends in biotechnology*, 32(7):356–362, 2014.

[27] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

[28] A. Sarkar and M. Stephens. Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis. *Nature Genetics*, pages 1–8, 2021.

[29] M. Schwob, J. Zhan, and A. Dempsey. Modeling cell communication with time-dependent signaling hypergraphs. *IEEE/ACM transactions on computational biology and bioinformatics*, 2019.

[30] Rosanna CG Smith and Ben D MacArthur. Information-theoretic approaches to understanding stem cell variability. *Current Stem Cell Reports*, 3(3):225–231, 2017.

[31] R.J. Sánchez-García. Exploiting symmetry in network analysis. *Commun. Phys.*, 3(87), 2020.

[32] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W.M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of single-cell data. *Cell*, 2019.

[33] P.S. Stumpf, H. Imanishi, Y. Kunisaki, Y. Semba, T. Noble, R. Smith, M. Rose-Zerilli, J. West, R. Oreffo, K. Farrahi, et al. Transfer learning efficiently maps bone marrow cell types from mouse to human using single-cell rna sequencing. *Communications Biology*, 2020.

[34] V. Svensson. Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, 2020.

[35] V. Svensson, K.N. Natarajan, L.-H. Ly, R.J. Miragaia, C. Labalette, I.C. Macaulay, A. Cvejic, and S.A. Teichmann. Power analysis of single-cell rna-sequencing experiments. *Nature methods*, 14(4):381, 2017.

[36] L. Tian, X. Dong, S. Freytag, K.-A. Lê Cao, S. Su, A. JalalAbadi, D. Amann-Zalcenstein, T.S. Weber, A. Seidi, J. S. Jabbari, et al. Benchmarking single cell rna-sequencing analysis pipelines using mixture control experiments. *Nature methods*, 16(6):479–487, 2019.

[37] F.W. Townes, S.C. Hicks, M.J. Aryee, and R.A. Irizarry. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome biology*, 20(1):1–16, 2019.

[38] V.A. Traag, L. Waltman, and N.J. Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

[39] C. Trapnell. Defining cell types and states with single-cell genomics. *Genome research*, 25(10):1491–1498, 2015.

[40] F.A. Wolf, F.K. Hamey, M. Plass, J. Solana, J.S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, and F.J. Theis. Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology*, 20(1):1–9, 2019.

[41] L. Zappia, B. Phipson, and A. Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):1–15, 2017.