LIFE: International Journal of Health and Life-Sciences
ISSN 2454-5872

Donovan et al., 2024

Volume 9, pp. 21-33

Received: 05th June 2024

Revised: 10th June 2024, 11th June 2024

Accepted: 10th June 2024

Date of Publication: 15th June 2024

DOI- https://dx.doi.org/10.20319/lijhls.2024.9.2133

This paper can be cited as: Barton, S., Coster, A., Donovan D., Lefevre, J (2024). Applying Hypergraphs to Studies in Quantitative Biology. LIFE: International Journal of Health and Life-Sciences, 09, 21-33.

APPLYING HYPERGRAPHS TO STUDIES IN QUANTITATIVE BIOLOGY

Samuel Barton
ARC Centre of Excellence, Plant Success in Nature and Agriculture, School of Mathematics and Physics, University of Queensland, Brisbane, 4072, Australia
s.barton@uq.net.au

Adelle Coster
School of Mathematics & Statistics, The University of New South Wales, NSW 2052, Australia,
a.coster@unsw.edu.au

Diane Donovan
ARC Centre of Excellence, Plant Success in Nature and Agriculture, School of Mathematics and Physics, University of Queensland, Brisbane, 4072, Australia
dmd@maths.uq.edu.au

James Lefevre
ARC Centre of Excellence, Plant Success in Nature and Agriculture, School of Mathematics and Physics, University of Queensland, Brisbane, 4072, Australia
j.lefevre@uq.edu.au

Abstract

The objective of this research is to demonstrate hypergraph versatility and applicability for modeling diverse biological systems. The inherent structure of hypergraphs allows for encoding

of higher-order feature interactions, providing a flexible framework for efficient models that can enhance our understanding of physical phenomena and one that can be generalized across various datasets. By adopting innovative methods including centrality measure and populations of models rather than singular instances, biases and overfitting tendencies are mitigated, again presenting promise for application across a broad spectrum of biological systems. Furthermore, emphasis is placed on the significance of probabilistic distribution analysis in elucidating threshold selection and feature relevance while maintaining high levels of accuracy. Our results demonstrate the advantages of hypergraph models on two different datasets; with the first on gene expression and the identification of outlier genes and the second on classifying starch grains. There is significant scope in the application of the hypergraph to a wider class of biological systems, with the potential to improve understanding of the biological processes.

Keywords:

Hypergraph, Hypergraph Model, Hypergraph Classifier, Graph

1. Background and Research Objectives

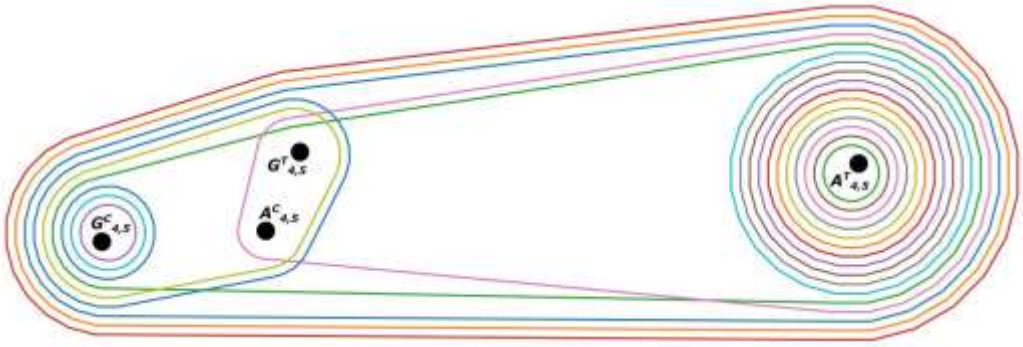
Hypergraphs have been applied across a broad range of studies, including bioinformatics (Di et. al., 2021), (Barton et. al., 2023) social networks (Li et. al., 2013) pattern recognition (Zhou et. al., 2006) Hypergraphs move beyond the pairwise comparisons given by graphs or networks and are designed to succinctly model higher-order relationships in complex, irregular data structures. The objective of the current article is to demonstrate hypergraph versatility and applicability in two distinct biological settings. The first involves differential gene expression data gathered to study a gravitropism trait in the Australian coastal plant Senecio lautus. The second involves plant microfossil data gathered from starch grains for species classification.

Gravitropism is the specific response of plants to gravity, resulting in upright growth even when the orientation of the plant is perturbed. It is known that in Senecio lautus the gravitropic response varies between populations adapted to different environments. In sand dunes, Senecio lautus responds strongly to gravity, with plants growing tall and erect, while on headlands the response is weak or absent, with low, prostrate growth. An experiment was designed to study gravitropism, recording differential gene expression over a short time series (5 points) in the presence and absence of experimental treatment (Broad et. al., 2023).

Using this data we may construct the hypergraph shown in Figure, with node set $\{G^C_{4,5}, G^T_{4,5}, A^C_{4,5}, A^T_{4,5}\}$. These nodes represent data from time points 4 to 5 in the experimental

categories gravitropic (*G*) and agravitropic (*A*) and under the presence (*T*) and absence (*C*) of treatment, respectively. The closed curves represent hyperedges aligned with contrasts (gene expression patterns) seen across different experimental categories. In this example, a node is incident with a hyperedge if the given experimental category shows the required response with respect to the contrast of interest: hyperedges $\{G_{4,5}^{C}\}$, $\{A_{4,5}^{T}\}$, $\{G_{4,5}^{C}, A_{4,5}^{C}, G_{4,5}^{T}\}$, $\{A_{4,5}^{C}, G_{4,5}^{T}, A_{4,5}^{T}\}$, $\{G_{4,5}^{C}, A_{4,5}^{C}, G_{4,5}^{T}, A_{4,5}^{T}\}$ of multiplicity 3, 11, 2, 1, 4, respectively.

Figure 1: *An Example Hypergraph on 4 Nodes, with 21 Hyperedges Incident with 1, 3 Or 4 Nodes*



(Source: Authors' Illustration).

In the above example, the hyperedges can be expanded to cliques giving a dense complex network that records pairwise interactions but with much of the detail obscured. Summary statistics, such as node degree, shortest path lengths and node centrality, can help but we will show that much is obtained by applying these statistics in the hypergraph setting.

In a second application, we construct a hypergraph model to compile a plant microfossil reference library for the classification of microfossils of unknown species. Here the data set records measurements taken from starch grains that grow between cells, (Coster & Field, 2015). Generally, immature grains are small and spherical, but over time crowding results in constrained growth and trait homogeneity is reduced across species. Shape metrics and supplementary Fourier signatures can be used to capture this differentiation and to formulate a hypergraph classifier that encodes higher-order relationships. One challenge that arises is the lack of encoding for grain maturity.

2. Methods

We will briefly describe two distinct methods applied here and refer the reader to (Barton et. al., 2023) and (Barton et. al., 2024) for full details. Formally a *hypergraph*, $H = (V, E)$ is given by a node set $V = \{v_1, v_2, \dots, v_n\}$ and hyperedge set $E = \{e_1, e_2, \dots, e_n\}$, where hyperedges are subsets of the node set V (Berge, 1984). A graph or network is a hypergraph where all hyperedges contain precisely 1 or 2 nodes.

The first step in constructing a hypergraph, from a given data set, is to determine a list of conditions of interest that will act as nodes. Then if a condition (node *v*) is satisfied by a given data unit (hyperedge *e*), node *v* is included in hyperedge *e*. Recording "Yes"" as 1 and "No" as 0 we obtain a Boolean sequence profiling the data unit (hyperedge) with respect to the list of conditions (node set V). An *incident matrix*, $B = [b_{i,j}]$, has the Boolean sequences as rows, one per node, where entry $b_{i,j} = 1$ if node v_i is in hyperedge e_j, otherwise $b_{i,j} = 0$. Summary statistics are now accessible: row sums give the *degree distribution* (number of hyperedges incident with each node); column sums give the *hyperedge size distribution* (number of nodes in each hyperedge).

These distributions are proxies for the number of experimental units that test positive for the conditions of interest and provide information on the topology of the hypergraph. For instance, scale-free hypergraphs have a small central hub of nodes of high degree exhibiting interactions with a large number of low-degree nodes. The *s*-intersection line graph identifies experimental units that concurrently satisfy *s* conditions of interest. It also gives access to *centrality measures*, e.g. *closeness, betweenness and eigencentrality*: high closeness indicates a hyperedge that is similar to other hyperedges (with respect to graph distance); higher than trend betweenness indicates a hyperedge that is similar in nature to at least two distinct classes of hyperedges; higher eigencentrality tends to indicate an "influential" hyperedge which has common elements with other "influential" hyperedges.

2.1 A Hypergraph Model for Studying Differential Gene Expression

In this study, we derived a list of hypergraph test conditions from experimental contrasts with respect to a threshold in differences for normalized (log transform) gene expression levels. The conditions were designed to test for differences in gene expression given a threshold. For example, "*Given a gene, is there a two-fold difference in adjusted expression levels as measured*

in the agravitropic control experiment (A^C) and the gravitropic treatment experiment (G^T) at time point 5." In addition, given the large gene set with low rates of true expression difference, family-wide false discovery rates were controlled by calibrating differences in expression levels against a significance criteria (adjusted *p*-value of less than 0.05). Incorporating dispersion information from the full set of genes allowed for optimal estimates of variance, while still tailoring the *p*-value criteria to individual conditions of interest and single genes. Summarizing, for a given gene and an experimental contrast, fold change and adjusted *p*-value scores were used to determine incidence in hyperedges.

1. Input gene expression data and normalize with a log transform. Determine a list of hypergraph conditions, $C_1, \dots, C_m$, aligned with experimental contrasts and determine adjusted *p*-values.
2. Construct hypergraph H with nodes $C_1, \dots, C_m$ (conditions) and hyperedge g_j containing node C_i if gene g_j satisfies condition C_i.
3. Determine the degree and hyperedge size distribution from the incidence matrix and where possible, characterize the topology of the hypergraph, e.g. scale free.
4. Construct the *s*-intersection line graph and determine closeness, betweenness and eigencentrality.
5. Use the above to identify genes that show functional variation across experimental contrasts.

2.2 A Hypergraph Algorithm to Classify Discrete Data Units

In this study, we derived a hypergraph model to classify plant microfossils (starch grains) of unknown species by referencing a known library of microfossils. As mentioned earlier, environmental challenges are present: microfossils can range in maturity from early to fully developed, causing overlapping in measurements and creating significant variance, see (Coster & Field, 2015). The compounding of these issues can restrict the performance of classifiers. To meet these challenges, we have constructed a classifier that differentiates species by referencing main effects and multi-way interactions. The performance will be compared against a random forest classifier, a robust and accurate tool, (Breiman, 2001) (Díaz-Uriarte & Alvarez de Andrés, 2006). Fuzzy boundaries are introduced to account for data variance and are achieved through a population of hypergraph models that allow for variance across interval lengths and central values for classes. Sensitivity analysis is used to identify significant and remove irrelevant or redundant features, as per the literature (Dash & Liu, 1997). Incorporating this information into confidence

thresholds for the classification of unidentified starch grains can reduce false positives and negatives. The overall approach is summarized below with full details given in (Barton et. al., 2024).

1. Input microfossil feature measurement data, e.g. shape metrics, and Fourier signature, and normalize transforming units using *p*-values and *z*-scores, with mean zero and standard deviation one. Label units by a standardized vector $\hat{v}_i = (\hat{v}_{i1}, \dots, \hat{v}_{im})$.
2. Select population of model parameters, $l \in U(0.2,1.5)$ and $\alpha \in U(-0.5,0.5)$ from uniform distributions, and discretize the components of $\hat{v}_i$to $v\_i = (\left\lceil \frac{\hat{v}_{i1}-\alpha}{l} \right\rceil, \cdots, \left\lceil \frac{\hat{v}_{im}-\alpha}{l} \right\rceil)$.
3. Construct the hypergraph *H*, where the microfossils form the node set, and hyperedges are indexed by pairs (j,t), where *j* is a feature and *t* is the associated discretized value. Then node *i* is incident with hyperedge (j,t) if and only if $t = v_{ij}$.
4. Repeat Steps 1, 2 and 3 generating probability distribution across features, their discretized classes and the population of models.
5. Set a threshold, and only predict a species that occurs in the set percentage of class forecasts. That is, in the final selection process, the predicted class must be selected by at least *k*% of models, otherwise no prediction is made.
6. Based on training data, use Steps 1 to 4 to create a reference library. Given an unknown microfossil $\hat{v}$ use the probability distributions stored in the reference library to determine the most likely species, say *s*. If the prediction is above the threshold, say $P(p(\hat{v} = s) > k$, predict species *s* for grain $\hat{v}$.

3. Results

In both studies, the goal is to use hypergraphs to interrogate data and identify both general trends as well as outliers and other units of data that warrant further investigation. Understanding the associated abstract concepts is enhanced by contextualizing and interpreting them within the biological setting.

First, gravitropic processes were studied through expression data gathered from 269,210 candidate gene sequences (referred to as genes). The experiment involved rotating individual seedlings by 90° and determining expression levels at 5 time points. A more complete analysis of the study is given in (Barton et. al., 2023) Here we focus on a representative example,

demonstrating the strengths of centrality measures in hypergraphs as interrogation tools for a multi-dimensional dataset.

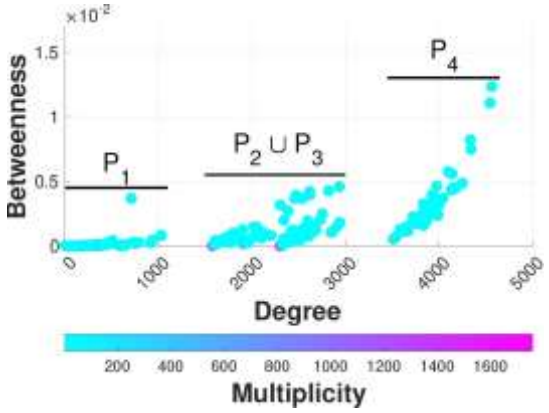
As discussed in Methods, a list of conditions was nominated and tested to identify links between differential gene expression levels and the gravitropic response. This information was captured in hypergraph models and the associated *s*-intersection line graphs, with betweenness and eigencentrality scores emerging as key indicators of potentially significant genes.

The data points in Figures 2 and 3 correspond to hyperedges or equivalently genes, with a natural 3-way partitioning of the gene pool emerging. Further, the eigencentrality scores (Figure 3) provide a much clearer delineation into four distinct gene classes $P_1,\ P_2,\ P_3$ and P_4, allowing us to distinguish genes based on key conditions. The betweenness centrality scores (Figure 2) allow us to identify genes of interest within each class, highlighting possible gene associations. The full biological significance of results is discussed in (Barton et. al., 2023) but for now, let us consider the two outlier genes with high centrality and degree (number of contrast/conditions satisfied) seen in Figure 2. Investigations suggest that these are the ubiquitin gene and a polyubiquitin gene with functions related to vegetative growth, auxin signaling and ethylene production.

Interestingly, betweenness does not correlate strongly with degree, and it does not differentiate the four classes. However, it does identify outlier genes, for instance, in class P_1 (with a small degree). This node represents the gene that was annotated to the *A.thaliana* shaggy-like kinase group 2 (AtSK2-2), also known as brassinosteroid insensitive 2-like. This gene shows strong differentiation across three distinct experimental contrasts/conditions as well as some commonality with a large number of other genes. Biologically, it is known that this gene is associated with the brassinosteroid-mediated signaling pathways and related to light-regulated

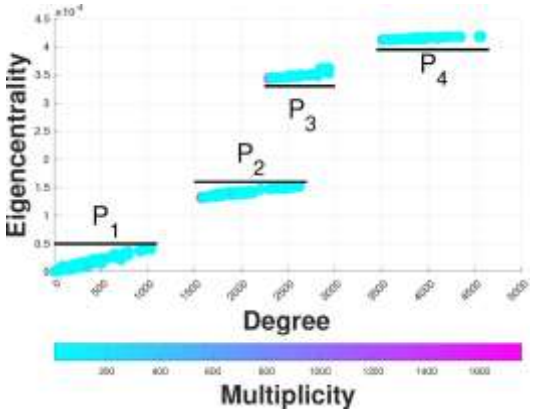
hypocotyl elongation and multiple stress responses including responses to drought, see (Barton et. al., 2023).

Figure 2: *The Betweenness and Degree Scores for the Hypergraph Model*



(Source: Authors' Illustration)

Figure 3: *The Eigencentrality and Degree Scores for the Hypergraph Model*

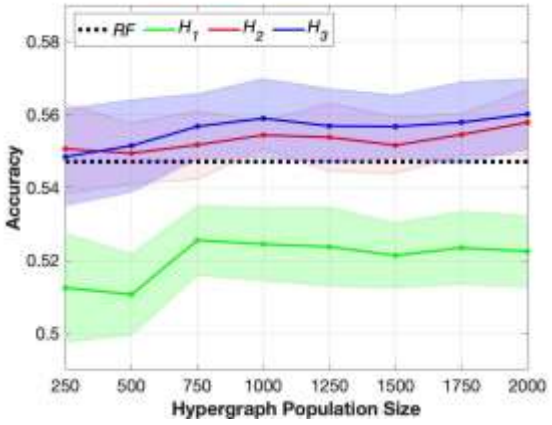


(Source: Authors' Illustration)

In the second study, the hypergraph classifier was tested on a relatively small dataset of 944 starch grains from seven distinct species, with each grain calibrated against 16 morphological features: the 5 shape metrics: length, area, perimeter, circularity, hilum position, and 11 Fourier coefficients. The small number of data units and the environmental challenges did impact performance; however, results show that the proposed hypergraph classifier performs well when compared to a random forest classifier.

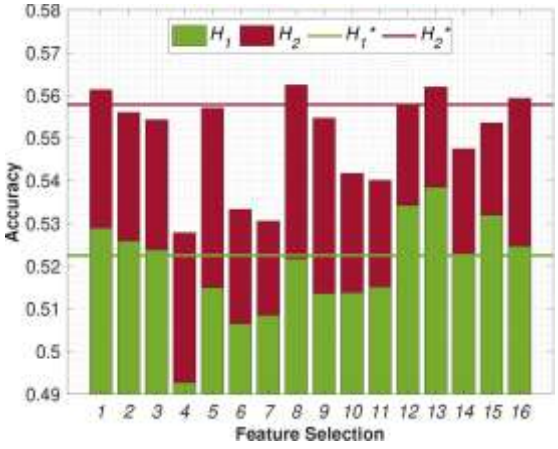
Figure 4 summarizes accuracy and standard error (shaded regions) for three hypergraph models (main effects only H_1 including 2-way interactions H_2, including 3-way interactions H_3) and a random forest classifier (*RF*). The hypergraph results are mapped against increases in the size of the hypergraph population. Overall H_3 achieves an accuracy score of 0.5601 and a standard error of 0.0097, while the random forest classifier achieves a slightly lower accuracy score of 0.5471 with a standard error of 0.0092. For this type of study, and given the specific challenges mentioned earlier, this is quite a respectable result especially when benchmarked against the random forest algorithm.

Figure 4: *Accuracy Scores across Different Hypergraph Population Sizes, with Shaded Regions Indicating a Standard Error*



(Source: Authors' Illustration)

Figure 5: *Adjusted and Original Accuracy Scores for the H_1 and H_2 Algorithms. Feature Selection Refers to Which Feature is removed, with the Label Coming from (Barton Et. Al., 2024)*



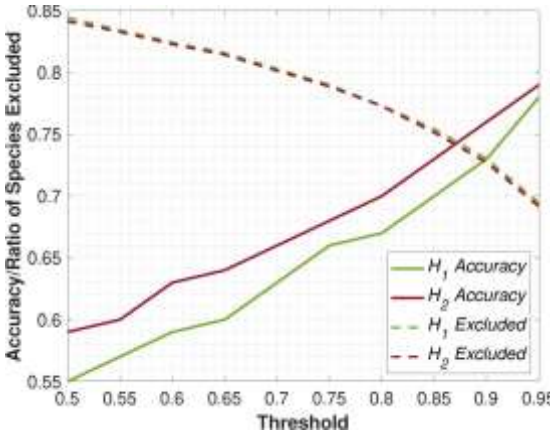
(Source: Authors' Illustration)

Leave one out sensitivity analysis of features is reported in Figure 5 (green main effects and red 2-way interaction models) with horizontal lines indicating overall accuracy. Little is gained by removing features from the H_2 model, however, the removal of the *Circularity* data (Feature 4 in Figure 5) markedly reduces the performance in both models, suggesting that circularity is a good classifier despite the uncertainty in the measurement of immature grains. Not surprisingly, removing the early Fourier coefficients also impacts on accuracy.

To further address the uncertainty introduced by immature spherical grains, we implemented a decision threshold that leaves some starch grains unclassified. Indeed, Figure 6 indicates that performance is improved by suppressing some information and reducing homogeneity in the training data; the H_2-algorithm continues to outperform the H_1-algorithm although the difference drops moderately for higher thresholds. Increasing the threshold also increases the number of unclassified grains. Figure 7 illustrates this trade-off. For thresholds up to 40%, most grains in this dataset can still be classified. When the threshold is increased to 70%, the fraction of classified grains reduces to approximately 0.4 and drops off with almost all grains unclassified for close to 100% accuracy.

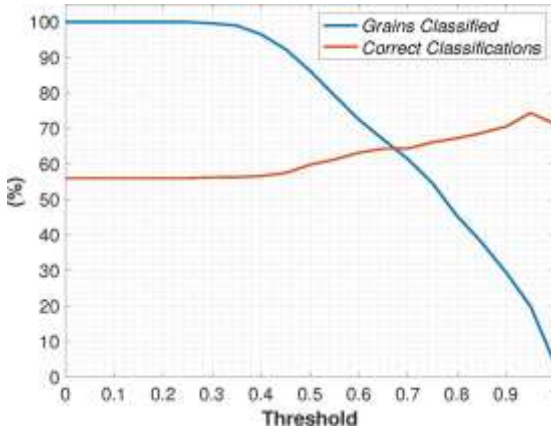
This approach minimizes the rate of both false positives and false negatives and maximizes the rate of both true positives and true negatives (Table 3.2). The hypergraph model also allows us to rule out unlikely species for an unclassified grain. This is valuable information

Figure 6: *The accuracy scores (solid lines) and the ratio of species excluded (dashed lines) for different threshold values when using the class prediction technique for the H_1 and H_2 algorithms*



(Source: Authors' Illustration)

Figure 7: *The accuracy of the H_3 algorithm (red) and the percentage of units classified (blue) across decision threshold values*



(Source: Authors' Illustration)

for immature spherical starch grains, where the inability to identify the specific species with high accuracy is often unavoidable.

Figure 6 displays results on the average number of species eliminated across different threshold values, showing that even at the highest threshold value considered we can eliminate over half of species from predictions on average. The threshold can be chosen to achieve a target accuracy for

ruling out unlikely species (Table 3.1 A threshold value of around 0.5 will give 90% accuracy with respect to the number of species ruled out and around 0.65 gives 95% accuracy.

Table 3.1 *The minimum threshold required to achieve given accuracy for each algorithm and the number the species ruled out*

Accuracy	Algorithm	Threshold	Number of Species Ruled Out
90%	H_1	0.5	3.8313
90%	H_2	0.52	4.14
95%	H_1	0.65	2.8038
95%	H_2	0.66	3.3

(Source: Authors' Calculations)

Full details on the performance of the classifiers including confusion matrices are given in (Barton et. al., 2024) Here Table 3.2 shows the performance of three classifiers across the species: $H_3(75)$ with a threshold of 75%, H_3 and without a threshold, and a random forest (*RF*). True positive and false negative rates are defined as proportions of the specified class, while false positive and true negative rates are proportions of all data excluding the class. For each of the four rates, $H_3(75)$ outperforms the other classifiers across most species.

Table 3.2 *True and false positive and negative rates for each classifier RF,* $\boldsymbol{H_3}$ *and* $\boldsymbol{H_3(75)}$. *The best result in each case is displayed in bold.*

Species/Algorithm	$\boldsymbol{RF: H_3: H_3(75)}$			
	False Negative	**False Positive**	**True Positive**	**True Negative**
Dysphania kalpari (DK)	16.1:7.3:**5**	**6.4:**8.9:11.6	83.9:92.7:**95**	**93.6**:91.1:88.4
Acacia aneura (AA)	45:21:**7.6**	**6:**7.7:8.2	55:79:**92.4**	**94:**92.3:91.8
Acacia victoriae (AV)	62.6:**53.7**:55.2	7.9:8.9:**5.5**	37.4:**46.3**:44.8	92.1:91.1:**94.5**
Brachiaria miliiformis BM)	48.5:41:**23.3**	9.7:10.5:**7.7**	51.5:59:**76.7**	90.3:89.5:**92.3**
Eragrostis eriopoda (EE)	**57.7:**71.4:77.6	8.4:5.1:**1.3**	42.3:28.6:22.4	91.6:94.9:98.7
Yakirra australiensis (YA)	44.4:52.8:**40.3**	6.6:5.1:**0.9**	55.6:47.2:**59.7**	93.4:94.9:**99.1**
Brachychiton populneus (BP)	43.7:48.5:**41.6**	8.4:5:**4.8**	56.3:51.5:**58.4**	91.6:95:**95.2**

(Source: Authors' Calculations)

4. Conclusions

This article explores the application of hypergraph models in two distinct biological settings, demonstrating that hypergraphs have the potential to create flexible models that can be generalized to other datasets. By taking the novel approach of working over populations of models, rather than a single model, there is a potential to reduce biases and overfitting. Both these facts suggest the need for future studies that apply hypergraph to a wider class of biological systems thus harnessing their potential to improve understanding of the physical processes.

In addition, greater emphasis needs to be placed on developing an understanding of the setting of thresholds and the selection of features through probability distribution analysis.

In conclusion, it has been demonstrated that hypergraph models provide a foundation for exploring multi-way interactions within datasets, while populations of hypergraphs provide a robust framework for the classification of unseen data units for complex datasets.

REFERENCES

Barton, S., Coster, A., Donovan, D. and Lefevre, J. (2024). A classification model based on a population of hypergraphs. arXiv, 2405.15063. URL https://arxiv.org/abs/2405.15063

Barton, S., Broad, Z., Ortiz-Barrientos, D., Donovan, D. and Lefevre, J.(2023). Hypergraphs and centrality measures identifying key features in gene expression data. Mathematical Biosciences, 366, 109089. DOI https://doi.org/10.1016/j.mbs.2023.109089

Breiman, L. (2001) Random forests. Machine Learning, 45, 5-32. DOI https://doi.org/10.1023/A:1010933404324

Berge, C. (1984) Hypergraphs: combinatorics of finite sets. Elsevier 45. URL https://books.google.com.au/books/about/Hypergraphs.html?id=jEyfse-EKf8C&redir_esc=y

Broad, Z., Lefevre, J., Wilkinson, M., Barton, S., Barbier, F., Jung, H., Donovan, D. and Ortiz-Barrientos, D. (2023). Gene expression divergence during adaptation to contrasting environments. arXiv. DOI 10.22541/au.169823548.87378722/v1

Coster, A.C.F. and Field, J.H. (2015) What starch grain is that? A geometric morphometric approach to determining plant species origin. Journal of Archaeological Science, 58, 9-25. DOI https://doi.org/10.1016/j.jas.2015.03.014

Dash, M. and Liu, H. (1997) Feature selection for classification. Intelligent data analysis, 1(1-4), 131-156. DOI https://doi.org/10.1016/S1088-467X(97)00008-5

Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006) Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7, 1-13. DOI https://doi.org/10.1186/1471-2105-7-3

Di D., Shi F., Yan F., Xia L., Mo Z., Ding Z., Shan F., Song B., Li S., Wei Y., Shao Y., Han M., Gao Y., Sui H., Gao Y. and Shen D. (2021). Hypergraph learning for identification of COVID-19 with CT imaging, Medical Image Analysis, 68, 101910. DOI 10.1016/j.media.2020.101910

Li, D., Xu, Z., Li, S. and Sun, X. (2013). Link prediction in social networks based on hypergraph. Proceedings of the 22nd international conference on world wide web, 41-42. URL https://api.semanticscholar.org/CorpusID:15302229

Zhou, D., Huang, J. and Schölkopf, B. (2006) Learning with hypergraphs: Clustering, classification, and embedding. Advances in neural information processing systems, 19. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/dff8e9c2ac33381546d96deea9922999-Paper.pdf