Py/ExplorReg: Exploration of Transcriptome for Potential Regulon Structure Detection

1st Patrícia Janigová
Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication, Brno University of Technology
Brno, Czech Republic
231027@vut.cz

2nd Wolfram Weckwerth
Molecular Systems Biology (MOSYS), University of Vienna
Vienna, Austria
Vienna Metabolomics Center (VIME), University of Vienna
Vienna, Austria
wolfram.weckwerth@univie.ac.at

3rd Jana Schwarzerová
Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication, Brno University of Technology
Brno, Czech Republic
Department of Molecular and Clinical Pathology and Medical Genetics, University Hospital Ostrava
Ostrava, Czech Republic
Molecular System Biology (MOSYS), University of Vienna
Vienna, Austria
Jana.Schwarzerova@vut.cz

***Abstract*—This study introduces Py/ExplorReg, a tool designed for exploring transcriptomic landscapes and identifying potential regulon structures, with a particular focus on its utility in the context of *Arabidopsis thaliana* research. Developed using Python, Py/ExplorReg demonstrates its effectiveness in identifying potential regulons through a pioneering approach rooted in the analysis of gene expression data. Leveraging publicly available datasets from the EMBL-EBI project PRJNA779072, this case study highlights its adaptability to *Arabidopsis thaliana* datasets.**

***Keywords* — *Arabidopsis thaliana*, Gene expression analysis, Regulon, Transcription units**

I. Introduction

Understanding the intricate regulatory mechanisms governing gene expression is fundamental to deciphering the complexities of biological systems [1]. Regulatory networks, composed of interconnected genes and their regulatory elements, orchestrate precise control over cellular processes in response to internal and external cues. However, accurately detecting the regulon across the organism is a pivotal step for the precise derivation of regulatory networks [2].

In the past, comparable tools were predominantly utilized for identifying operons in bacterial systems [3], [4]. However, in this case study, our aim is to repurpose a similar methodology by recalibrating its parameters. Our objective is to apply this adapted approach to analyze regulatory structures within eukaryotic organisms, including plants. This endeavor aims to pave the way for novel avenues within ecological engineering.

Py/ExplorReg represents a novel approach to interrogating gene expression data, particularly focusing on its applicability within the context of *Arabidopsis thaliana* research. By harnessing methodologies such as correlation coefficients in gene expression analysis [5], Py/ExplorReg pioneers a new frontier in the investigation of gene regulatory networks from an in-silico perspective.

Py/ExplorReg, implemented in Python, is capable of analyzing transcriptomic data to elucidate regulatory relationships within *Arabidopsis thaliana* and beyond.

II. Dataset

This study, centered on *Arabidopsis thaliana*, employed the Columbia (Col-0) ecotype and specific mutants sourced from the Arabidopsis Biological Resource Center. All specimens were cultivated under controlled conditions to ensure data reliability.

By leveraging publicly available datasets, such as those from the EMBL-EBI project PRJNA779072 [6], Py/ExplorReg demonstrates its adaptability and efficacy in uncovering regulon structures. This introduction provides an overview of Py/ExplorReg's significance in advancing our understanding of gene regulation and highlights its potential contributions to molecular biology research.

Various genomic and epigenomic techniques, including Whole-genome bisulfite sequencing (BS-seq), Transcription Start Site sequencing (TSS-seq), and Chromatin Immunoprecipitation sequencing (ChIP-seq), are available to unveil the dynamics of gene expression regulation and DNA methylation patterns [6]. Subsequently, our analysis primarily focused on gene expression data derived from ChIP-Seq and RNA-Seq (ID GSE188493) [6] as the used dataset for this study.

Different phenotypic expressions in various *Arabidopsis thaliana* mutants may arise from different regulons influencing plant traits. This underscores the significance of regulons in shaping plant morphology and adaptive characteristics. Fig. 1 highlights morphological differences in *Arabidopsis thaliana* resulting from distinct regulon structures. By comparing variants such as hira-1, asf1a-1, and asf1b-1 with the Col-0 control, distinct phenotypic expressions were observed [6].

These differences are crucial for understanding how regulons modulate specific plant features and adaptations.

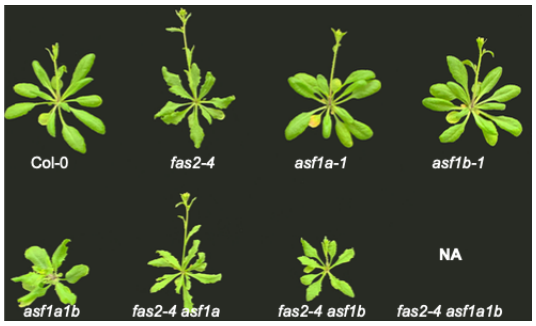

Fig. 1. Morphology of Col-0, hira-1, asf1a-1, asf1b-1, asf1a1b, hira-1 asf1a-1, hira-1 asf1b-1, and hira-1 asf1a1b. The figure was taken over from [6].

III. Methods

The identification and analysis of regulon structures within the transcriptome of *Arabidopsis thaliana* constitute the primary focus of Py/ExplorReg. The potential regulons, defined as functional clusters of genes regulated under a shared promoter region and controlled by specific transcription factors, play a pivotal role in understanding gene regulatory networks [7].

Regulatory networks are essential for the organism's response to various environmental cues and developmental processes. Unlike operons, which are characterized by their continuous genomic location, regulons are not confined to a specific definitive order and may be distributed in different regions throughout the genome, Fig. 2 [8].

This dispersal allows for a coordinated response to external stimuli, reflecting the complexity of gene regulation in eukaryotic organisms [2]. This regulatory feature permits genes within a regulon to be coordinately activated or repressed, even when dispersed throughout the genome.

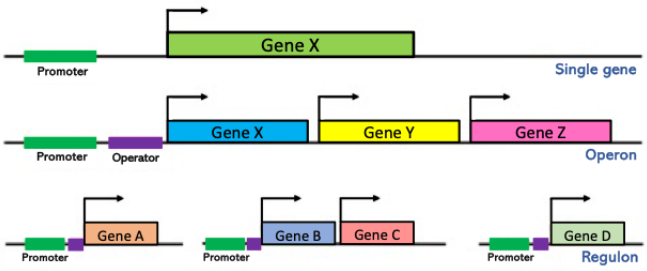

Fig. 2. Illustration of three types of regulatory structures: Single Gene (top), Operon (middle), and Regulon (bottom). In the Single Gene structure, a gene is controlled by its own promoter. The Operon depicted includes genes X, Y, and Z, which share a common promoter and operator, indicative of co-regulated gene expression. The Regulon consists of genes A, B, C, and D, each with individual promoters but collectively regulated, exemplifying the concept of a regulon.

IV. Py/ExplorReg: Exploration of Transcriptome for Potential Regulon Structure Detection

Py/ExplorReg presents a computational case study developed to analyze gene expression data with the objective of identifying potential regulons — groups of co-expressed genes confirmed post-hoc by literature and regulon databases such as [9]. Utilizing Python packages such as Pandas [10], NumPy [11], SciPy [12], and Scikit-learn [13], Py/ExplorReg navigates the complex landscape of gene regulation by combining Pearson's correlation coefficient [5] and mutual information [14], followed by visualization using a heatmap and cluster dendrogram.

The correlation coefficient calculates the linear dependency between gene expressions, identifying genes that co-vary under various conditions. This measure is crucial for identifying genes with synchronized expression changes, suggesting they might belong to the same regulon. Meanwhile, mutual information extends this analysis to capture non-linear relationships, revealing deeper layers of gene regulation and providing insights into more complex gene interactions that correlation alone might miss.

Py/ExplorReg combines these two metrics to create an intersection matrix that normalizes and integrates the calculated correlation coefficients and mutual information scores. By applying a threshold (set at 0.8 for this study), the tool identifies the most correlated or dependent gene pairs, suggesting a high likelihood of belonging to the same potential regulon. The final step involves visualizing these relationships through a binary image representation, where values above the threshold are marked, highlighting the identified potential regulons.

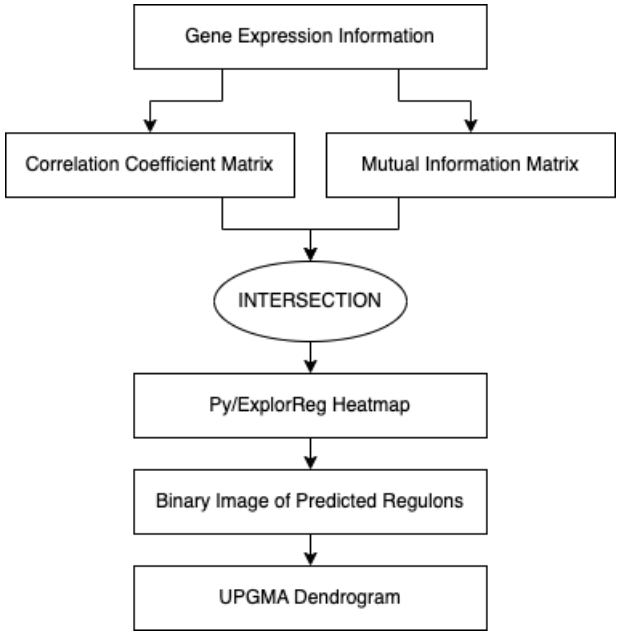

Fig. 3. Flowchart illustrating the workflow of Py/ExplorReg, such as a computational analysis for transcriptome analysis and potential regulon identification.

V. Results and Discussion

Computational exploration using Py/ExplorReg identified a strong correlation between specific genes in *Arabidopsis thaliana*, revealing key insights into their genetic coordination. Specifically, the algorithm revealed that the strongest linkage was notably observed between AT4G28520 and AT4G27150. This correlation is slightly less significant when involving AT4G25140, indicating a nuanced hierarchy of interaction within the potential regulon. These gene relationships were illustrated in a detailed heatmap, Fig. 4, providing a compelling graphical representation of their interconnectedness.

The heatmap visualizes the intersection of correlation coefficients and mutual information, ranging from 0 to 1. Values close to 1 indicate a strong linear correlation or dependency, suggesting significant regulation between genes. Importantly, the heatmap's diagonal always shows values of 1, as it represents each gene's correlation with itself, demonstrating the heatmap's utility in depicting complex gene interactions within biological systems. This concise visualization facilitates understanding of the intricate co-expressed networks influencing gene expression in *Arabidopsis thaliana*, serving as a preliminary step before exploring regulatory gene networks, which can be confirmed based on literature information.

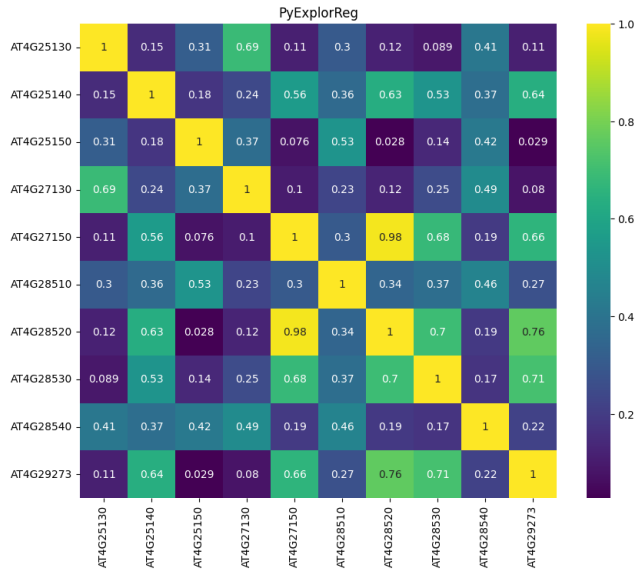

Fig. 4. Visualization of correlation coefficients and mutual information scores in gene expression analysis using Py/ExplorReg – a heatmap representation.

After the heatmap visualization highlighted the intricate dependencies among genes AT4G28520, AT4G27150, and AT4G25140, this complex interaction is further discernible in the dendrogram, Fig. 5, generated through the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm [15]. UPGMA is a hierarchical clustering method that builds a phylogenetic tree by sequentially merging pairs of clusters based on similarity, determined by the average distance between elements in each cluster. Within the dendrogram, these genes form a distinct cluster, illustrating their close regulatory relationship and mutual dependency.

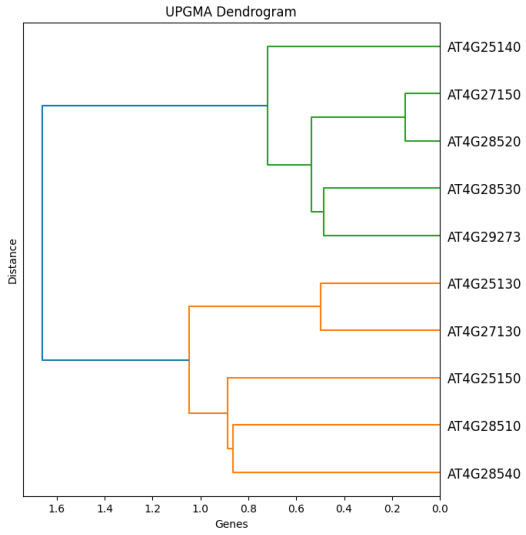

Fig. 5. Visualization of gene regulatory relationships using UPGMA dendrogram: analysis of gene dependency among AT4G28520, AT4G27150, and AT4G25140.

This finding is supported by insights from research on the ABI3 regulon [9] and validates Py/ExplorReg's capability in identifying complex gene regulatory networks. The ABI3 regulon, a pivotal factor influencing these genes, plays a central role in seed development and maturation processes. ABI3 is known for its comprehensive involvement in regulating genes critical for seed storage, desiccation tolerance, and embryo development.

VI. Conclusion

In conclusion, our case study has introduced Py/ExplorReg for exploring gene expression and regulation. By incorporating both Pearson's correlation coefficient and mutual information, Py/ExplorReg offers a unique approach to uncovering potential regulons, providing a deeper understanding of gene regulatory networks. Unlike traditional methods that rely solely on database searches, Py/ExplorReg integrates gene expression data, thereby enhancing the probability of regulon prediction.

Through testing on meaningful datasets and implementation in Python, Py/ExplorReg sets a new standard in algorithmic analysis for potential regulon structures. It can explore and identify a significant number of co-expressed genes, which can subsequently be confirmed through literature, database searches, or new wet-lab experiments as genes belonging to a single regulon. This capability provides a new computational approach for easier data mining within gene expression and enables faster focus on specific *in-silico* potential regulon structures. This study presents Py/ExplorReg, which, by reproducing potential regulon structures, opens up new possibilities for further research unifying cis-regulatory elements in *Arabidopsis thaliana*.

Overall, Py/ExplorReg is applicable for researchers studying gene expression and regulation, offering enhanced capabilities for exploring complex regulatory networks. Its availability as an open-source implementation on GitHub (https://github.com/pa3cka/PyExplorReg) ensures accessibility within the field of computational biology.

ACKNOWLEDGMENT

This work has been supported by grant project FEKT/FIT-J23-8274.

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

REFERENCES

[1] V. M. Weake and J. L. Workman, "Inducible gene expression: diverse regulatory mechanisms", *Nature Reviews Genetics*, vol. 11, no. 6, pp. 426-437, 2010.

[2] H. Zare, D. Sangurdekar, P. Srivastava, M. Kaveh, and A. Khodursky, "Reconstruction of Escherichia coli transcriptional regulatory networks via regulon-based associations", *BMC Systems Biology*, vol. 3, no. 1, 2009.

[3] J. Schwarzerová, "Operon-Expresser: The Innovated Gene Expression-Based Algorithm For Operon Structures In-Ference", in *Proceedings I of the 27st Conference STUDENT EEICT 2021: General papers*, 2021, pp. 301-305.

[4] J. Nejezchlebová and J. Schwarzerová, "Operon identifier: Identification of operon structures in the whole genome", in *Proceedings II of the 28st Conference STUDENT EEICT 2022: Selected papers*, 2022, pp. 80-83.

[5] A. Almudevar, L. B. Klebanov, X. Qiu, P. Salzman, and A. Y. Yakovlev, "Utility of correlation measures in analysis of gene expression", *NeuroRX*, vol. 3, no. 3, pp. 384-395, 2006.

[6] Z. Zhong, Y. Wang, M. Wang, F. Yang, Q. A. Thomas, Y. Xue, Y. Zhang, W. Liu, Y. Jami-Alahmadi, L. Xu, S. Feng, S. Marquardt, J. A. Wohlschlegel, I. Ausin, and S. E. Jacobsen, "Histone chaperone ASF1 mediates H3.3-H4 deposition in Arabidopsis", *Nature Communications*, vol. 13, no. 1, 2022.

[7] *Regulon.* Online. National Library of Medicine. 1993.

[8] RODIONOV, Dmitry A. Comparative Genomic Reconstruction of Transcriptional Regulatory Networks in Bacteria. Online. *Chemical Reviews.* 2007, s. 3467-3497.

[9] G. Mönke, M. Seifert, J. Keilwagen, M. Mohr, I. Grosse, U. Hähnel, A. Junker, B. Weisshaar, U. Conrad, H. Bäumlein, and L. Altschmied, "Toward the identification and regulation of the Arabidopsis thaliana ABI3 regulon", *Nucleic Acids Research*, vol. 40, no. 17, pp. 8240-8254, Sep. 2012.

[10] J. Bernard and J. Bernard, "Python Data Analysis with pandas", in *Python Recipes Handbook*, Berkeley, CA: Apress, 2016, pp. 37-48.

[11] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy", *Nature*, vol. 585, no. 7825, pp. 357-362, Sep. 2020.

[12] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. P. Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G. -L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko, and Y. Vázquez-Baeza, "SciPy 1.0: fundamental algorithms for scientific computing in Python", *Nature Methods*, vol. 17, no. 3, pp. 261-272, Mar. 2020.

[13] O. Kramer and O. Kramer, "Scikit-Learn", in *Machine Learning for Evolution Strategies*, Cham: Springer International Publishing, 2016, pp. 45-53.

[14] I. Priness, O. Maimon, and I. Ben-Gal, "Evaluation of gene-expression clustering via mutual information distance measure", *BMC Bioinformatics*, vol. 8, no. 1, 2007.

[15] I. Gronau and S. Moran, "Optimal implementations of UPGMA and other common clustering algorithms", *Information Processing Letters*, vol. 104, no. 6, pp. 205-210, 2007.