Re_actShap : détection de rebranchements des réseaux de régulation d'expression génique à l'aide des valeurs SHAP

Lisa Chabrier[**,*,****] Anton Crombach[*,**]
Sergio Peignier[***] Christophe Rigotti[**,*]

[*]Inria, Centre de Lyon
69603 Villeurbanne, France
[**]INSA Lyon, CNRS, Universite Claude Bernard Lyon 1,
LIRIS, UMR5205,
69621, Villeurbanne, France
[***]INSA Lyon, INRAE, BF2i, UMR0203
69621, Villeurbanne, France
[****]lisa.chabrier@inria.fr

Résumé. Nous présentons dans cet article une méthode permettant de détecter les modifications dans les réseaux de régulation géniques, modélisant les interactions de régulations entre les gènes. Ces interactions varient d'un type cellulaire à un autre, et ces changements sont fondamentaux pour comprendre les mécanismes en jeux dans les cellules. Nous proposons une méthode de détection qui requière uniquement les données issues de la technique dite du single-cell RNA-Sequencing. Cette restriction permet de rendre la méthode applicable à un grand nombre d'espèces et de types cellulaires. En premier lieu, des modèles d'apprentissage automatiques sont entraînés à prédire l'expression d'un gène cible à partir de l'expression d'autres gènes appelés facteurs de transcriptions. Grâce à une méthode d'IA explicable basée sur les valeurs SHAP nous décrivons l'activité du réseau de régulation génique dans chacune des cellules. En détectant les interactions de régulation différentiellement exprimées dans un type cellulaire, nous faisons des hypothèses sur les modifications de la régulation.

1 Introduction

Les données obtenues avec la technique dite single-cell RNA sequencing (scRNA-seq) permettent d'observer l'hétérogénéité de l'expression des gènes entre les cellules d'une même espèce. Comprendre cette hétérogénéité est important, car elle est à l'origine de nombreux mécanismes biologiques. Cette hétérogénéité est en partie due à la régulation des gènes par des facteurs de transcriptions, des protéines produite à partir des gènes du même nom, et qui ont la capacité d'influencer l'expression d'autre gènes, les gènes cible. Les facteurs de transcription, ou Transcription Factors en anglais, seront abrégés TFs, et les gènes cibles, ou Target Genes en anglais, seront abrégés TGs. L'ensemble des liens de régulation entre TFs et TGs forme un réseau appelé réseau de régulation génique (Gene Regulatory Network en anglais,

abrégé GRN). Il a été montré que ces réseaux ne sont pas statiques (Srivastava et Mahony, 2020; Whyte et al., 2013), un même gène pouvant être régulé par des TFs différents dans des contextes différents (par exemple des types cellulaires différents). Ceci est représenté dans la figure 1. Ces changements sont appelés les rebranchements des GRNs.

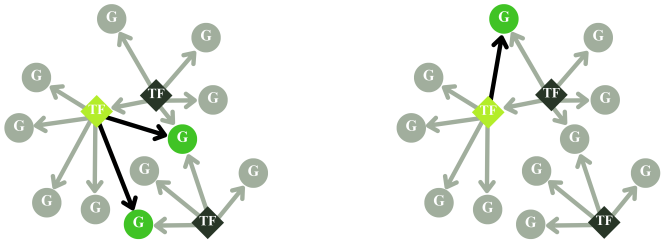

FIG. 1 – *Rebranchement des liens de régulation entre deux types cellulaires.*

Il existe différentes méthodes computationnelles qui s'intéressent à ces rebranchements dans les GRNs. Ces méthodes nécessitent souvent plusieurs types de données d'entrées différents, en plus des données scRNA-seq. Ces formes de données d'entrées sont issues des techniques dites sc-ATAC-Seq et sc-ChIP-Seq, ou encore des réseaux d'interaction entre protéines. Or, pour la plupart des espèces, ces types de données ne sont pas accessibles. En effet, pour les techniques sc-ATAC-Seq et sc-ChIP-Seq, l'adaptation des protocoles d'expérience à une nouvelle espèce, ou a un nouveau type cellulaire, est coûteux et long, sans garantie de réussite. Pour les réseaux d'interactions entre protéines, ils consistent en un archivage méticuleux de nombreuses sources d'information pour constituer un ensemble cohérent. La quantité d'études nécessaires permettant d'obtenir ces réseaux n'existe que pour quelques espèces très étudiées : l'humain, les souris ou la drosophile par exemple.

Nous proposons dans cet article une méthode - re_actShap - qui utilise uniquement des données issues du scRNA-seq pour la détection des rebranchements des GRNs. Cette méthode est compatible avec l'environnement de développement "scanpy", qui est très utilisé pour l'analyse de ce type des données scRNA-seq.

2 Etat de l'art

Des études suggèrent que les événements de rebranchements ont un fort impact phénotypique (Bhardwaj et al., 2010). Ils sont impliqués dans la prise de décisions cellulaire (Davis et Rebay, 2017), l'adaptation à l'environnement (Brooks et al., 2011), la réaction au stress (Califano, 2011) et la division cellulaire (Karlebach et Shamir, 2008). Le nombre croissant d'études publiées mentionnant les relations de régulation nous aide à comprendre l'importance du rebranchement.

La détection des événements de rebranchements est un défi majeur pour la compréhension des processus cellulaires. Elle pourrait permettre de mieux comprendre les différences entre les types cellulaires, les états cellulaires, les états pathologiques ou encore les effets des traitements.

Cependant, cette tâche est encore exploratoire. Plusieurs articles ont été publiés dans le but de détecter des rebranchements sur un ensemble de données pré-défini, avec une question biologique spécifique. Afin de détecter des rebranchements entre deux conditions biologiques différentes, la plupart des articles se limitent à appliquer des opérations ensemblistes sur les GRNs inférés pour les deux conditions. Cependant, la revue de Badia-i Mompel et al. (2023) suggère que de telles approches ne sont pas assez robustes.

D'autres articles proposent des méthodes générales pouvant être utilisées dans différents contextes d'étude. Ces méthodes proposent d'enrichir l'approche des opérations ensemblistes en usant d'outils issus de l'apprentissage automatique et des statistiques. Dans le tableau 1 sont listés les méthodes que nous avons identifiées.

Méthode	données d'entrées requises	article
ANANSE	scRNA-seq, APH, sc-ATAC-Seq, sc-ChIP-Seq	Xu et al. (2021)
sc-compreg	pour deux populations : scRNA-seq, sc-ChIP-Seq	Duren et al. (2021)
RNCMI	scRNA-seq, RIPP	Xie et al. (2021)
CEFCON	scRNA-seq, RIPP	Wang et al. (2023)
TopicNet	scRNA-seq, sc-ChIP-Seq	Lou et al. (2020)

TAB. 1 – *Differentes methodes de detection des rebranchements des GRNs. APH signifie Atlas des Protéines Humaines, et RIPP signifie Réseau d'Interaction Proteine-Proteine.*

Comme l'indique ce tableau, toutes les méthodes utilisent des données scRNA-seq, ainsi que des jeux de données complémentaires, dans une approche dite multi-omique. Pour la plupart des espèces, il n'est pas aisé d'obtenir tous ces jeux de données, car ils nécessitent l'adaptation des techniques d'acquisitions à chaque nouvelle espèce, et cette adaptation est longue et coûteuse. Par conséquent, ces données ne sont souvent pas disponibles, en particulier pour des organismes qui ne font pas partie des quelques organismes modèles communément étudiés.

3 Présentation de re-actShap

Dans cet article, nous souhaitons nous appuyer uniquement sur l'utilisation des données de type scRNA-seq. La méthode que nous proposons, re_actShap, est inspirée d'un autre outil nommé Arboreto présentées dans Moerman et al. (2019). Cette méthode est très populaire, et permet l'inférence des GRNs. Le fonctionnement de re_actShap est le suivant : un modèle d'apprentissage automatique est entraîné à la prédiction de l'expression d'un gène, en utilisant comme paramètre pour la prédiction les expressions des facteurs de transcription (TFs). Ce modèle apprend les relations complexes qui lient les TFs au gène cible.

À la différence d'Arboreto, nous proposons d'utiliser une méthode d'explicabilité locale. Pour chacune des cellules, les TFs qui jouent un rôle dans la prédiction du modèle se voient attribuer un score d'importance. Nous sélectionnons les 10 meilleurs TFs pour chacun des gènes (avec ex aequo). Ces liens TF $\rightarrow$ TG sont considérés comme *activés* dans la cellule. Nous avons choisi l'emploi des valeurs SHAP (Lundberg et Lee, 2017) appliquées à l'explicabilité de l'apprentissage automatique qui est une méthode d'explicabilité locale qui propose des garanties d'équité, ce qui nous semble adapté. Dans la pratique, le calcul de ces valeurs est très

coûteux, et nous proposons d'en calculer des approximations. De plus, nous sommes intéressés par les TFs les plus importants, et nous sélectionnons donc les k TFs associés aux valeurs SHAP les plus élevées. Afin de réduire le coût de calcul des valeurs SHAP, nous utilisons la méthode TopShap, décrite dans Chabrier et al. (2024) qui permet d'obtenir les top-k valeurs SHAP en élaguant l'espace de recherche lors d'un processus d'approximation itératif. Nous recueillons toutes ces valeurs SHAP qui correspondent chacune au niveau d'activation d'un des liens du GRN dans une des cellules de la population. Les cellules sont maintenant toutes représentées par un vecteur indiquant l'activité de chacun des liens de régulation, et sont donc représentées dans un nouvel espace.

Nous utilisons cet espace pour détecter des tendances dans les activations des cellules de différents types cellulaires. Pour ce faire, nous utilisons un test statistique de Student. Les liens de régulation qui sont statistiquement plus activés dans un type cellulaire comparé au reste de la population sont qualifiés de liens de régulation *marqueurs* de ce type cellulaire. En comparant les liens détectés pour chacun des types cellulaires, il est possible de détecter les TGs qui seraient régulés par des TFs différents dans des types cellulaires différents.

4 Démonstration

Nous proposons dans cette partie une démonstration de l'usage qui peut être fait de cette méthode. Pour ce faire, nous utiliserons le jeu de données PBMC 3k mis à disposition par 10X Genomics [1]. Le jupyter notebook utilisé pour cette démonstration, ainsi que les données sont accessibles au lien suivant : `https://gitlab.inria.fr/topshap/Re_actShap_demonstration`

Nous distinguons plusieurs étapes. En premier lieu, il faut nettoyer les données. Ensuite, interviens une étape de clustering des cellules et d'identification des types cellulaires parmi les différents groupes formés par le clustering. Enfin, une fois les groupes identifiés, nous calculons les représentations des cellules dans l'espace des liens de régulation [2]. La dernière consiste à appliquer un test de Student. Pour chacun des liens de régulation, ce test permet l'obtention d'un t-score et d'un p-value. Plus le t-score est élevé, plus le lien de régulation est sur-activé dans le type cellulaire, et sous-activé ailleurs. Plus la p-value est faible, plus le résultat du test est significatif. Ceci permet le classement des liens de régulation par t-score pour chacun des types cellulaires. Nous sélectionnons les 20 premiers liens et les nommons liens de régulations marqueur du type cellulaire. Nous présentons dans les figures 2, 3 et 4 les différentes visualisations que nous utilisons pour analyser les résultats.

Dans la figure 2, nous représentons les liens marqueurs de chacun des types cellulaires avec leurs expressions dans les différents groupes de cellules. Chacun des diagrammes montre les 20 liens de régulation marqueurs (axe y), avec la valeur du t-score associé représenté comme la largeur de la barre (axe x, gauche), et les points représentant par leur taille la proportion de cellules du groupe pour lesquelles ce lien est actif, et par leur couleur la moyenne de la valeur

1. les données sont accessibles au lien suivant : `https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k`

2. Les étapes de pré-processing et de clustering des cellules sont des taches non-triviales, mais classique pour l'étude des jeux de données single-cell. Classiquement, la production d'un jeu de données prend plusieurs mois, et les traitements bio-informatiques des données produites plusieurs mois supplémentaires. Ici, le temps de calcul d'une matrice d'association est de l'ordre de la journée, ce qui en fait un outil compatible avec les délais usuel de ce type d'études.

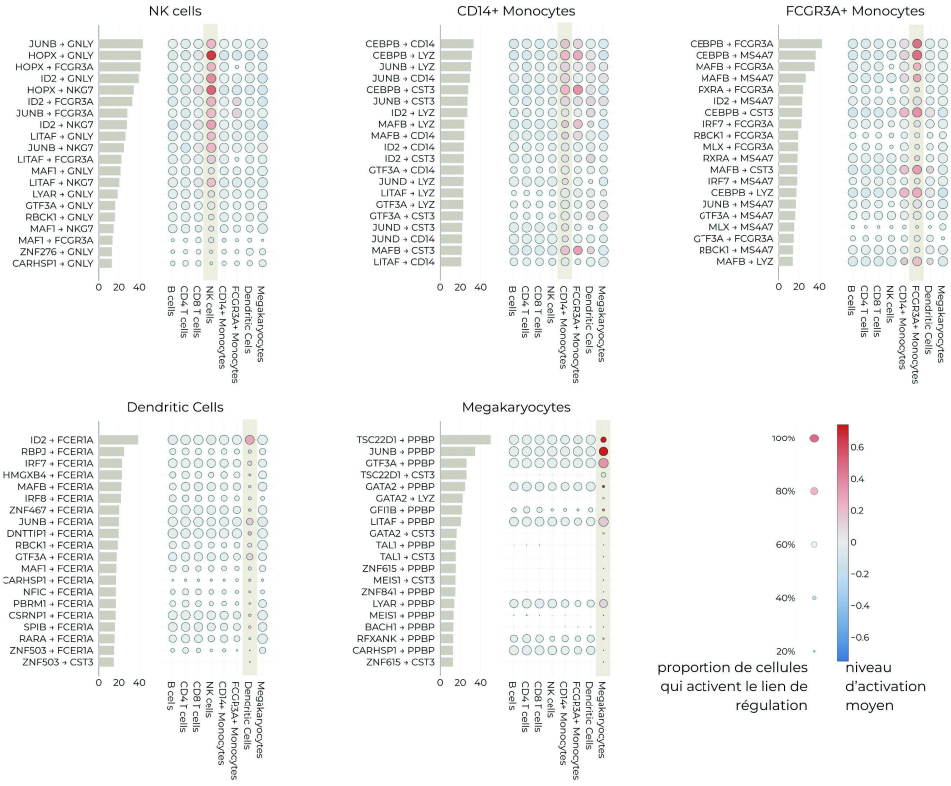

FIG. 2 – *Les niveaux d'activités des liens de régulations détectés comme marqueurs dans chacun des types cellulaires. Les types cellulaires B cells, CD4 T cells et CD8 T cells ne sont pas représentés ici, mais les figures correspondantes sont disponibles sur le dépot Git.*

d'activation dans le groupe de cellule (axe x, droite). On peut voir sur les graphiques que les niveaux d'activation des différents liens marqueurs sont plus important dans le type cellulaire pour lequel ils sont détectés comme marqueurs. Certains types proches biologiquement ont des profils d'activation similaire, par exemples les monocytes CD14+ et FCGR3A+. Au contraire, certains types cellulaires ont des profils d'activation unique, tels que les Megakaryocytes. On remarque par ailleurs que les gènes cible des liens de régulation marqueurs (le TG dans le lien TF→TG) sont souvent très uniformes parmi les liens détectés pour un même type cellulaire.

Dans la figure 3, nous observons la représentation des liens marqueurs sous la forme de liens dans un réseau. La couleur du lien correspond au type cellulaire pour lequel le lien est détecté comme statistiquement différent. Certains liens sont représentés par des lignes pointillées quand ils sont détectés dans plusieurs types cellulaires.

La figure 4 est une représentation synthétique du réseau. La couleur des barres représente le type cellulaire dans lequel est détecté le lien de régulation qui a pour cible le gène. Cette figure permet d'identifier rapidement les gènes qui sont impliqués dans des liens de régulation dans plusieurs types cellulaires. C'est le cas pour les gènes NKG7 et FCGR3A par exemple. Nous

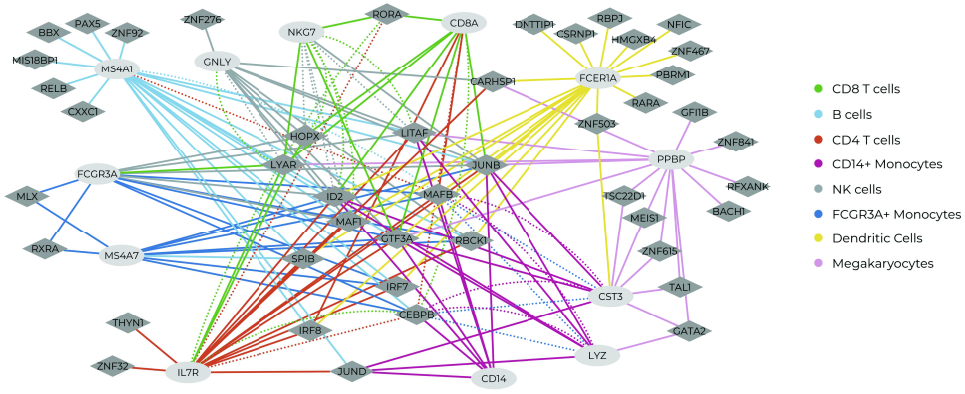

FIG. 3 – *Les liens de régulation détéctés pour chacun des types cellulaires représentés dans un réseau. La couleur des arretes correspond au type cellulaire dans lequel ce lien est detecté. Certains liens sont détecté dans plusieurs types cellulaires, et les arretes correspondantes sont représentées en pointillé.*

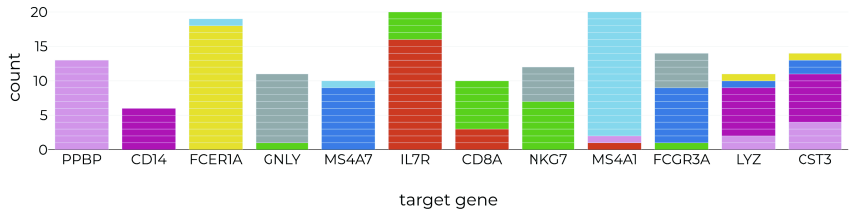

FIG. 4 – *Pour chacun des gènes cible, sont représentés le nombre de régulateurs détectés. Les couleurs des barres représentent le nombre de régulateurs détectés dans chacun des types cellulaires.*

faisons ici l'analyse suivante : ces gènes apparaissent comme étant régulés par des groupes de TFs différents dans deux types cellulaires. Il est possible que ceci soit la matérialisation des rebranchements dans le GRN. Il faudra vérifier ces résultats expérimentalement avant de pouvoir affirmer que c'est bien le cas. Néanmoins, cet outil permet de générer des hypothèses et de limiter le nombre d'expériences à envisager.

5 Conclusion

Cet article présente la méthode re_actShap, qui permet la détection d'événement de rebranchements dans les réseaux de régulation de l'expression génique (GRNs). La nouveauté proposée par cette méthode réside dans le fait que seul un jeu de données single-cell RNA sequencing (scRNA-seq), et une liste de gènes spécifiques, dit facteurs de transcription (TFs). Avec seulement ces deux entées, nous proposons d'entraîner des modèles d'apprentissage au-

tomatiques à prédire l'expression des gènes cible à partir de l'expression des TFs. Grâce à une méthode d'explicabilité des prédictions des modèles basées sur les valeurs SHAP nommée topShap, nous construisons une matrice d'association dans laquelle chaque cellule du jeu de données est représentée par son état d'activation du GRN. Puis grâce à un test de Student, nous détectons les liens de régulation les plus différentiellements activés dans chacun des types cellulaires. Grâce à plusieurs visualisations complémentaires, nous pouvons proposer des hypothèses sur les rebranchements du GRN à l'œuvre dans le jeu de données étudié. Cette méthode est donc applicable dans un grand nombre d'études. Afin de faciliter l'accès à la méthode re_actShap, nous avons rendu l'implémentation compatible avec l'outil scanpy, qui est très utilisé pour l'analyse des données de type scRNA-seq. La méthode, ainsi qu'un jupyter notebook sont accessible sur un dépôt Git : `https://gitlab.inria.fr/topshap/Re_actShap_demonstration`

Références

Badia-i Mompel, P., L. Wessels, S. Müller-Dott, R. Trimbour, R. O. Ramirez Flores, R. Argelaguet, et J. Saez-Rodriguez (2023). Gene regulatory network inference in the era of single-cell multi-omics. *Nature Reviews Genetics 24*(11), 739–754. Publisher : Nature Publishing Group UK London.

Bhardwaj, N., P. M. Kim, et M. B. Gerstein (2010). Rewiring of transcriptional regulatory networks : Hierarchy, rather than connectivity, better reflects the importance of regulators. *Science Signaling 3*(146), –, doi: 10.1126/scisignal.2001014.
en

Brooks, A. N., S. Turkarslan, K. D. Beer, F. Yin Lo, et N. S. Baliga (2011). Adaptation of cells to new environments. *WIREs Systems Biology and Medicine 3*(5), 544–561, doi: 10.1002/wsbm.136.

Califano, A. (2011). Rewiring makes the difference. *Molecular Systems Biology 7*(1), 463, doi: 10.1038/msb.2010.117. Publisher : John Wiley & Sons, Ltd.

Chabrier, L., A. Crombach, S. Peignier, et C. Rigotti (2024). Effective pruning for top-k feature search on the basis of SHAP values. *IEEE Access 12*, 1–1, doi: 10.1109/ACCESS.2024.3489958. Conference Name : IEEE Access.
en

Davis, T. L. et I. Rebay (2017). Master regulators in development : Views from the Drosophila retinal determination and mammalian pluripotency gene networks. *Developmental Biology 421*(2), 93–107, doi: 10.1016/j.ydbio.2016.12.005.

Duren, Z., W. S. Lu, J. G. Arthur, P. Shah, J. Xin, F. Meschi, M. L. Li, C. M. Nemec, Y. Yin, et W. H. Wong (2021). Sc-compReg enables the comparison of gene regulatory networks between conditions using single-cell data. *Nature Communications 12*(1), 4763, doi: 10.1038/s41467-021-25089-2.
en

Karlebach, G. et R. Shamir (2008). Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology 9*(10), 770–780, doi: 10.1038/nrm2503. Publisher : Nature Publishing Group.

Lou, S., T. Li, X. Kong, J. Zhang, J. Liu, D. Lee, et M. Gerstein (2020). TopicNet : a framework for measuring transcriptional regulatory network change. *Bioinformatics 36*, i474–i481, doi: 10.1093/bioinformatics/btaa403.

Lundberg, S. M. et S.-I. Lee (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.
eng

Moerman, T., S. Aibar Santos, C. Bravo González-Blas, J. Simm, Y. Moreau, J. Aerts, et S. Aerts (2019). GRNBoost2 and Arboreto : efficient and scalable inference of gene regulatory networks. *Bioinformatics (Oxford, England) 35*(12), 2159–2161, doi: 10.1093/bioinformatics/bty916.
eng

Srivastava, D. et S. Mahony (2020). Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns. *Biochimica Et Biophysica Acta. Gene Regulatory Mechanisms 1863*(6), 194443, doi: 10.1016/j.bbagrm.2019.194443.

Wang, P., X. Wen, H. Li, P. Lang, S. Li, Y. Lei, H. Shu, L. Gao, D. Zhao, et J. Zeng (2023). Deciphering driver regulators of cell fate decisions from single-cell transcriptomics data with CEFCON. *Nature Communications 14*(1), 8459, doi: 10.1038/s41467-023-44103-3.

Whyte, W. A., D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee, et R. A. Young (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell 153*(2), 307–319. Publisher : Elsevier.

Xie, J., Y. Yin, F. Yang, J. Sun, et J. Wang (2021). Differential network analysis reveals regulatory patterns in neural stem cell fate decision. *Interdisciplinary Sciences : Computational Life Sciences 13*(1), 91–102, doi: 10.1007/s12539-020-00415-2.

Xu, Q., G. Georgiou, S. Frölich, M. van der Sande, G. J. C. Veenstra, H. Zhou, et S. J. van Heeringen (2021). ANANSE : an enhancer network-based computational approach for predicting key transcription factors in cell fate determination. *Nucleic Acids Research 49*(14), 7966–7985, doi: 10.1093/nar/gkab598.

Summary

In this article, we present the re_actShap method that allows the detection of rewiring events in the Gene Regulatory Networks (GRNs). The novelty of this method is that the only required inputs are a sc-RNA-Seq datasets dataset, and a list of transcription factors (TFs) for the species. We propose to train machine learning models to predict the expression of target genes with the expression of the TFs. Then, we use an explainability method based on the SHAP values: topShap. With the output of topShap, we build an association matrix, in which each cell is represented by its activation of the GRNs regulatory links. We use a Student test to detect the regulatory associations that are differentially activated in each cell type. We use several visualizations to generate hypotheses on the rewiring events in the GRN. The method is compatible with the tool scanpy, widely used to analyze sc-RNA-Seq datasets.