Gene networks: Inference, evaluation, usage, and beyond

Nathalie Vialaneix

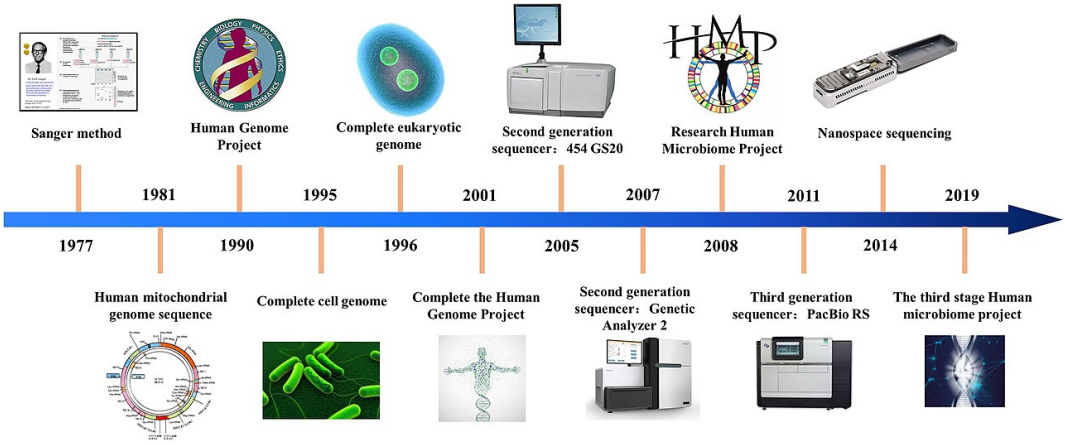
nathalie.vialaneix@inrae.fr
http://www.nathalievialaneix.eu

Journées de Statistique de la SFdS, 28 mai 2024

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

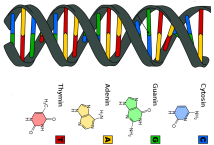
INRAE

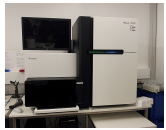
The revolution of "next Generation" sequencing technology



[Yang et al., 2020]

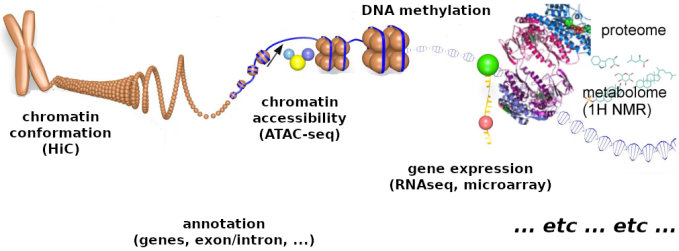
The revolution of “next Generation” sequencing technology



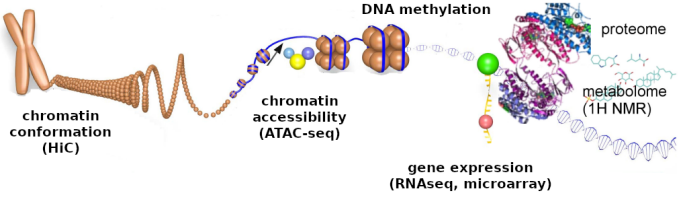


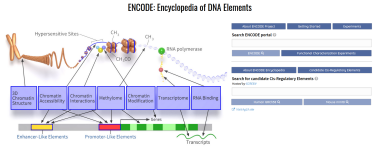
```
@A00318:471:HN2VVDRX3:1:1101:2636:1016 1:N:0:TAAGCAACTG+TTGAGTATAG
CNATTTCAGAGTGTGCAAAATTAGTCGG
+
,#FFFFFFFF:F:FFFFF,,FFFFFFFF
@A00318:471:HN2VVDRX3:1:1101:3360:1016 1:N:0:TAAGCAACTG+TTGAGTATAG
TNTGATGAGGCCATAGGTTATCTTTTAA
+
F#:FFFFF:FFFF:F,:FFFFFFFF,FF
@A00318:471:HN2VVDRX3:1:1101:3830:1016 1:N:0:TAAGCAACTG+TTGAGTATAG
CNAGCTATCGCAACATGCCCATAGATTT
+
F#FFFFFFFFFFFFFF,FFFFFFFFFF
@A00318:471:HN2VVDRX3:1:1101:4499:1016 1:N:0:TAAGCAACTG+TTGAGTATAG
TNCGTGTGTTCTCGTCTGGCATACTAAG
+
F#FFFFFFFFFFFFFFFFFFFFFFFFF
@A00318:471:HN2VVDRX3:1:1101:4806:1016 1:N:0:TAAGCAACTG+TTGAGTATAG
TNCAGGACACGAAGACGACAAGTGTTAC
+
```

Collected data at genomic level are increasingly available



Collected data at genomic level are increasingly available





TCGA

GTExPortal

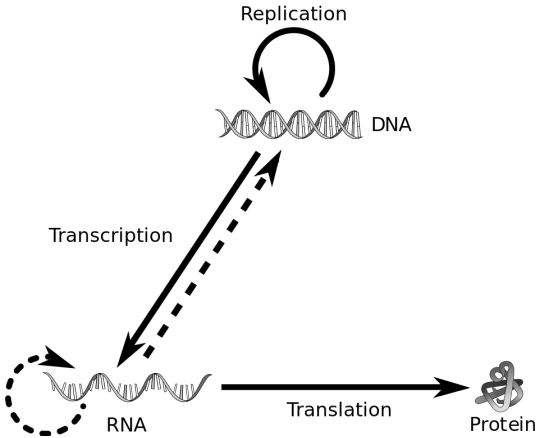
TARA OCEANS

FAANG
Functional Annotation of Animal Genomes

FR-AgENCODE

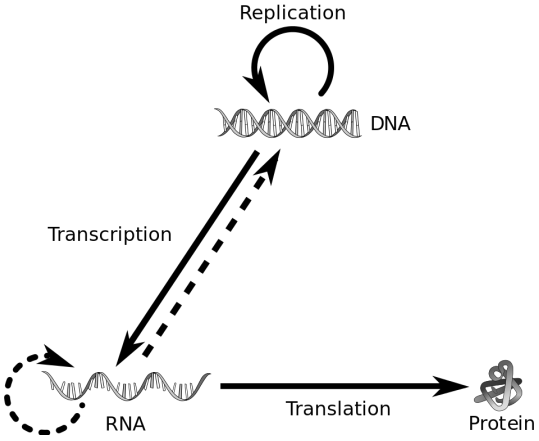
[Foissac et al., 2019]

Molecular biology dogma

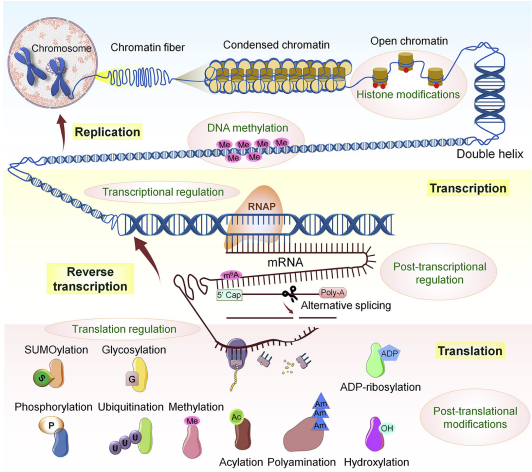


[Barillot et al., 2012]

Molecular biology dogma... and beyond

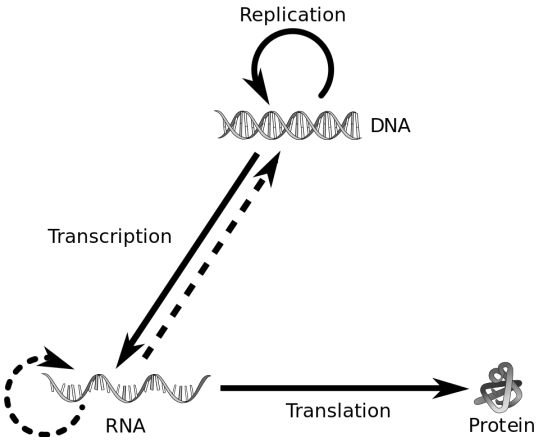


[Barillot et al., 2012]

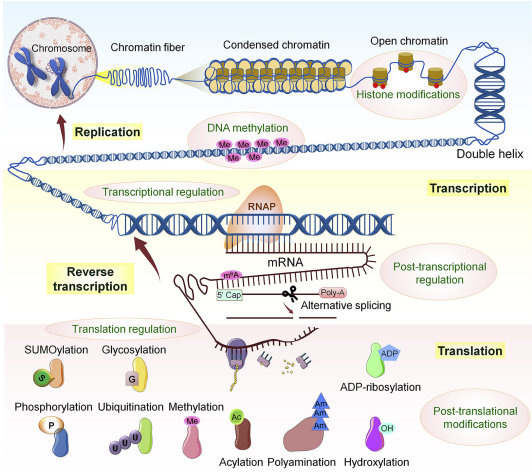


[Pramanik et al., 2021]

Molecular biology dogma... and beyond



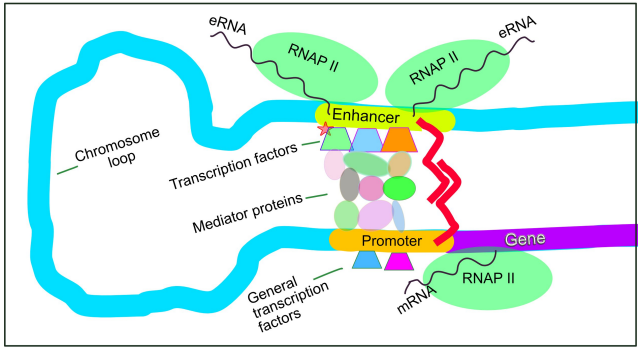
[Barillot et al., 2012]



[Pramanik et al., 2021]

Main rule in biology: "There's no rule!"

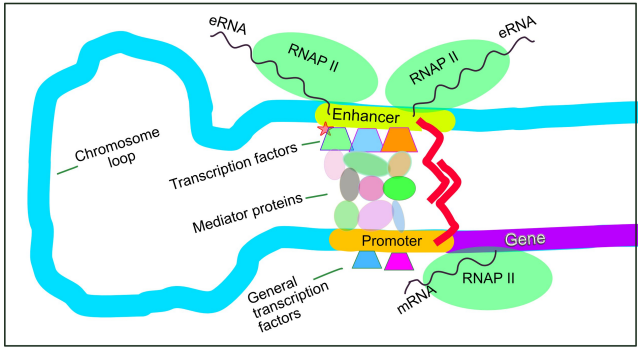
Interactions at a molecular levels: gene regulation(s)



What do we expect in the data?

Level of transcription (mRNA) of TF ↗ ⇒ mRNA of target gene ↗

Interactions at a molecular levels: gene regulation(s)

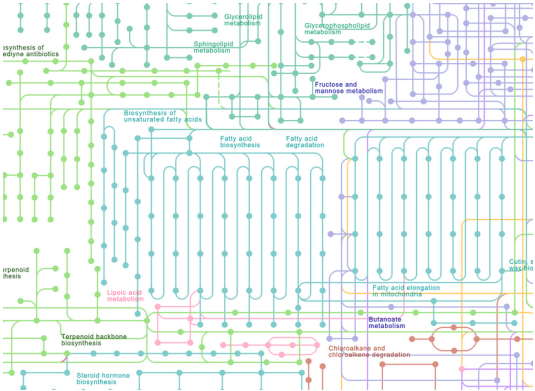


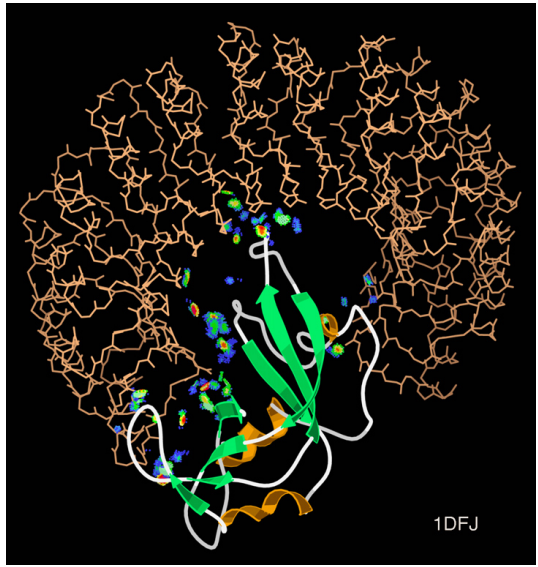
What do we expect in the data?

Level of transcription (mRNA) of TF ↗ ⇒ mRNA of target gene ↗

Network inference: Recover these dependency structures from gene expression

Other gene interaction networks: gene pathways, PPI

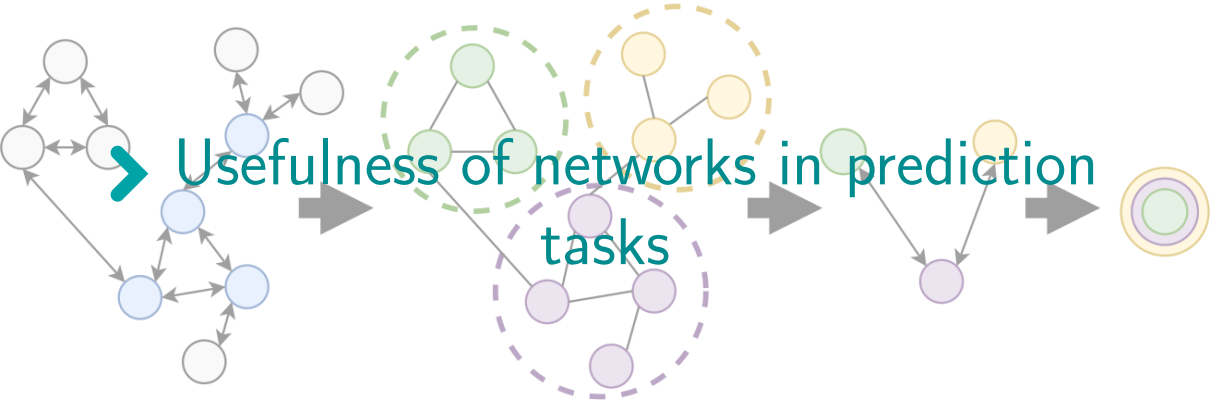




In this talk

- Can (and how) we use network information in predictive biology? Discussion on GNN

- What is the current quality of network inference methods in real-life situations?

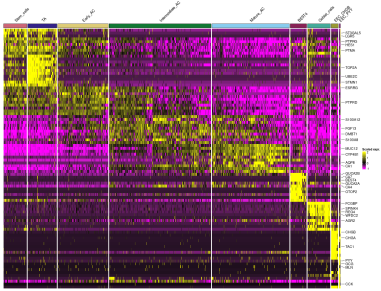


Usefulness of networks in prediction tasks

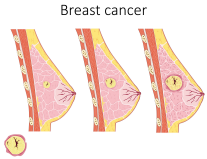
Predictive biology

Learn a classification / regression function:

$\underbrace{\mathbf{X}}_{n \times p}$ → $\underbrace{\mathbf{y}}_{n}$



gene expression



phenotype

Using the graph Laplacian in linear prediction models

Knowing a network, $\mathcal{G}$, with p nodes, $v_1, \ldots, v_p$:

$$\arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \left(\beta^\top \mathbf{x}_i - \mathbf{y}_i \right)^2 + C \beta^\top L \beta \underbrace{+ C' \|\beta\|_1}_{\text{to enforce sparsity}}$$

- **[Li and Li, 2008]**: Y is time to death (Glioblastoma)
- $\Rightarrow$ implemented in R package **glmgraph** (not maintained, archived on CRAN) **[Chen et al., 2015]**

Similar idea using eigendecomposition of the Laplacian

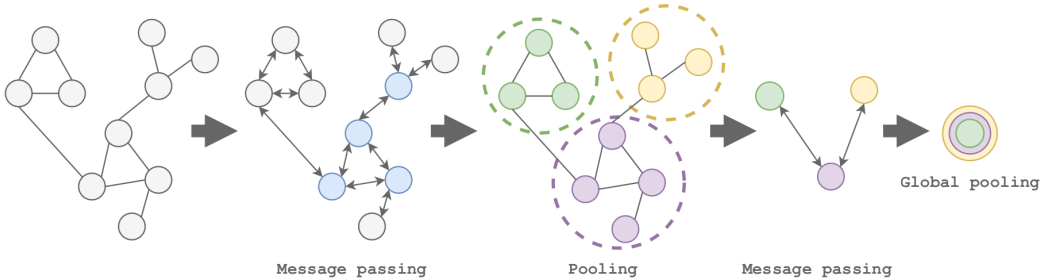
[Rapaport et al., 2007] (Y is irridiated / not irridiated)
L is symmetric and non-negative with $(\lambda_j)_{j=1,\dots,p}$ eigenvalues (in increasing order) and $(e_j)_{j=1,\dots,p}$ orthonormal eigenvectors

$$\arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \left(\beta^\top \mathcal{S}(\mathbf{x}_i) - \mathbf{y}_i\right)^2 + C\|\beta\|_2^2 \qquad \text{with } \mathcal{S}(\mathbf{x}_i) = \sum_{j=1}^{p} \mathbf{x}_{ij}\phi(\lambda_j)e_j$$

Penalize non-smoothness of $\mathbf{x}_i$ over the network:

- low pass filter: $\phi(\lambda_j) = \lambda_j$ for $\lambda_j < \lambda^*$ and 0 otherwise
- attenuation of high frequencies $\phi(\lambda_j) = e^{-\beta\lambda_i}$

Alternative methods based on DL: Graph Neural Networks (GNN)



[Grattarola and Alippi, 2020]

Implemented in: Python libraries Spektral **[Grattarola and Alippi, 2020]** and PyTorch Geometric **[Fey and Lenssen, 2019]**

Message passing layers

- are the generalization of convolutional layers to graph data
- general concept introduced in **[Gilmer et al., 2017]**

Node features: $\mathbf{x}_j$ (for node v_j) $\longrightarrow$ Node latent representations: $h_j^t \in \mathbb{R}^K$ (computed iteratively for layers $t = 1, \ldots, T$)

$$h_j^{t+1} = F\left(h_j^t;\ \square_{j' \in \mathcal{N}(v_j)} \phi_t(h_j^t, h_{j'}^t)\right)$$

with

- $\square$: differential permutation invariant function (mean, sum...)
- ϕ_t: different possible shapes, involving weights learned during training phase

Message passing layers

- are the generalization of convolutional layers to graph data
- general concept introduced in **[Gilmer et al., 2017]**

Node features: $\mathbf{x}_j$ (for node v_j) $\longrightarrow$ Node latent representations: $h_j^t \in \mathbb{R}^K$ (computed iteratively for layers $t = 1, \ldots, T$)

$$h_j^{t+1} = F\left(h_j^t;\ \square_{j' \in \mathcal{N}(v_j)} \phi_t(h_j^t, h_{j'}^t)\right)$$

with

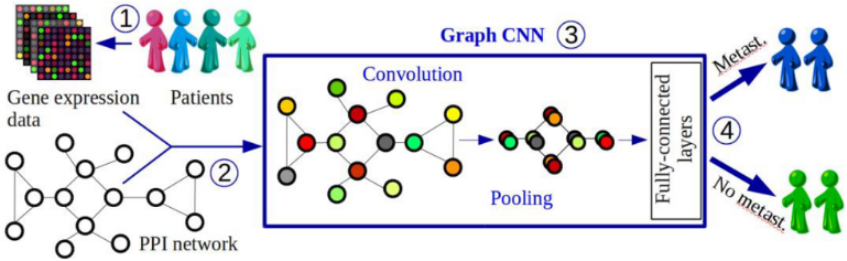
- $\square$: differential permutation invariant function (mean, sum...)
- ϕ_t: different possible shapes, involving weights learned during training phase

In particular: ChebNets **[Defferrard et al., 2016]** (based on Laplacian low band filtering)

Details

Several references show usefulness of GNN in predictive biology

[Chereda et al., 2019, Ramirez et al., 2020, McDermott et al., 2020, Chereda et al., 2021]



Replication study with negative control and simulated data



[Brouard et al., 2024]

Datasets:

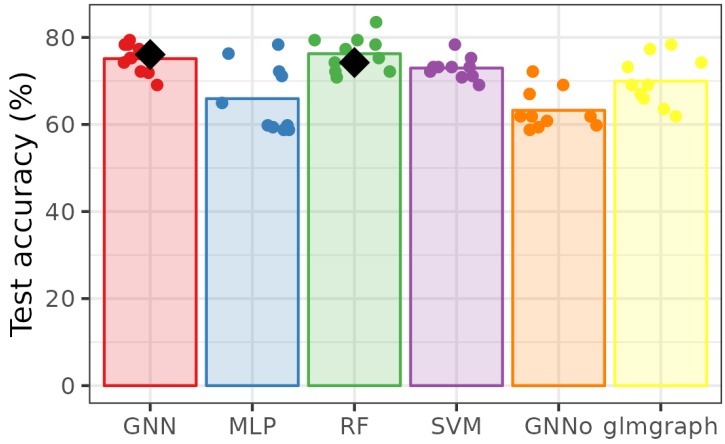
- **BreastCancer** **[Chereda et al., 2019, Chereda et al., 2021]**
  - 969 breast cancer patients belonging to 2 classes:
    - 393 with distant metastasis within the first five years
    - 576 without metastasis having the last follow-up between 5 and 10 years
  - graph : PPI network with 6888 nodes (main connected component)
- CancerType, F1000 ($\times 3$)
- + simulated data: with a mecanistic model and known gene regulatory network **sismonr** **[Angelin-Bonnet et al., 2020]** and DREAM5 dataset **[Marbach et al., 2012]**

Methodology of the replication study

- tested methods: GNN, RF, SVC, perceptron (with or without regularization), glmgraph (including a 5-fold CV to tune hyperparameters)
- with different networks (as negative control): PPI network, correlation network, random network, complete network
- methodology: 10-fold CV (same folds for all methods)

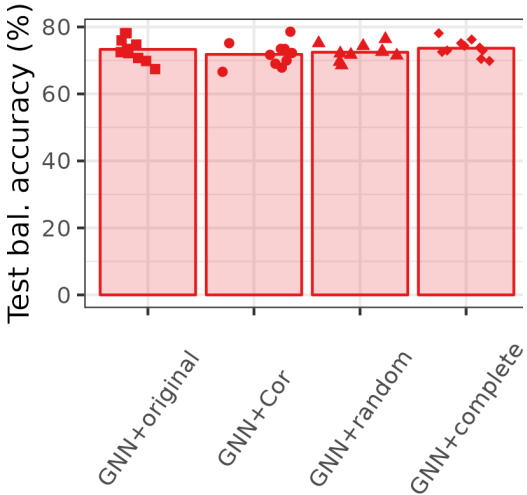
Take home messages from BreastCancer data [Chereda et al., 2021]

GNN results are reproducible but other (less computationnally demanding) methods are as good or better

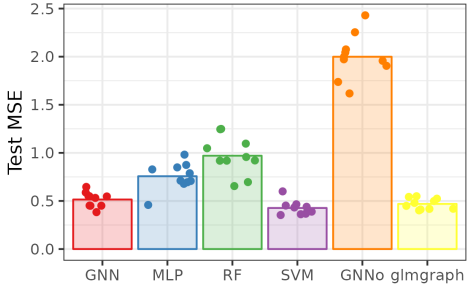


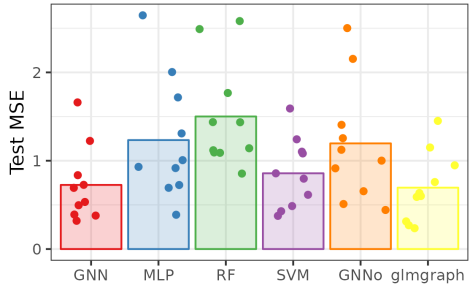
Take home messages from BreastCancer data [Chereda et al., 2021]

Provided network does not influence accuracy for GNN

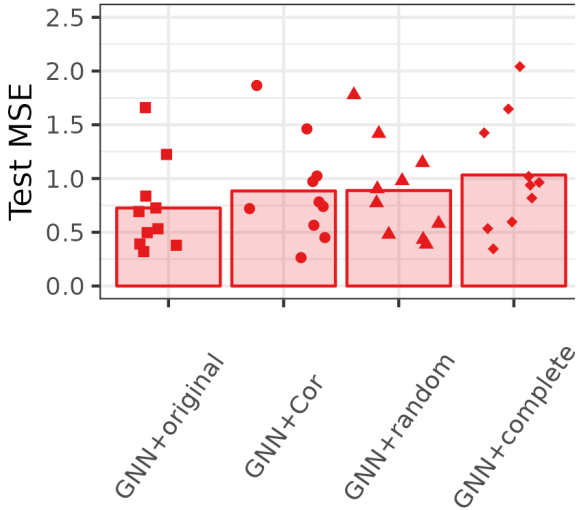


What about simulated datasets?





What about simulated datasets?



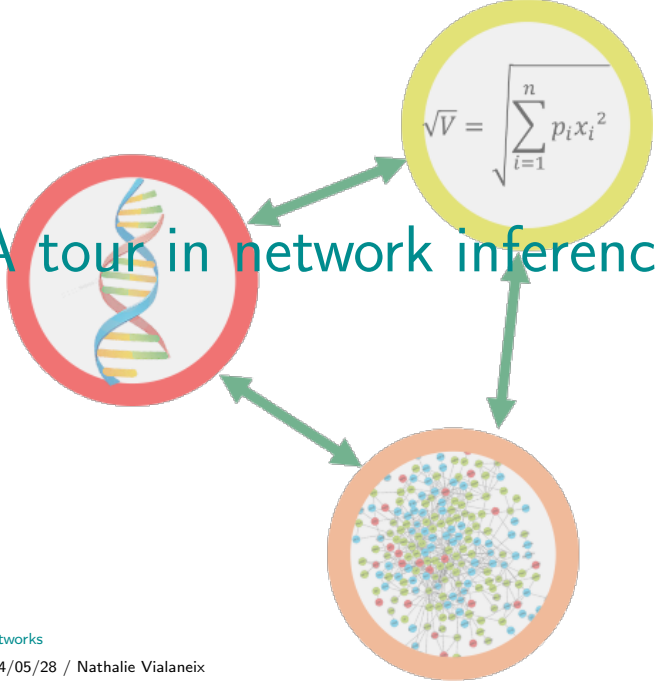
Take home messages

- Reproducibility / replication issues

- If knowing the regulation network can improve inference,
  - GNN are probably not the best candidate at the moment to use this information in an accurate way

  - knowledge on regulation network needs to be good

A tour in network inference



What is network inference?

$(n \times p)$ gene expression matrix $\mathbf{X}$

$\downarrow$

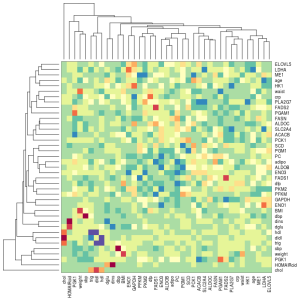
gene network (graph $\mathcal{G} = (V, E)$):

- nodes $V = \{1, \ldots, p\}$: genes
- edges $E \subset V \times V$: "dependency" between gene expressions

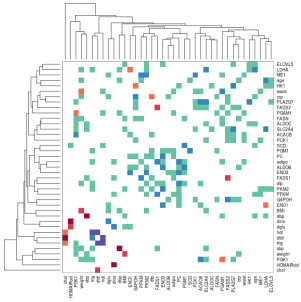
Usually: $n \ll p$

Main directions to address network inference

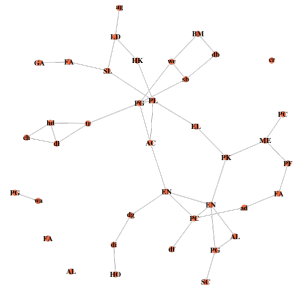
- Most used: "Relevance" network [**Butte and Kohane, 1999**, **Butte and Kohane, 2000**]: correlations (or MI) + threshold



"Correlations"



Thresholding



Network

Main directions to address network inference

- Most used: "Relevance" network **[Butte and Kohane, 1999, Butte and Kohane, 2000]**: correlations (or MI) + threshold

- GGM: Under $\mathbf{X}_i$ i.i.d. $\sim \mathcal{N}_p(0, \Sigma)$: edge between j and j' $\Leftrightarrow \mathbb{C}\mathrm{or}\left(X^j, X^{j'} | (X^k)_{k \neq j,j'}\right) \neq 0$
  - Edge between j and j' $\Leftrightarrow \left[\Sigma^{-1}\right]_{jj'} \neq 0$ **[Friedman et al., 2008]**
  - Edge between j and j' $\Leftrightarrow \beta_{jj'} \neq 0$ in

$$X^j = \sum_{j' \neq j} \beta_{jj'} X^{j'} + \epsilon_j$$

[Meinshausen and Bühlmann, 2006]

Many variants to account for discrete input data, design of the experiment, network structure, priors, ...

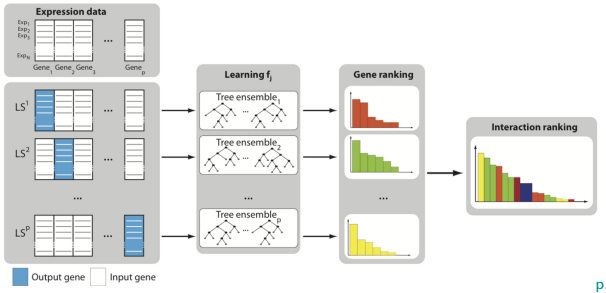
[Ambroise et al., 2009, Chiquet et al., 2011, Chiquet et al., 2016, Mohan et al., 2012, Gallopin et al., 2013, Villa-Vialaneix et al., 2014, Chiquet et al., 2021] ...

Main directions to address network inference

- Most used: "Relevance" network [**Butte and Kohane, 1999**, **Butte and Kohane, 2000**]: correlations (or MI) + threshold
- GGM: Under $\mathbf{X}_i$ i.i.d. $\sim \mathcal{N}_p(0, \Sigma)$: edge between j and j' $\Leftrightarrow \mathbb{C}\mathrm{or}\left(X^j, X^{j'} | (X^k)_{k \neq j,j'}\right) \neq 0$
- Random forest: [**Huynh-Thu et al., 2010**] **GENIE3**

$$\forall j,\ X^j = F_j\left(\{X^{j'}\}_{j' \neq j}\right) + \epsilon_j$$

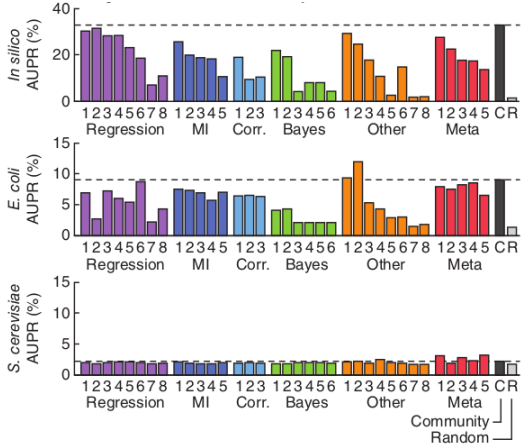
and many variants [**Aibar et al., 2017**, **Petralia et al., 2015**, **Cassan et al., 2023**]



Main directions to address network inference

- Most used: "Relevance" network [**Butte and Kohane, 1999**, **Butte and Kohane, 2000**]: correlations (or MI) + threshold
- GGM: Under $\mathbf{X}_i$ i.i.d. $\sim \mathcal{N}_p(0, \Sigma)$: edge between j and j' $\Leftrightarrow \mathbb{C}\mathrm{or}\left(X^j, X^{j'} | (X^k)_{k \neq j,j'}\right) \neq 0$
- Random forest: [**Huynh-Thu et al., 2010**] **GENIE3**
- Variational auto-encoder: [**Yu et al., 2019**, **Shu et al., 2021**] **DeepSEM** Details

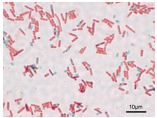
Method evaluation (DREAM4/5 challenges)



[Marbach et al., 2012]

- with known list of regulators, RF is often the best
- on higher organisms, no better than random guess

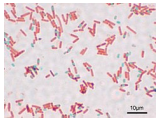
Insights into “methods / regulation mechanisms” relationships

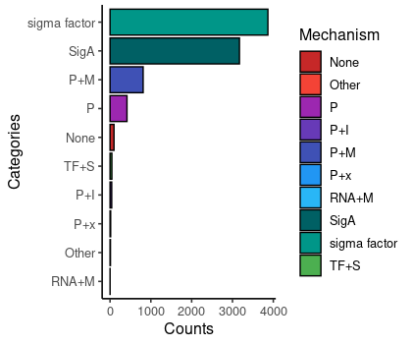



Bacillus subtilis:

- $p \simeq 3,900$ genes / $n = 269$ experiments **[Nicolas et al., 2012]**
- regulation network **[Faria et al., 2016]**

Insights into “methods / regulation mechanisms” relationships

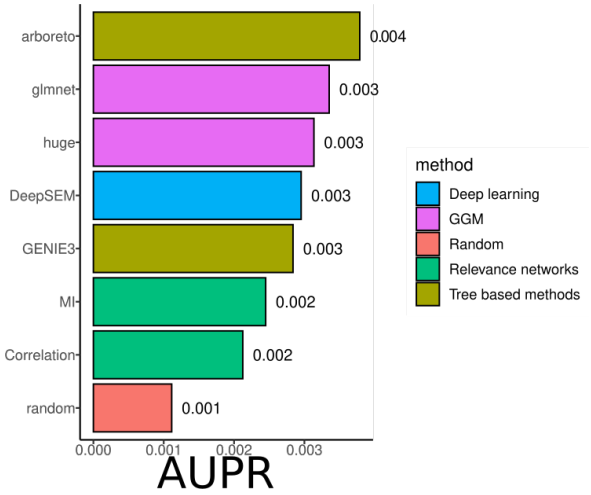


Bacillus subtilis:

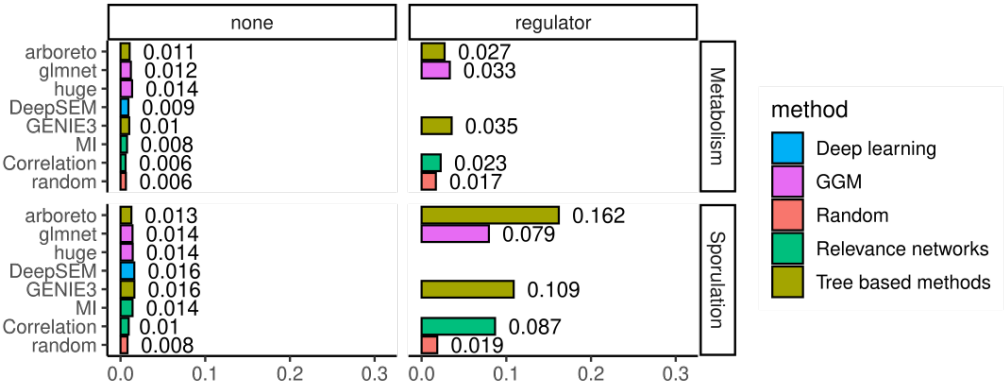
- $p \simeq 3,900$ genes / $n = 269$ experiments **[Nicolas et al., 2012]**
- regulation network **[Faria et al., 2016]**

- direct regulations: σ factors or P/TF (transcription factor)
- indirect regulations: involving something outside product of gene expression

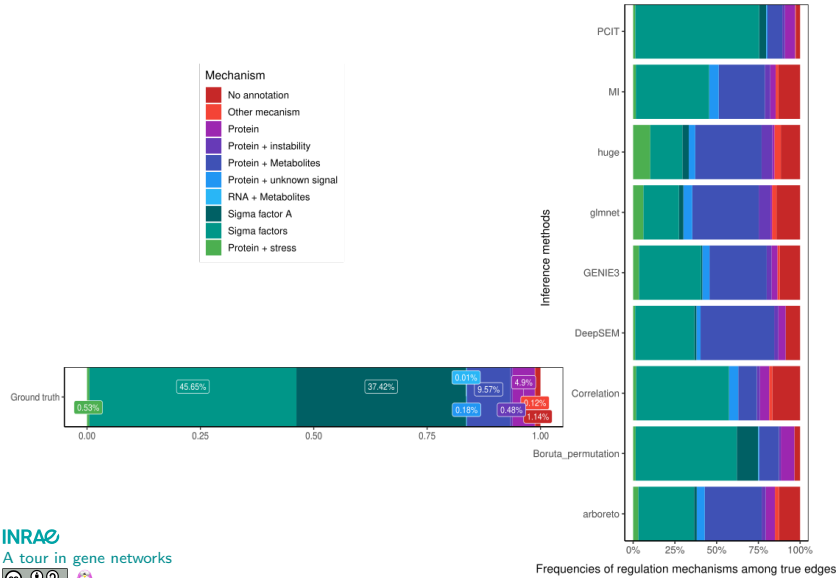
Insights into “methods / regulation mechanisms” relationships



Insights into “methods / regulation mechanisms” relationships



Insights into “methods / regulation mechanisms” relationships



Conclusion and possible ways to go...

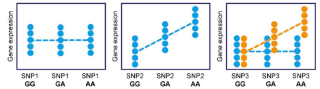
- When completely agnostic, network inference methods are not able to properly find gene-gene regulations ⇒ include other (experimental or knowledge) data **[Petralia et al., 2015, Cassan et al., 2023]**

- But, network clusters are biologically meaningfull

- Prediction beyond purely genetic is especially difficult ⇒ toward a more precise model of biological mechanisms? (hybrid statistical / deterministic models) **[Ventre et al., 2023]**

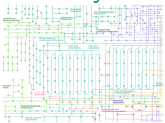
Thank you for your attention! More about omics?

Breeds/Ecotypes in GWAS



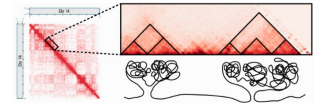
Jeong Hwan Ko "Données omiques" (10h20 today)

Pathway based metabolomics analysis



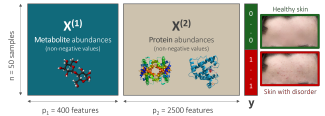
Camille Guilmineau "Multi-omique" (16h20 Thursday)

3D DNA structure



Élise Jorge "Données omiques" (11h05 today)

NMF for multi-omic integration



Aurélie Mercadié "Multi-omique" (15h35 Thursday)

References

Aibar, S., Gonzàlez-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z. K., Wouters, J., and Aerts, S. (2017).
SCENIC: single-cell regulatory network inference and clustering.
Nature Methods, 14:1083–1086.

Ambroise, C., Chiquet, J., and Matias, C. (2009).
Inferring sparse Gaussian graphical models with latent structure.
Electronic Journal of Statistics, 3:205–238.

Angelin-Bonnet, O., Biggs, P. J., Baldwin, S., Thomson, S., and Vignes, M. (2020).
sismonr: simulation of *in silico* multi-omic networks with adjustable ploidy and post-transcriptional regulation in R.
Bioinformatics, 36(9):2938–2940.

Barillot, E., Calzone, L., Hupé, P., Vert, J.-P., and Zinovyev, A. (2012).
Computational Systems Biology of Cancer.
CRC Press.

Brouard, C., Mourad, R., and Vialaneix, N. (2024).
Should we really use graph neural networks for transcriptomic prediction?
Briefings in Bioinformatics, 25(2):bbae027.

Butte, A. and Kohane, I. (1999).
Unsupervised knowledge discovery in medical databases using relevance networks.
In *Proceedings of the AMIA Symposium*, pages 711–715.

Butte, A. and Kohane, I. (2000).
Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.

In *Proceedings of the Pacific Symposium on Biocomputing*, pages 418–429.

Cassan, O., Lecellier, C.-H., Bréhélin, L., Martin, A., and Lèbre, S. (2023).
Integration of transcription factor binding sites to gene expression data improves regression-based gene regulatory network inference in Arabidopsis thaliana.
In preparation.

Chen, L., Liu, H., Kocher, J.-P. A., Li, H., and Chen, J. (2015).
glmgraph: an R package for variable selection and predictive modeling of structured genomic data.
Bioinformatics, 31(24):3991–3993.

Chereda, H., Bleckmann, A., Kramer, F., Leha, A., and Beissbarth, T. (2019).
Utilizing molecular network information via graph convolutional neural networks to predict metastatic event in breast cancer.
Studies in Health Technology and Informatics, 267:181–186.

Chereda, H., Bleckmann, A., Menck, K., Perera-Bel, J., Stegmaier, P., Auer, F., Kramer, F., Leha, A., and Beißbarth, T. (2021).
Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer.
Genome Medicine, 13:42.

Chiquet, J., Grandvalet, Y., and Ambroise, C. (2011).
Inferring multiple graphical structures.
Statistics and Computing, 21(4):537–553.

Chiquet, J., Mariadassou, M., and Robin, S. (2021).
The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances.
Frontiers in Ecology and Evolution, 9:188.

Chiquet, J., Mary-Huard, T., and Robin, S. (2016).

Structured regularization for conditional Gaussian graphical models.
Statistics and Computing, pages 789–804.

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016).
Convolutional neural networks on graphs with fast localized spectral filtering.
In Lee, D. D., von Luxburg, U., Garnett, R., Sugiyama, M., and Guyon, I., editors, *Advances in Neural Information Processing Systems (NIPS 2016)*, volume 29, pages 3844–3852, Red Hook, NY, USA. Curran Associates Inc.

Faria, J. P., Overbeek, R., Taylor, R. C., Conrad, N., Vonstein, V., Goelzer, A., Fromion, V., Rocha, M., Rocha, I., and Henry, C. S. (2016).
Reconstruction of the regulatory network for Bacillus subtilis and reconciliation with gene expression data.
Frontiers in Microbiology, 7:275.

Fey, M. and Lenssen, J. E. (2019).
Fast graph representation learning with pytorch geometric.
In *Proceedings of RLGM Workshop at ICLR 2019*.

Foissac, S., Djebali, S., Munyard, K., Vialaneix, N., Rau, A., Muret, K., Esquerre, D., Zytnicki, M., Derrien, T., Bardou, P., Blanc, F., Cabau, C., Crisci, E., Dhorne-Pollet, S., Drouet, F., Faraut, T., Gonzáles, I., Goubil, A., Lacroix-Lamande, S., Laurent, F., Marthey, S., Marti-Marimon, M., Mormal-Leisenring, R., Mompart, F., Quere, P., Robelin, D., SanCristobal, M., Tosser-Klopp, G., Vincent-Naulleau, S., Fabre, S., Pinard-Van der Laan, M.-H., Klopp, C., Tixier-Boichard, M., Acloque, H., Lagarrigue, S., and Giuffra, E. (2019).
Multi-species annotation of transcriptome and chromatin structure in domesticated animals.
BMC Biology, 17:108.

Friedman, J., Hastie, T., and Tibshirani, R. (2008).
Sparse inverse covariance estimation with the graphical lasso.
Biostatistics, 9(3):432–441.

Gallopin, M., Rau, A., and Jaffrézic, F. (2013).
A hierarchical Poisson log-normal model for network inference from RNA sequencing data.

PLoS ONE, 8(10).

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017).
Neural message passing for quantum chemistry.
In Precup, D. and The, Y. W., editors, *Proceedings of the 34 th International Conference on Machine Learning (ICML 2017)*, volume 70, pages 1263–1272, Sydney, Australia.

Grattarola, D. and Alippi, C. (2020).
Graph neural networks in TensorFlow and Keras with Spektral.
In *Proceedings of the Graph Representation Learning and Beyond – ICML 2020 Workshop*.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010).
Inferring regulatory networks from expression data using tree-based methods.
PLoS ONE, 5(9):e12776.

Li, C. and Li, H. (2008).
Network-constrained regularization and variable selection for analysis of genomic data.
Bioinformatics, 24(9):1175–1182.

Marbach, D., Costello, J. C., Küffner, R., Vega, N., Prill, R. J., Camacho, D. M., Allison, K. R., the DREAM5 Consortium, Kellis, M., and Collins, James J.and Stolovitsky, G. (2012).
Wisdom of crowds for robust gene network inference.
Nature Methods, 9(8):796–804.

McDermott, M. B., Wang, J., Zhao, W.-N., Sheridan, S. D., Szolovits, P., Kohane, I., Haggarty, S. J., and Perlis, R. H. (2020).
Deep learning benchmarks on L1000 gene expression data.
IEEE/ACM Transations on Computational Biology and Bioinformics, 17(6):1846–1857.

Meinshausen, N. and Bühlmann, P. (2006).

High dimensional graphs and variable selection with the Lasso.
Annals of Statistic, 34(3):1436–1462.

Mohan, K., Chung, J., Han, S., Witten, D., Lee, S., and Fazel, M. (2012).
Structured learning of Gaussian graphical models.
In *Proceedings of NIPS (Neural Information Processing Systems) 2012*, Lake Tahoe, Nevada, USA.

Nicolas, P., Mäder, U., Dervyn, E., Rochat, T., Leduc, A., Pigeonneau, N., Bidnenko, E., Marchadier, E., Hoebeke, M., Aymerich, S., Becher, D., Bisicchia, P., Botella, E., Delumeau, O., Doherty, G., Denham, E. L., Fogg, M. J., Fromion, V., Goelzer, A., Hansen, A., Härtig, E., Harwood, C. R., Homuth, G., Jarmer, H., Jules, M., Klipp, E., Le Chat, L., Lecointe, F., Lewis, P., Liebermeister, W., March, A., Mars, R. A., Nannapaneni, P., Noone, D., Pohl, S., Rinn, B., Rügheimer, F., Sappa, P. K., Samson, F., Schaffer, M., Schwikowski, B., Steil, L., Stülke, Wiegert, T., Devine, K. M., Wilkinson, Anthony J. ad van Dijl, J. M., Hecker, M., Völker, U., Bessières, P., and Noirot, P. (2012).
Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis.
Science, 335(6072):1103–1106.

Petralia, F., Wang, P., Yang, J., and Zhidong, T. (2015).
Integrative random forest for gene regulatory network inference.
Bioinformatics, 31(12):i197–i205.

Pramanik, D., Shelake, R. M., Kim, M. J., and Kim, J.-Y. (2021).
CRISPR-mediated engineering across the central dogma in plant biology for basic research and crop improvement.
Molecular Plant, 14(1):127–150.

Quach, H. and Quintana-Murci, L. (2017).
Living in an adaptive world: genomic dissection of the genus Homo and its immune response.
Journal of Experimental Medicine, 4(4):877–894.

Ramirez, R., Chiu, Y.-C., Hererra, A., Mostavi, M., Ramirez, J., Chen, Y., Huang, Y., and Jin, Y.-F. (2020).
Classification of cancer types using graph convolutional neural networks.

Frontiers in Physics, 8.

Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E., and Vert, J.-P. (2007).
Classification of microarray data using gene networks.
BMC Bioinformatics, 8:35.

Shu, H., Zhou, J., Lian, Q., Li, H., Zhao, D., Zeng, J., and Ma, J. (2021).
Modeling gene regulatory networks using neural network architectures.
Nature Computational Science, 1(7):491–501.

Ventre, E., Herbach, U., Espinasse, T., Benoit, G., and Gandrillon, O. (2023).
One model fits all: combining inference and simulation of gene regulatory networks.
PLOS Computational Biology, 19(3):e1010962.

Villa-Vialaneix, N., Vignes, M., Viguerie, N., and San Cristobal, M. (2014).
Inferring networks from multiple samples with concensus LASSO.
Quality Technology and Quantitative Management, 11(1):39–60.

Weinreb, C. and Raphael, B. J. (2016).
Identification of hierarchical chromatin domains.
Bioinformatics, 32(11):1601–1609.

Yang, A., Zhang, W., Wang, J., Yang, K., Han, Y., and Zhang, L. (2020).
Review on the application of machine learning algorithms in the sequence data mining of DNA.
Frontiers in Bioengineering and Biotechnology, 8.

Yu, Y., Chen, J., Gao, T., and Yu, M. (2019).
DAG-GNN: DAG structure learning with graph neural networks.

In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, volume 97 of *Proceedings of Machine Learning Research*, pages 7154–7163. PMLR.

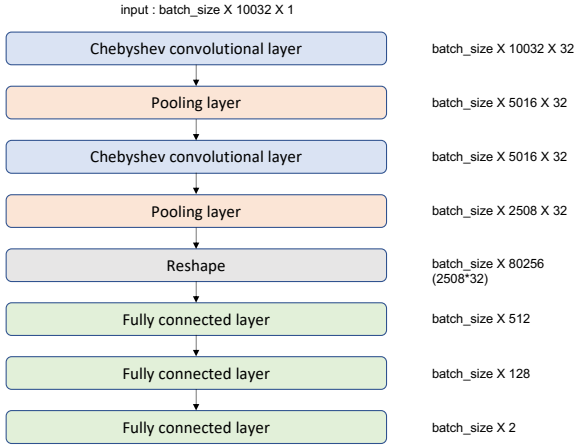
Credits

(unofficial) Beamer template made with the help of Thomas Schiex, Matthias Zytnicki and Andreea Dreau: https://forgemia.inra.fr/nathalie.villa-vialaneix/bainrae

- page 2: DNA image adapted from "double-stranded DNA" by MesserWoland, WikiMedia Commons and right arrow adapted from "Red short left arrow" from Ariel196, WikiMedia Commons
- page 6: image on expression regulation is from "Regulation of transcription in mammals" by Bernstein0275, Wikimedia Commons
- page 7: image from KEGG pathway is from https://www.genome.jp, image from Protein-Protein interaction (right) is from "RNAseInhibitor-RNAse complex" by Dcrjsr, Wikimedia Commons
- page 9: image of Breast Cancer is from "Breast cancer" by SMART-Servier Medical Art, part of Laboratoires Servier, Wikimedia Commons
- page 32: "microscopic image of the bacterial spore formation of Bacillus subtilis(ATCC 6633) Spore staining, magnification:1,000. (green) spores, (red) vegetatives" by Y tambe, WikiMedia Commons
- page 37: GWAS image from **[Quach and Quintana-Murci, 2017]** and 3D DNA image from **[Weinreb and Raphael, 2016]**

Architecture of the GCN used in Chereda et al., 2019, 2021

Back



Convolutional layer

Back

- **Chebyshev convolutional layer**:
  - Spectral convolution on graph based on Laplacian low band filtering

$$y = \sum_{k=0}^{K} \theta_k T_k \left(\frac{2L}{\lambda_{max}} - I \right) x$$

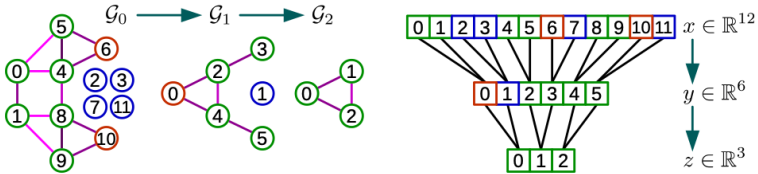
where T_k are Chebyshev polynomials .

- **Graph convolutional layer**:
  - linear formulation wrt L by considering the case where $K = 1$
  - approximation by restraining the number of parameters

$$y = \theta \tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} x, \quad \text{where } \tilde{A} = A + I$$

Graph coarsening [Defferrard et al., 2016]

Back



- Graclus algorithm: computes successive coarser versions of the graph
- clustering objective: normalized cut $W_{ij}\left(\frac{1}{d_i}+\frac{1}{d_j}\right)$
- Creation of a balanced binary tree: fake (disconnected) nodes are added to pair with singletons
- Vertices are then rearranged
  $\rightarrow$ pooling is analog to pooling a regular 1D signal

Implementation of the model of [Chereda et al., 2021] using Keras and Spektral

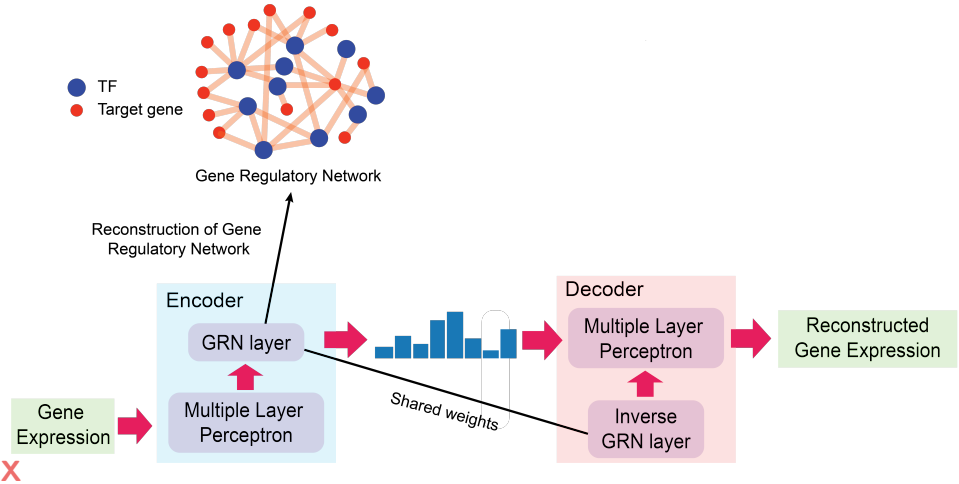
Back

- Layers:
  - Convolutional layers: Spektral (ChebConv)
  - Pooling layers: the coarsening from [Defferrard et al., 2016] is computed in the preprocessing and then a max pooling of size 2 is used.
  - Fully connected layers: Keras (dense) with ℓ_2 regularization
- For creating mini-batches data, we use the mixed data mode of Spektral (single graph and different node attributes)
- The GNN model had to be adapted to take into account the different coarsened graphs

DeepSEM

Back

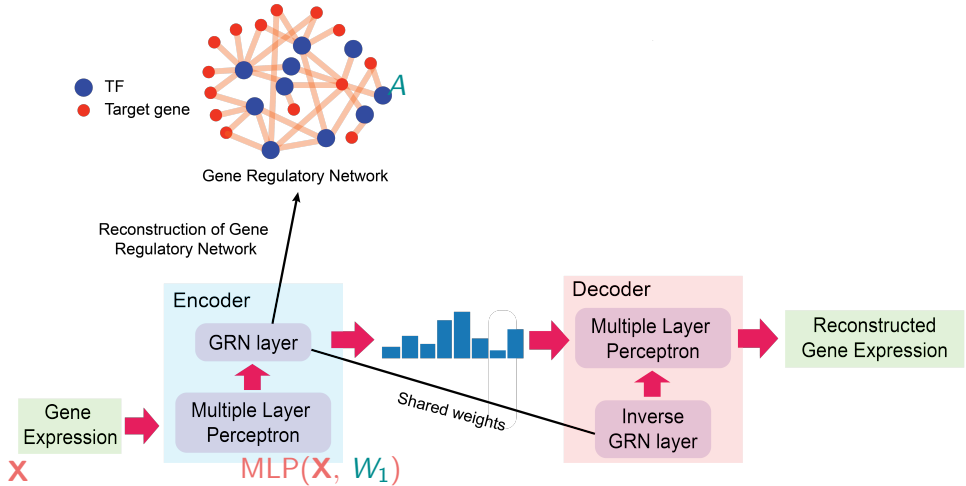
Variational auto-encoder: [Yu et al., 2019, Shu et al., 2021] DeepSEM



DeepSEM

Back

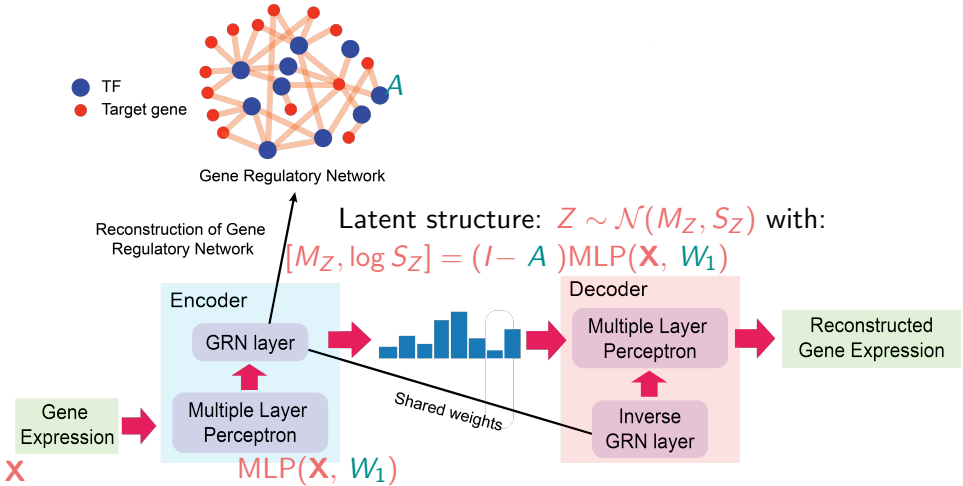
Variational auto-encoder: [Yu et al., 2019, Shu et al., 2021] DeepSEM



DeepSEM

Back

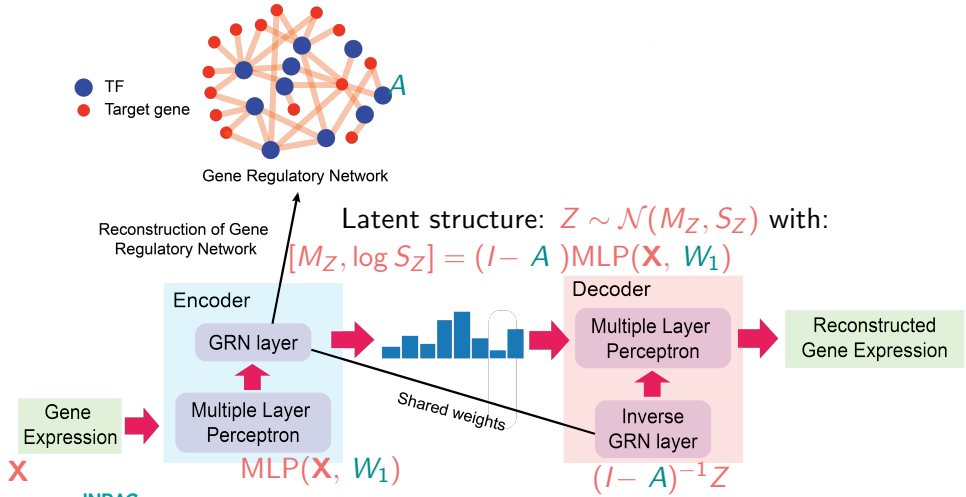
Variational auto-encoder: [Yu et al., 2019, Shu et al., 2021] DeepSEM



DeepSEM

Back

Variational auto-encoder: [Yu et al., 2019, Shu et al., 2021] DeepSEM



DeepSEM

Back

Variational auto-encoder: [Yu et al., 2019, Shu et al., 2021] DeepSEM