Prediction of gene expression using whole-genome epigenomic signals

Mathilde Bruguet[1] & Romane Leroy[2] & Santa Kirezi[3] & Romain Tavenard[4] & Magalie Houée-Bigot[5] & Gaël Le Trionnaire[6] & Nadia Ponts[7] & David Causeur[8]

[1] *Institut Agro, Rennes, France, mathilde.bruguet@agrocampus-ouest.fr*
[2] *Institut Agro, Rennes, France, romane.leroy@agrocampus-ouest.fr*
[3] *Institut Agro, Rennes, France, santa.kirezi@agrocampus-ouest.fr*
[4] *LETG - UMR CNRS 6554, France, romain.tavenard@univ-rennes2.fr*
[5] *Institut Agro, Rennes, France, magalie.houee@agrocampus-ouest.fr*
[6] *IGEPP - UMR INRAE 1349, France, gael.le-trionnaire@inrae.fr*
[7] *MycSA - UR INRAE 1264, France, nadia.ponts@inrae.fr*
[8] *IRMAR - UMR CNRS 6625, France, david.causeur@institut-agro.fr*

Résumé. Pour survivre et se développer, les agents phytopathogènes doivent s'adapter à une variété de stress environnementaux. A des horizons de temps courts, les variations épigénétiques peuvent permettre des réponses phénotypiques adaptatives en modifiant les réseaux d'expression génique, sans changement de séquence génomique. Ainsi, les génomes et les épigénomes interagissent avec l'environnement et contribuent à l'apparition de nouveaux phénotypes, dont l'adaptabilité est une caractéristique clé de la résilience. Si l'adaptation des espèces par variation génétique fait l'objet d'intenses recherches, les composantes épigénétiques de l'innovation phénotypique restent peu étudiées notamment du fait de la difficulté de s'affranchir de variations du fond génétique entre les générations. Dans ce contexte, les organismes se reproduisant de manière clonale sont d'excellents modèles pour étudier la contribution de l'épigénétique dans les processus adaptatifs. Parmi ceux-ci, le champignon filamenteux phytopathogène producteur de mycotoxines *Fusarium graminearum* est un bioagresseur de grandes cultures aux capacités de résilience remarquables.

Les variations d'accessibilité de la chromatine par repositionnement des nucléosomes sont des mécanismes épigénétiques clés qui modulent l'expression des gènes. Leur étude est possible grâce à des méthodes de séquençage à haut débit (MAINE-seq) générant des signaux complexes et hétérogènes dont la valorisation par les méthodes actuelles de la statistique génomique est limitée. L'apprentissage statistique des motifs d'association entre un signal épigénomique observé pour chaque gène sur une région large du génome incluant sa partie codante et sa région promotrice et l'expression de ce gène mesurée par séquençage de l'ARN (RNA-seq) offre des perspectives pour comprendre les capacités d'un organisme à s'adapter à un stress par des mécanismes épigénétiques. Or, les études comparatives menées sur *Fusarium graminearum* et portant sur une large gamme de méthodes d'apprentissage statistique, allant de celles basées sur des scores linéaires aux réseaux de neurones profonds en passant par les forêts aléatoires, conduisent à des résultats peu satisfaisants.

L'objectif de notre travail est de proposer une nouvelle approche dans laquelle l'expression des gènes est vue comme un signal le long de la séquence de nucléotides support de la partie codante du gène. Ce nouveau paradigme pour l'apprentissage statistique *function-to-function*

de données fonctionnelles massives par d'autres données fonctionnelles tire profit de supports communs des signaux épigénomiques et d'expression pour valoriser des corrélations spatiales induites par des mécanismes connus de régulation de l'expression par le niveau d'accessibilité de la chromatine autour du codon d'initiation de chaque gène. La présentation démontre l'intérêt de cette approche à la fois en termes de performance de prédiction mais aussi pour mieux comprendre les relations entre épigénome et transcriptome.

Mots-clés. Apprentissage statistique, Données fonctionnelles massives, Intégration de données -omiques, Régression pour données fonctionnelles.

Abstract. Plant pathogens have to adapt to a variety of environmental stresses in order to grow and survive. Epigenetic variations can drive short-term phenotypic responses to those environmental stresses by modifications of the gene regulatory network, without any changes of the genomic sequence. This interplay between genomes, epigenomes and environment contributes to the emergence of new phenotypes, whose adaptability ensures more resilience. Whereas genetic mutations involved in species adaptation are intensively studied, epigenetic determinants of phenotypic innovation are so far poorly investigated, a major pitfall being the difficulty to adjust for the variations of the genetic background between generations. In this context, organisms with clonal reproduction are convenient models to study the specific contribution of epigenetics to adaptation mechanisms. The filamentous fungus *Fusarium graminearum* is an example of a highly resilient clonal reproduction plant pathogen producing a mycotoxin responsible for heavy damages in crops.

Variations of chromatin accessibility by changes of nucleosome positions are key epigenetic mechanisms regulating gene expression. High-throughput sequencing technologies (MAINE-seq) are especially designed to measure those whole-genome epigenetic variations. They generate complex and heterogeneous signals for which statistical genomics does not provide any standard data analysis routine so far. Statistical learning of the patterns of association between the epigenomic signal observed for each gene within a large genomic region covering the promoter and coding regions and gene expression measured using RNA-sequencing technologies (RNA-seq) offers some perspectives to understand epigenetic drivers of adaptability. However, comparative studies conducted on *Fusarium graminearum* and including a large panel of statistical learning methods, based on linear scores, aggregations of tree-based predictions or neural networks, show limited prediction performance.

We propose a new approach in which gene expression is viewed as a continuous signal over the coding region of a gene. This new *function-to-function* paradigm for statistical learning of massive functional responses by functional predicting variables takes advantage of common spatial supports for epigenomic and gene expression signals to leverage spatial correlations induced by proven transcriptional regulation mechanisms by chromatin accessibility in the neighborhood of the start codon of a gene. The extent to which this approach improves prediction performance is discussed in the presentation and prospective leads to better understand epigenetic regulation of transcription are deduced.

Keywords. Massive functional data, Statistical learning, Multi-omic data integration, Regression for functional data.

1 Introduction

Transcriptomic variations induced by many kinds of environmental stresses, especially heat stress, have been studied for a large variety of organisms. Identification of genes or more generally regulation pathways involved in those stress-induced variations generally results from so-called differential analyses consisting in whole-genome statistical tests for the comparison of mean expressions between contrasted experimental conditions. For example, such studies demonstrate that approximately 43% of genes in the filamentous fungus *Fusarium granimearum*, a clonal reproduction plant pathogen showing a high resilience to heat stresses (see Clairet *et al.*, 2023) have a modified gene expression level when exposed to a strong heat stress (15 minutes long exposition to 37 degrees Celsius). Moreover, complementary studies introducing a whole-genome mapping of epigenetic marks highlight correlations between gene expression levels and the presence or not of such marks on the genomic sequence of nucleotides within the neighborhood of those genes. Deciphering patterns of association between transcriptomic variations and epigenetic mechanisms involving modifications of the chromatin structure is yet a current challenge for bioinformatics and statistical genomics.

Availability of binding sites to transcription factors is notoriously affected by the positioning of nucleosomes along the sequence of nucleotide pairs in a broad region covering the promoter, coding and terminator regions of genes. How nucleosomes are positioned along the chromatin therefore plays a fundamental role in the regulation of transcription and subsequently gene expression. Nucleosome occupancy can be measured by MNase-assisted isolation of nucleosomes sequencing (MAINE-seq) technology, mapping regions of the genomes protected by nucleosomes. The former high-throughput technology generates signals providing a numeric measurement of chromatin accessibility at every base pair (bp) along the genome. In most studies involving such MAINE-seq data, raw signals of chromatin accessibility are reduced to summary statistics giving a general overview of the nucleosome occupancy over specific regions of interest in the genome. However, the dynamics of the transcription machinery suggests that numbers, amplitudes and positions of peaks and throughs in signals of chromatin accessibility should not be ignored, to account for a spatial correspondence between intervals of base pairs where the chromatin is accessible and some delayed transcriptional activity within the coding region of genes.

Identifying the specific patterns in curves of chromatin accessibility responsible for the regulation of gene expression is a key step to understand the epigenetic mechanisms driving phenotypes for improved adaptation of organisms to environmental stresses. The focus of our study (supported by the national research program Digit-Bio, INRAE, see https://digitbio-ia.github.io/) is on *F. granimearum*, whose epigenetic drivers are poorly known. Our goal is to investigate this question using statistical learning methods to establish association rules between patterns of MAINE-seq epigenomic data and RNA-seq measurements of gene expressions.

2 Data preparation and quality control

For each of the 14145 genes of *Fusarium granimearum*, three replicates of epigenomic signals are available on a region starting from 800 bp before the start codon to 800 bp after the stop codon. Those fixed values of 800 bp in the promoter and terminator regions of each gene has been chosen as a compromise between the necessity of leveraging epigenomic information outside of the coding region and the risk of interference with signals of too close genes, the genome of *Fusarium granimearum* being very dense. A data quality control procedure is first designed based on those three replicates, leading to the exclusion of 3350 genes, either showing inconsistencies across replicates, or with abnormal intervals of zeroes, or whose distance to other genes is smaller than 400 bp. Finally, a curve of log-transformed chromatin accessibility indices is obtained after averaging over the three replicates for the 10795 remaining genes. For each gene, three replicates of RNA-seq read counts within the coding region are also available. Gene expressions are log-transformed averages of those three replicates. The logarithm transformation applied on the raw chromatin accessibility indices and on the RNA-seq read counts is motivated by a strong over-dispersion of the data.

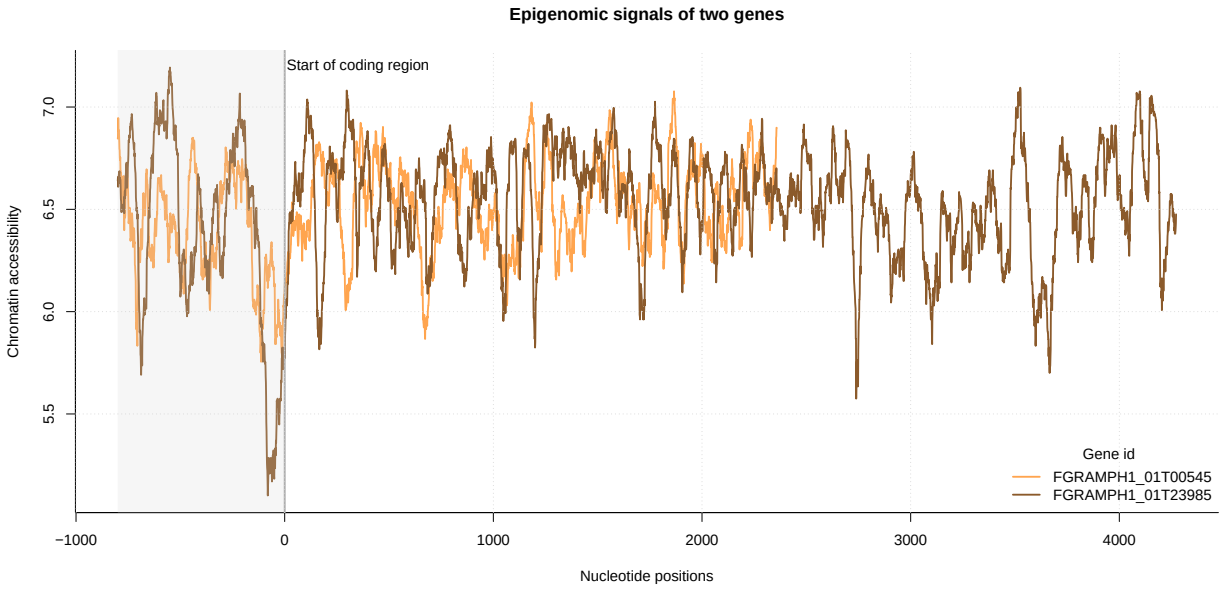

Figure 1: Epigenomic signals for two genes (`FGRAMPH1_01T00545` and `FGRAMPH1_01T23985`). The coding region of each gene starts at zero (plain vertical line). The promoter region is coloured in grey.

For illustration, Figure 1 displays the resulting epigenomic signals for two genes, whose identifiers are `FGRAMPH1_01T00545` and `FGRAMPH1_01T23985`, starting 800 bp before the start codon. First, note that the lengths of those genes and consequently of their epigenomic signals are different (1559 and 3473 bp), which raises a first major issue for designing statistical learning procedures based on those curves. More generally, gene lengths over the genome of

Fusarium graminearum are highly variable, as shown by the histogram displayed in Figure 2.

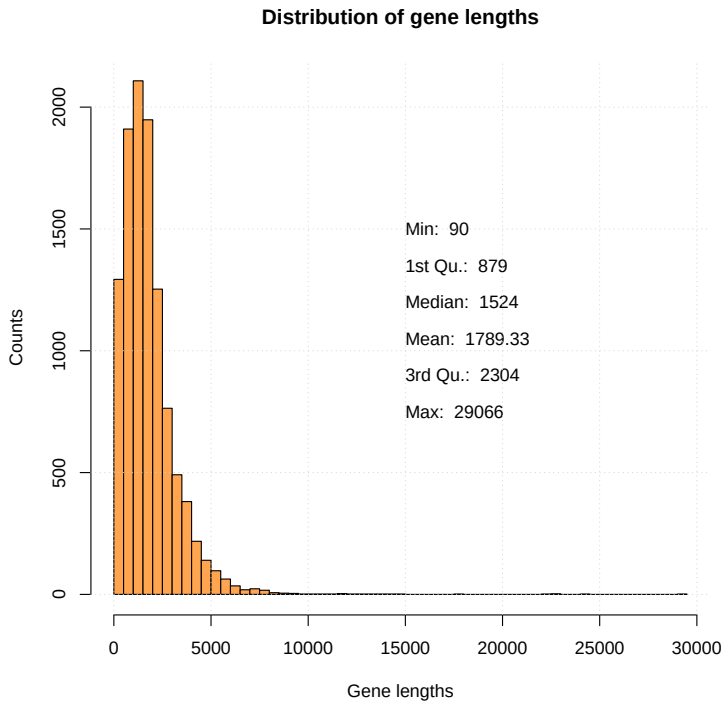

Figure 2: Histogram of gene lengths over the genome of *Fusarium graminearum*.

Many options have been considered to address this issue. First, reducing all signals to the same list of counts for histone modifications within bins of base pairs (see Singh *et al.*, 2016) or of geometrical summary statistics, such as latencies and amplitudes of peaks, or areas under the curve over some targeted regions, leads to a common definition of all explanatory variables for each gene. Unfortunately, how to draw this list of summary statistics is not guided by any biological consideration, which exposes to the risk of missing important patterns of association with gene expression. A second option preserving the potential prediction ability of whole epigenomic curves is to align all signals within the coding region, by taking a common grid of points for all genes, regularly spaced by the same fraction of the whole gene lengths. This option turns out to reduce one-to-one correlations between chromatin accessibility indices and gene expressions, raising suspicion about the relevance of such a transformation of epigenomic signals. The third option consists in restricting the study of epigenomic signals over an interval centered on the start codon, for example [-800 bp ; 800 bp] for all genes whose coding regions is at least longer than 800 bp. Indeed, as shown by the curve of correlations between chromatin accessibility indices and gene expressions displayed in Figure 3, the largest correlations are observed around the start codon of genes. More precisely, the negative peak of correlations at the end of the promoter region confirms that chromatin accessibility within this particular region favors the transcriptional activity.

In the presentation, we will focus on this last option, leading to a training dataset of joint

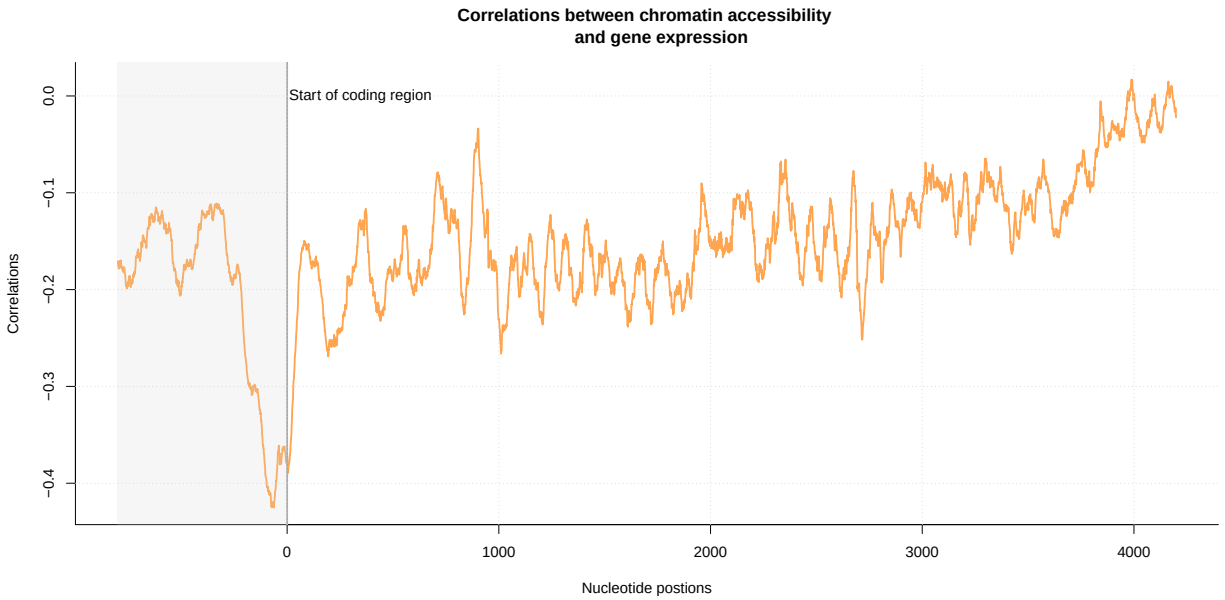

Figure 3: Correlations between chromatin accessibility indices at all nucleotide positions (limited to 4200 bp after the start codon, only 5% of genes being longer than 4200 bp) and gene expressions. The coding region of each gene starts at zero (plain vertical line). The promoter region is coloured in grey.

observations of the RNA-seq measurement of gene expression and discretized observations of the corresponding epigenomic signal on a grid of 1600 nucleotide positions around the start codon of each of 8440 genes. Note also that individuals in the present statistical learning issue are genes, which questions the usual independence assumption required to guarantee good properties of most standard estimation procedures. Indeed, gene expressions are notoriously driven by a regulatory network inducing a graph-structured stochastic dependence pattern across genes. This point will be discussed as prospective directions for improved prediction.

3 Statistical learning of gene expression based on epigenomic signals

In most prediction issues, the ideal objective in terms of prediction performance is to reach a cross-validated mean-squared error of prediction (MSEP) as close as possible to zero. In the present context, the ideal MSEP is unknown. It is indeed reasonable to consider that perfect prediction of gene expression from epigenomic signals exclusively would be a biological nonsense. Integration of additional -omic data, especially Hi-C data describing the three-dimensional conformation of chromatin, will be discussed as promising perspectives in the presentation.

A variety of statistical learning methods have been implemented to set up prediction

rules of gene expression from either raw epigenomic signals or reduced signals using B-spline coefficients in spline approximations with a large scope of dimensions for the basis:

- Penalized regression methods (Lasso, Ridge, Adaptive lasso and Elastic net with a fixed combination of the ℓ_1 (0.9) and ℓ_2 (0.1) norm of coefficients in the penalty term), where the penalty parameter minimizes a 10-fold cross-validated MSEP;
- Partial Least Squares (PLS) regression, where the number of PLS components minimizes a 10-fold cross-validated MSEP;
- Random forests (RF), where the number of variables to randomly sample at each split and the number of trees minimize the MSEP over a grid in a 10-fold cross-validation setup;
- Support Vector Regression (SVR);

The prediction performance of the state-of-the-art machine learning methods listed above and the CNN are reproduced in Table 1.

	Ridge	Lasso	Adaptive Lasso	Elastic net	PLS	RF	SVR
RMSEP	3.69	3.60	3.57	3.58	3.62	2.02	1.80

Table 1: Root Mean Squared Error of Prediction (RMSEP) of state-of-the-art machine learning methods for the prediction of gene expression using [-800 bb ; 800 bp] epigenomic signals. RMSEP is calculated in a 80% train-20% test cross-validation setup.

It turns out that the best prediction performance is reached by the random forests and the Support Vector Regression method, implemented either with the raw epigenomic signals or equivalently using a spline approximation with a large number of B-spline functions. Indeed, using a spline approximation does not improve prediction but speed up calculations. Yet, the former methods show a limited prediction performance.

4 The function-to-function prediction approach

One of the leads to improve the prediction performance of gene expression using epigenomic signals is motivated by the observation that the RNA-seq gene expression measurement is a total reads count over the coding region of genes, which can be viewed as a summary statistic masking a variety of dynamics of the transcriptional activity. For example, the two genes whose epigenomic signals are displayed in Figure 1 have indeed different epigenomic signals, especially at the end of the promoter region where the amplitude of the negative peaks identified as a possible marker of a large expression are different, but they have the same gene expression measurement. However, the locations within the coding region where the reads map exons of genes are distributed differently, as shown by their curves of cumulative

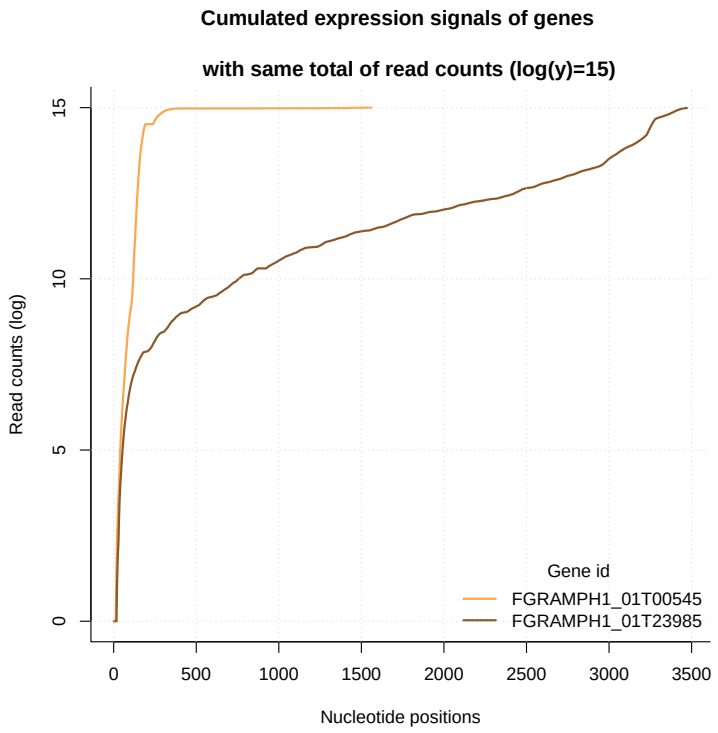

Figure 4: Cumulative numbers of reads within the coding region for the two genes whose epigenomic signals are displayed in Figure 1. The maximum value of each curve is reached before the stop codon and equals the total RNA-seq reads count.

log-transformed read counts, starting from the start codon up to the stop codon (see Figure 4).

In order to take advantage of a common spatial support of both epigenomic and transcriptional information, we propose to measure gene expression by the curve of cumulative reads count starting from the start codon and ending at 800 bp. Moreover, it turns out that the cumulative gene expression curve over this interval [1;800 bp] within the coding region shows an excellent ability to predict the total read counts. Indeed, Figure 5 demonstrates the accuracy of prediction of the total read counts by the cumulative gene expression curve over [1;800 bp], using PLS regression, with a 10-fold cross-validated correlation between observed and predicted value of 0.98.

The method we propose is two-step:

- Step 1: set up a prediction rule for the curve of cumulative gene expression in [1;800 bp] using epigenomic signals;
- Step 2: use PLS regression, whose performance is shown in Figure 5, to deduce the predicted total RNA-seq reads count.

The first step of the above method consists in predicting a curve, describing the cumulative

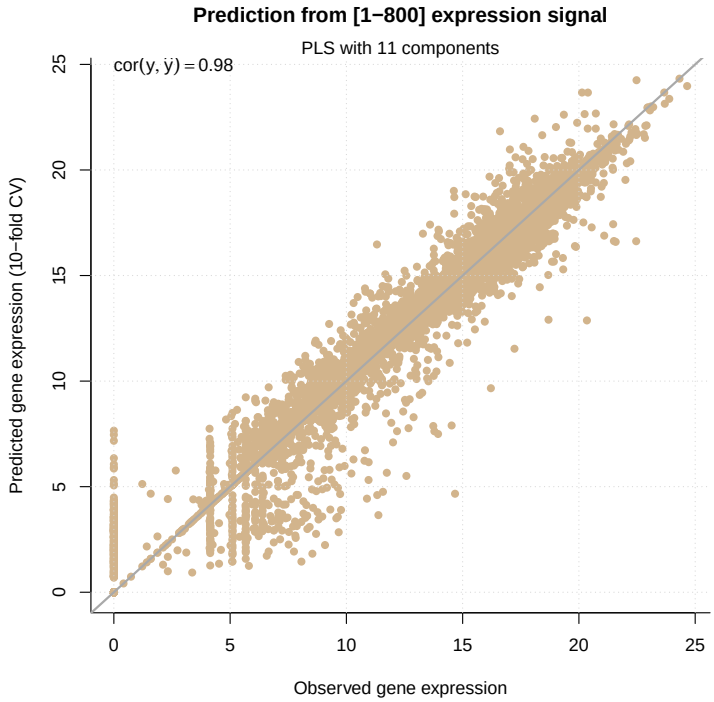

Figure 5: Performance of PLS regression in the prediction of the total RNA-seq reads count by the curve of cumulative gene expression in the interval [1;800 bp] starting at the start codon.

gene expression in [1;800 bp] by another curve, the epigenomic signal. For this *function-to-function* prediction task, two methods show good performance: random forests, where the loss function is the mean of squared differences between discretized observations and predictions of the curve of cumulative gene expressions, and a multi-head self-attention neural network. The added-value of using neural networks was previously mentioned in Singh *et al.* (2016), reporting a study in which a two-group expression level (high or low) is predicted by the counts for five pre-chosen histone modifications within adjacent 100 bp-bins around the start codon of each gene using Convolution Neural Networks (CNNs). Introducing self-attention layers in the neural network is inspired by the desirable properties of such mechanisms in language models, where the prediction of sequences by other sequences is similar to the present function-to-function prediction task.

The presentation will discuss the implementation of the above two-step method and show the improvements with respect to the direct prediction of the total RNA-seq reads count using the epigenomic signals. Promising perspectives for the current work will also be proposed, including integration of additional -omic data for a more complete description of the association between epigenome and transcriptome.

Bibliography

Clairet, C., Lapalu, N., Simon, A., Soyer, J.L., Viaud, M., Zehraoui, E., Dalmais, B., Fudal, I. and Ponts, N. (2023) Nucleosome patterns in four plant pathogenic fungi with contrasted genome structures, *Peer Community Journal*, 3: e13.

Singh, R., Lanchantin, J., Robins, G., Qi, Y. (2016) DeepChrome: deep-learning for predicting gene expression from histone modifications, *Bioinformatics*, Volume 32, Issue 17, September 2016, Pages 639–648, https://doi.org/10.1093/bioinformatics/btw427