Unveiling Gene Expression Histonic Regulative Patterns by Hyperplanes Clustering

Fabrizio Frasca$^{(1)}$, Matteo Matteucci$^{(1)}$, Marco Morelli$^{(2)}$ and Marco Masseroli$^{(1)}$

(1) Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano, Piazza Leonardo Da Vinci 32, 20133 Milan, Italy
[fabrizio.frasca@mail., matteo.matteucci@, marco.masseroli@]polimi.it

(2) Center for Genomic Science of IIT@SEMM
Istituto Italiano di Tecnologia (IIT), 20139 Milan, Italy
morelli.marco@hsr.it

Keywords: Gene expression, Epigenetic transcriptional regulation, K-plane regression.

Abstract. In targeted cancer therapy, great relevance is assumed by data-driven investigations on the fundamental mechanisms by which epigenetic modifications cooperate to regulate the transcriptional status of genes. At the high resolution level of genome-wide studies, only general, mean regulative motifs are drawn, with possible multi-functional co-regulative roles remaining concealed. In order to retrieve sharper and more reliable regulative patterns, in this work we propose the application of K-plane regression to partition the set of protein coding genes into clusters with shared regulative mechanisms. Completely data-driven, the approach has computed clusters of genes significantly better fitted by specific linear models than by single regression, and characterized by distinct histonic input patterns and mean measured expression values.

1 Scientific Background

Understanding the fundamental mechanisms by which histone marks (HM) and transcription factors (TF) operate to regulate the expression of specific genes is of great interest. Studies revealing the main role played in cancer etiology by gene expression alterations from epigenetic aberrations [1] have recently paved the way to the promising field of targeted cancer therapy, where epigenetic approaches are used to treat cancers in a personalized manner, by kick-starting particular immune responses or bringing back the gene expression levels to the expected ones.

Leveraging the large amount of publicly available high-throughput sequencing data, statistical models have been conceived to study the association between gene-related epigenetic signals and messenger RNA (mRNA) abundance at a genome-wide scale [2], with the problem usually framed as a regression task. Genes are *samples*, signals from HMs and/or TFs are *input features* and the aim is to predict the *response value*, i.e., mRNA abundance quantifications.

At a genome-wide level, HMs and TFs have been shown to be predictive for mRNA abundance [2], but also to exhibit certain *statistical redundancy* within themselves, with few works trying to break this last in a *data-driven* manner. Avoiding the inclusion of biological prior knowledge in statistical modeling is a relevant concern in the context of targeted cancer therapy and personalized medicine: possibly uncharted epigenetic aberrations and anomalies in their regulative effects represent the main objects of analyses.

A notable data-driven attempt has been made in [3], where a mixture of Bayesian linear elastic nets revealed to better fit transcriptional regulation w.r.t. a *single* regression model and to expose distinct predictive relevance of the epigenetic features. Though the models accounting to the mixture in [3] are distinctly defined, genes in the dataset are, however, only *softly* clustered, as the expression for a gene is the weighted sum of the outputs of *all* models.

As this *soft* approach renders interpretative analyses trickier, in this work we inquire into the possibility of performing a *hard* partitioning of the whole gene set in a data-driven manner, defining clusters where specific linear regression models are fit to learn the regulative dynamics of those gene sub-groups. In such a setting, a one-to-one association between linear models and gene clusters follows, and interpretative analyses are supported at best: regulative patterns can be investigated both at a gene-specific level and, statistically, at a gene-cluster level, and the regulative behavior can be matched with the most represented biological processes within a group.

Considered our problem to learn different linear models in a scenario where dynamics are likely to be overlapped, discontinuous, and partially lying on sub-dimensional manifolds, a suited tool is represented by *K-plane regression* [4] rather than *piece-wise linear affine model fitting* methods, such as that proposed in [5]. Despite its name, this method is based on a clustering approach; it finds a fixed number of (K) hyperplanes in order to have each point in the training set close to one of the hyperplanes, and all points in a partition as closest as possible in the input feature space. Given the capability of K-plane regression to tackle discontinuous functions and the more flexibility offered by a clustering approach, we built upon this last work to solve our problem.

2 Materials and Methods

The aim of this work is modeling epigenetic transcriptional regulation by means of a *hard* ensemble of linear regression models, each explaining mRNA abundance as a function of epigenetic signals for a specific gene sub-group, i.e., a cluster of genes.

All considered measurements are over the K562 immortalized cell line (human blood tissue), and only involve protein coding genes. GENCODE v10 reference annotation for the hg19 assembly was used to retrieve their transcription start sites (TSSs). The Roadmap Epigenomics Mapping Consortium's (REMC) repository was chosen as the only data source in this work.

Genes are *epigenetically* characterized by data in the form of processed ChIP-seq called peaks only for the $m = 12$ histone modifications assayed over K562 in REMC (no TF was accounted for). The epigenetic status of the generic gene g is, numerically, an m-dimensional *input vector* $\boldsymbol{x}_g$. Its elements summarize, each, the g-related status of a specific monitored HM, as the maximum peak enrichment value attained within a symmetric window region of 10 kbases centered on g's TSS. In accordance to [2], signals closer to the genes' TSSs (roughly, within promoters) are, indeed, the most valuable for the prediction of gene expression. Considered together for all our $n = 19,794$ genes, such vectors form *input matrix* X, with dimensions $n \times m$.

As for the *transcriptional* characterization of genes, we consider mRNA quantifications, measured by means of RNA-sequencing. The transcriptional status of gene g is encoded by $t_g = \sqrt{\ln(1+\tau_g)}$, where τ_g is the original mRNA quantification, and the application of two sub-linear, monotonically increasing functions aims at reducing the heteroskedasticity in regression residuals. Finally, consider t_g as the $(g+1)$-th element in n-dimensional target vector T collecting the transcriptional statuses of all the genes.

Together, X and T form our dataset $D = \langle X, T \rangle$, which is going to be partitioned by K-plane regression algorithm that, in our work, is designed to minimize the following objective function:

$$E(\Theta) = \sum_{k=0}^{K-1} \sum_{i \in \Theta(k)} (t_i - \tilde{\boldsymbol{w}}_k^T \tilde{\boldsymbol{x}}_i)^2 \tag{1}$$

where K is a pre-defined number of clusters, Θ defines the partitioning over the dataset ($\Theta(k)$ is the set of samples in Cluster k), $\tilde{\boldsymbol{x}}_i$ and t_i are, respectively, the input feature and target value for sample $i \in \Theta(k)$, and $\boldsymbol{w}_k$ is the weight vector of the least square solution for those points (the 'tilde' indicates inclusion of the bias term in the regression).

Here, we resort to multiple re-initializations to tackle possible sub-optimality, and

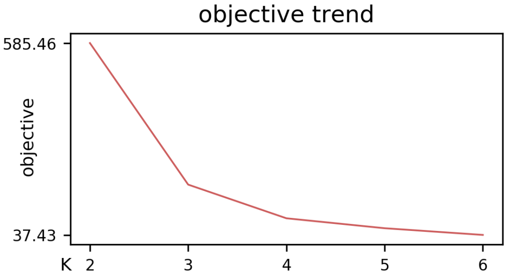

Figure 1: Values of best solutions (objective) from re-initialized K-plane regression as a function of number of clusters (K).

drop the additive Euclidean ‘closeness term’ added in [4] to force feature space contiguity of sample partitions - an inadequate pre-assumption in our domain. For each of the R runs the procedure is called, it starts by a random partition and optimizes Equation 1 by iteratively alternating a Maximization step - hyperplanes to clusters fitting - and an Expectation one - gene-cluster reassignments. Initializations are designed to construct a completely random partitioning made up of equally sized clusters. In the end, among the R yielded solutions, the one attaining best objective value is returned.

3 Results

Parameter K has been made to range in $[2 \dots 6]$, as the best value is indeed hardly guessable *a priori* and might depend on the nature of the specific problem. Better solutions, in terms of cost functions, have been observed for larger values of K: Figure 1 depicts the trend of the convergence objective value as a function of this parameter. In the following, results for the setting $K = 4$ are discussed; this mild setting is less prone to overfit spurious correlations, still yielding a good value of the objective function. It represents a trade-off between goodness of fit and, in light of the current knowledge about HM (co-)activity, reasonable biological interpretability.

Let $\Theta = \{\vartheta_0, \dots, \vartheta_{K-1}\}$ be our obtained solution, with $K = 4$ and ϑ_k representing the $(k+1)$-th cluster of genes computed by the algorithm. In correspondence with this partitioning, an ensemble of *cluster-wise* linear models can be considered as $M = \{\mu_0, \dots \mu_{K-1}\}$, where μ_k represents the hyperplane being the least square solution over genes in ϑ_k. Our solution Θ is contrasted against $\Theta_{gw} = \{\vartheta_{gw}\}, \vartheta_{gw} = \{0, 1, \dots, n-1\}$, the degenerate partitioning made up of a single cluster indexing the whole dataset D. This solution corresponds to setting $K = 1$, that is, to the use of a single linear model fitted over the entire dataset D. In the following, such a model is referred to as the *genome-wide* one and is labeled as μ_{gw}.

3.1 *Enhanced (Cluster-wise) Fitting*

Our K-Plane Regression managed to cluster genes with common regulative behaviors, as the obtained model ensemble effectively enhanced data fitting. Not only the objective value associated with Θ_{gw} is way larger than that associated with our solution Θ (3892.08 vs. 339.83), but, also, fitting is better at the level of *all* the computed clusters. The regression scores computed specifically over clusters in Θ, for both cluster-wise and genome-wide models, are reported in Table 1 in terms of residual sum of squares (RSS) and coefficients of determination (R^2); the row i of the table comprises scores for models μ_i and μ_{gw} over Cluster ϑ_i - subscripts ‘$_{cw}$’ and ‘$_{gw}$’, respectively.

In Table 1, the effectiveness of the proposed approach is confirmed by the fact that clusters not only are always better fitted by cluster-wise models than by μ_{gw}, but, also, that the specific linear models are acceptable, if not very good, in explaining the epigenetic transcriptional regulation of a large part of genes (refer to column “R$^2_{cw}$”). In 3 clusters out of 4 the R^2 scores from μ_{gw} are negative, implying the fitting, over the genes of each of those clusters, is worse than the constant mean model.

cluster (cardinality)	RSS$_{cw}$	RSS$_{gw}$	R$^2_{cw}$	R$^2_{gw}$
$0(2,717)$	79.22	1393.00	0.80	-2.50
$1(7,547)$	82.05	1514.44	0.54	-7.57
$2(5,045)$	85.64	714.70	0.84	-0.37
$3(4,485)$	92.90	269.92	0.92	0.76

Table 1: Cluster specific figures of merit. For cluster k, RSS$_{cw}$ and RSS$_{gw}$ are the residual sum of squares of, μ_k, μ_{gw} over ϑ_k, while R$^2_{cw}$ and R$^2_{gw}$ refer to the coefficients of determination of μ_k, μ_{gw} over ϑ_k.

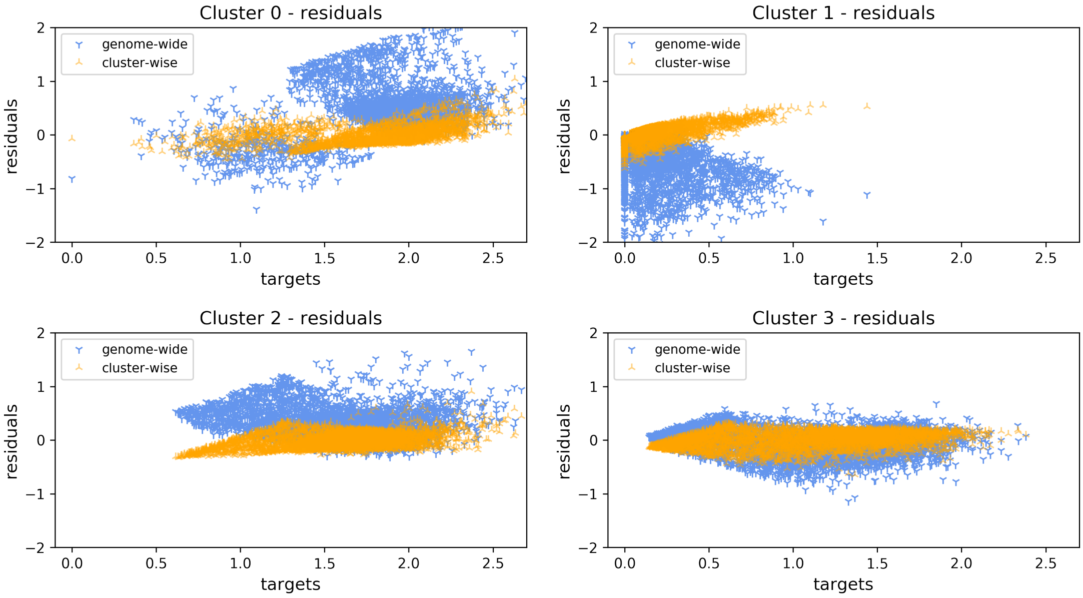

Figure 2: Cluster specific residuals from cluster-wise (orange) and genome-wide (blue) models.

The intuition that μ_{gw} is likely to only capture the regulative mechanisms of genes with “intermediate” regulative behaviour, such as those in Cluster 3, is supported by what observed in Figure 2, where cluster specific residuals ($\vec{y}$-axis) from cluster-wise and genome-wide models are plotted against target values ($\vec{x}$-axis).

Residuals from the genome-wide model are generally more disperse and heteroskedastic except for Cluster 3 - the only where μ_{gw} attains positive R^2 - where they are similar to those from the cluster-wise model μ_3, very well fitting the comprised genes. The overall R^2 of 0.66 attained by μ_{gw} on the whole dataset D is, consequently, an intermediate value resulting from considering together mildly modeled genes (Cluster 3) with the remaining ones, where the genome-wide model seems to be rather inadequate.

Hard hyperplanes clustering has revealed the criticality of single genome-wide regression by exposing subsets of genes under-fitted by μ_{gw}. In a real setting such that of targeted cancer therapy, unacceptable is to reasonably fit only 23% of protein coding genes (Cluster 3), as conceptually wrong conclusions might be drawn about the epigenetic regulative behaviors of the remaining ones.

3.2 *Cluster Characterization*

The effectiveness of hyperplanes clustering also emerges by observing how the obtained clusters are distinct in terms of the input patterns and mean expression value for the genes they contain. In this sub-section we leverage the enhanced interpretability coming from a *hard* gene partitioning to characterize the computed clusters.

For the generic cluster-wise model μ_i, let $\tilde{\boldsymbol{w}}_i$ be its weight vector, comprising the learnt intercept and regression coefficients ($m+1$ elements). For gene g in ϑ_i, let $\boldsymbol{\psi}_g$ be

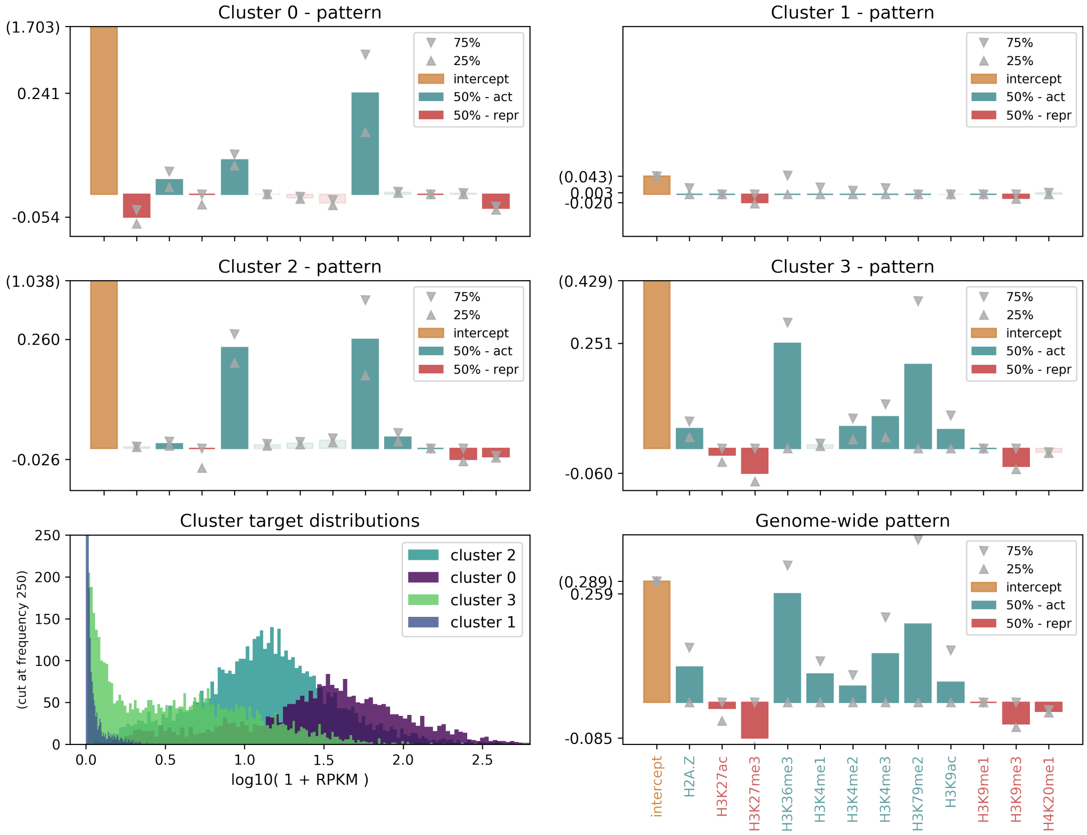

Figure 3: Bottom row: cluster specific target distribution (left) and genome-wide input pattern (right). Rows 1 through 3: cluster specific input patterns; the HMs on the $\vec{x}$-axis are in the same order of the genome-wide case; in green *computed activators* (positive weight), in red *computed repressors* (negative weight).

its *weighted input vector*, obtained by an element-wise multiplication between its input vector $\tilde{\boldsymbol{x}}_g$ and $\tilde{\boldsymbol{w}}_i$ - input vector is 1-edged to account for bias. Weighted input vectors are an effective means to quickly grasp the responsibilities of single features in determining the predicted response value, as $y_g = \sum_{j=0}^{m}(\boldsymbol{\psi}_{g_j})$. Weighted input vectors for *all* the genes in a cluster generate feature-wise boxplots that can be used to investigate the frequency distributions of cluster specific histone contributions and their associated dispersion, tracking the features which vary the most and those which, on the contrary, are more constant within the cluster.

Figure 3 depicts the patterns obtained by considering the feature-wise medians of the weighted input vectors, specifically for each cluster, along with the 25-th and 75-th percentiles of their distributions. Cluster specific target distributions and the ϑ_{gw}-associated pattern are also reported. In the patterns, intercepts are in orange, whilst HMs are green if associated with positive regression weight and red otherwise, with semi-transparent rendering for weights not passing a statistical F-test with significance $\alpha = 0.01$. In this way simpler and more robust patterns are provided, as fictitious correlations are pruned.

Cluster 1 is the most populated one and has a quite clear characterization. It comprises genes with a feeble histonic activity and usual null expression: this suggests the comprised genes are likely to never be activated during a cell life, and to be repressed at a chromatin level, e.g., embryonic genes. The flat-like input and the low intercept are consistent with the related expression distribution (0.0 RPKM median value). In such a scenario, a lower signal-to-noise-ratio is the probable cause of the mild attained R^2 score in this cluster (see Table 1).

Clusters 2 and 0 comprise active genes, with the highest expressed ones in the latter cluster (RPKM medians 12.73 and 32.23). This characterization is confirmed by high intercepts and the predominant roles assumed by activator H3K79me2, and H3K36me3 specifically in Cluster 2. Their large variations explain higher expression levels the most. Although being similar, the two clusters show different relative regulative relevance from repressors H2A.Z and H3K9me3, and activators H3K27ac and H3K9ac.

Cluster 3 embraces null to low transcriptional activity (RPKM median 2.32) and is characterized by a richer input pattern: more relevant than in other clusters are H3K27me3, H3K4me2 and H3K4me3. Bemusing are, however, activating and repressive roles attributed to, respectively, H2A.Z and H3K27ac. Despite this is in contrast with the functions commonly accredited to these features separately, single HMs might counterbalance one another and/or co-work to induce particular effects. Whether this observation suggests the cluster comprises genes of heterogeneous nature or a non-standard specific regulation pattern, this is still to be investigated.

Interesting is to notice the resemblance between the pattern of Cluster 3 with the genome-wide one. This is a further confirmation the algorithm has managed to expose the sub-group of genes possessing the largest leverage in bending one single regression hyperplane. Also, it has set apart the remaining population in a well *differentiated* manner: genes lacking the single genome-wide fit have been naturally stratified according to their expression value and in groups with distinct characteristic input patterns.

4 Conclusion

We proposed the application of a randomly re-initialized version of K-plane regression to expose sub-populations of protein coding genes commonly regulated at an epigenetic-histonic level. The proposed approach has revealed how single regression only captures the fit of a sub-group of genes with null to low expression and how poor scores from μ_{gw} on the remaining genes are due to *unfitting* rather than linear underfitting. The *hard* gene partitioning produced by the method allowed a statistical characterization of the computed clusters in terms of input contribution patterns, revealing how clusters stratify for higher and higher expression levels, with histone marks assuming specific roles of different relevance. Future developments will involve biological characterizations of the found gene clusters and investigations on the optimal choice for the value of hyperparameter K.

Acknowledgments

This work was supported by the Advanced ERC Grant “Data-Driven Genomic Computing (GeCo)” project (2016-2021), funded by the European Research Council.

References

[1] J. Vaquerizas, S. Kummerfeld, S. Teichmann, and N. Luscombe, “A census of human transcription factors: function, expression and evolution,” *Nature Reviews. Genetics*, vol. 10, pp. 252–263, 2009.

[2] C. Cheng, K.-K. Yan, K. Yip, J. Rozowsky, R. Alexander, C. Shou, and M. Gerstein, “A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets,” *Genome Biology*, vol. 12, p. R15, 2011.

[3] T. G. do Rego, H. G. Roider, F. A. T. de Carvalho, and I. G. Costa, “Inferring epigenetic and transcriptional regulation during blood cell development with a mixture of sparse linear models,” *Bioinformatics*, vol. 28, no. 18, pp. 2297–2303, 2012.

[4] N. Manwani and P. Sastry, “K-plane regression,” *Information Sciences*, vol. 292, pp. 39–56, 2015.

[5] L. Breiman, “Hinging hyperplanes for regression, classification, and function approximation,” *IEEE Trans. on Information Theory*, vol. 39, pp. 999–1013, 1993.