PREDICTION OF GENE EXPRESSION FROM TRANSCRIPTION FACTORS AFFINITIES: AN APPLICATION OF BAYESIAN NON-LINEAR MODELLING

Federico Marotta[1], Paolo Provero[1] and Silvia Montagna[2,3]

[1] Dipartimento di Neuroscienze "Rita Levi Montalcini", Università degli Studi di Torino, Via Cherasco, 15, 10126, Torino, Italy (e-mail: federico.marotta@edu.unito.it, paolo.provero@unito.it)

[2] Dipartimento di Scienze Economico-sociali e Matematico-statistiche, Università degli Studi di Torino, Corso Unione Sovietica, 218/bis, 10134 Torino, Italy, (e-mail: silvia.montagna@unito.it)

[3] Collegio Carlo Alberto, Piazza Vincenzo Arbarello, 8, 10122 Torino, Italy

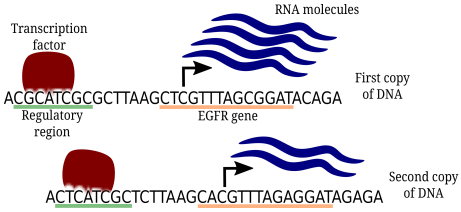
ABSTRACT: The prediction of gene expressions from DNA sequences is a relevant problem in biology. While most of the existing methods dedicated to this task use genotypes as predictors, here we propose a method based on transcription factor affinities, which have a clearer biological interpretation. This novelty, however, introduces new challenges for modelling, which we address leveraging on Bayesian non-linear modelling techniques.

KEYWORDS: Bayesian Methods, Gene Expressions, Non-linear Predictive Modelling.

1 Introduction

Scientists are often interested in predicting differences in the expression of a gene in different individuals solely from the DNA sequence of the individuals. The predicted expression can then be used in place of the real one when measuring the latter is too expensive, and the learnt relationship between DNA and expression can lead to a better understanding of how genes are regulated (Manor & Segal, 2013). The expression of a gene is the amount of RNA molecules it produces. Humans have two independent sets of DNA molecules, one coming from the father and one from the mother, therefore there are two copies of each gene. When measuring the expression, one simply sums the RNA molecules produced by each copy.

When associating DNA to gene expressions, the first problem we face is how to encode the DNA (a 3-billion letter string from the alphabet $\{A, C, G, T\}$)



Person	log Expr. of EGFR	Affinity of TF 1 for copy 1 of region 1	...	Affinity of TF l for copy 2 of region r
Alice	3.5	8.4	...	1.1
Bob	4.1	7.7	...	0.6
Craig	3.3	9.4	...	0.5
Dave	3.8	10.2	...	0.8
Eve	3.4	8.1	...	1.2
Frank	4.2	9.5	...	0.5
	Response (log $\mathbf{y}$)		Predictors ($\mathbf{A}$)	

Figure 1. *Left: Humans have two copies of DNA in each cell; the expression of a gene is the amount of RNA it produces. Transcription factors bind the DNA at the regulatory regions from where they activate or inhibit the expression of their target gene. Right: A plausible instance of our data set.*

into numbers to be used in a regression model. Most existing methods rely on genotypes (discrete variables taking values 0, 1, or 2 encoding single-letter differences in the DNA of different people), which do not allow for easy interpretation (*e.g.*, "If the DNA has an 'A' instead of a 'T', the expression of the gene will be higher"). Our first goal is to develop a more interpretable model.

Gene expression is mainly controlled by specialised proteins called transcription factors, which bind the DNA at particular locations (regulatory regions) by establishing weak chemical bonds. Different DNA sequences will have, therefore, different chemical *affinities* for the transcription factors. Since different individuals have different DNA sequences, it is possible to use the affinities for transcription factors as numerical (continuous) predictors in the predictive model of gene expression. Affinities have a far superior interpretation, exemplified by statements such as "If the affinity for this transcription factor is higher, the expression will be higher."

However, one needs to make an assumption about the relationship (*e.g.*, linear) between affinities and gene expression. de Boer *et al.* , 2020 models the logarithm of the expression as a linear function of the affinities. The model is developed for a type of yeast and achieves a good performance, but is still too simple for our application. Indeed, yeast has two important distinguishing features: 1) it is haploid, meaning that it has only one copy of DNA, whereas humans have two; and 2) its genes are regulated primarily by one regulatory region, whereas human genes typically have more than one.

In this paper, we set up a predictive model for the expression of the EGFR (Epidermal Growth Factor Receptor) gene, and explicitly address both limitations in de Boer *et al.* , 2020. Figure 1 provides a schematic of our application.

2 Methodology and results

Our dataset consists in the expression values of the EGFR gene for 414 individuals (from The GTEx Consortium, 2020), and in the affinity of each regulatory region for all transcription factors, for a total of 358 predictors.

We can take multiple regulatory regions into account (goal 2 above) via a straightforward modification of the model in de Boer *et al.* , 2020, which becomes: $\log(y) = \beta_0 + \sum_{g=1}^{r} \sum_{f=1}^{l} A_{fg} \beta_{fg}$. Here y denotes the gene expression, $\{A_{fg}\}_{f=1,g=1}^{l,r}$ are the affinities, and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{rl})^\top$ is a vector of model parameters. Similarly to de Boer *et al.* , 2020, we sum over all transcriptions factors, indexed by f, but now also along the regulatory regions, g, of the gene.

Accommodating for both copies of DNA (goal 1 above) is more challenging. Biologically, we know that the effects of the two copies should be additive in the original scale of the expression, not in the log-transformed expression. At the same time, working with the expression in the original scale can be troublesome, for it is often not normally distributed. Therefore, we propose the following model for the expression of a single gene:

$$\log(\boldsymbol{y}) \sim \text{mvnormal}\left(\log\left(e^{\boldsymbol{A}^{(1)}\boldsymbol{\beta}} + e^{\boldsymbol{A}^{(2)}\boldsymbol{\beta}}\right), \sigma^2 \boldsymbol{I}\right). \tag{1}$$

Here $\boldsymbol{y}$ is an n-vector of expression values (one for each individual), $\boldsymbol{A}^{(i)}$, with $i \in \{1,2\}$, is the $n \times rl$ affinity matrix for copy i, where each column represents a transcription factor-regulatory region pair (r is the number of regions, l the number of transcription factors), and vector $\boldsymbol{\beta}$ ($lr \times 1$) encapsulates the coefficients of the affinities. By computing the exponential of $\boldsymbol{A}^{(i)}\boldsymbol{\beta}$, with $i \in \{1,2\}$, we obtain the effect of copy i on the expression in the original scale. We subsequently sum the two effects, and take the log of the sum to go back to the log-scale response. Importantly, the coefficient of a given transcription factor in a given regulatory region is the same for the two copies of DNA. We notice that for this reason our model does not fall in the class of generalised linear models (at least not obviously), as each coefficient β_j appears two times independently for two different predictors.

Model (1) is embedded in a Bayesian framework by placing a normal prior (with mean zero and variance τ) on all coefficients $\boldsymbol{\beta}$ independently, and an inverse-gamma prior on σ^2. We reparameterise τ as $\frac{\sigma^2}{b}$, so that b can also be interpreted as the parameter of a Ridge penalty.

To carry out an unbiased evaluation of the performance, we implemented a nested cross-validation strategy where the outer 5-fold loop evaluates the

Table 1. *Results of the nested-cross validation. MSE is the mean squared error,* ρ *the correlation between true and predicted expression; averages and standard deviations of these quantities are computed across the 5-folds. Avg* R^2 *is the average of the squared correlations. Z is the Z-score computed via Stouffer's method, which combines the* ρ *of the five folds, and pval Z is the p-value of the Z-score.*

Gene	**Avg MSE**	**Sd MSE**	**Avg** ρ	**Sd** ρ	**Avg** R^2	Z	**pval** Z
EGFR	0.012	0.001	0.140	0.128	0.033	2.852	0.002

performance, and the inner 10-fold loop tunes the parameter b. Table 1 summarises the results. While the average R^2 may seem small, we emphasise that low values are common in the prediction of gene expression and our model outperforms recently published genotype-based models (for instance, the R^2 achieved by Nagpal *et al.* , 2019 is only 0.005).

Thus, our method can model the underlying biological problem in a realistic way and provide meaningful results thanks to its interpretable predictors. In the future, it could be improved by considering interactions between transcription factors, which are also biologically important. Nevertheless, for the time being, we hope that non-linear models will find their way in the field of gene expression prediction, which currently is dominated by genotype-based linear models.

References

DE BOER, CARL G., VAISHNAV, EESHIT DHAVAL, SADEH, RONEN, *et al.* . 2020. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.*, **38**(1), 56–65.

MANOR, OHAD, & SEGAL, ERAN. 2013. Robust Prediction of Expression Differences among Human Individuals Using Only Genotype Information. *PLoS Genet.*, **9**(3), e1003396.

NAGPAL, SINI, MENG, XIAORAN, EPSTEIN, MICHAEL P., *et al.* . 2019. TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. *Am. J. Hum. Genet.*, **105**(2), 258–266.

THE GTEX CONSORTIUM. 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, **369**(6509), 1318–1330.