Probing instructions for gene expression regulation in gene nucleotide compositions

Gene expression regulation

Taha, Bessière et al.
Probing DNA instructions for expression regulation
Previous work

- Predicting Epigenetics data from DNA sequence
- Predicting gene expression from epigenetics data

Question: Can we identify directly the DNA determinants involved in gene regulation?
Plan

1. Model building
2. Comparison with experimental data (Chip-Seq)
3. Advanced model
4. Biological interpretation
Our Work

Originality:
- Modeling gene expression using DNA sequence data only
- ONE model per patient (Cancer tumors)

Data
- Gene expression measurements for each patient (RNA-Seq)
- DNA sequence (Genome Reference GRCh38/hg38)
  - Nucleotide and di-nucleotide compositions: \( \%CG = \frac{\#CG}{\text{length}-1} \)
  - TF binding motifs: PWM scores
  - DNA shapes (computed with the Bioconductor package DNAshapeR)

Sequence variations affect histone modifications
Response variable: RNA-Seq (log transformed values)

- Gene expression measured by RNA-seq (reads count)
- 12 different types of cancer from TCGA: Breast, Leukemia, Liver...
We built a global linear regression model to explain the expression of genes using DNA/RNA features associated with their regulatory regions (e.g. nucleotide composition, TF motifs, DNA shapes):

\[ Y = X\beta + \varepsilon \]

where

- \( Y_{[n \times 1]} = (y_1, \ldots, y_n)' \) is the vector of observed gene expression,
- \( X_{[n \times p]} = (x_{ij}) \) is the feature matrix (\( x_{ij} \) is feature \( j \) for gene \( i \)),
- \( \beta_{[p \times 1]} = (\beta_1, \ldots, \beta_p)' \) is the vector of regression coefficients
- \( \varepsilon_{[n \times 1]} = (\varepsilon_1, \ldots, \varepsilon_n)' \) is the vector of the residual errors.

\((n \sim 20000)\)
Variable selection with Lasso

- Linear regression with $\ell_1$-norm penalty or Lasso (Tibshirani, 1996) applied to standardized data:

  $$\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} \left( \sum_{g=0}^{n} (Y - X \beta)^2 + \lambda \sum_{i=0}^{p} |\beta_i| \right)$$

- The penalty $\lambda$ is chosen by 10-fold cross-validation to minimize the mean square prediction error.

- Some coefficients $\beta_i$ are set to 0 exactly ($\ell_1$-norm geometry).

- $\lambda$ defines the number of selected variables.
Model evaluation

- **Criterion:**
  1. Mean square error (MSE)
  2. Correlation coefficient \( \text{Corr}(Y, \hat{Y}) \) between the measured expression \( Y \) and the predicted expression \( \hat{Y} \)

in a 10-fold **cross-validation** procedure:

  1. Model is learned in the training data
  2. MSE/\( \text{Corr}(Y, \hat{Y}) \) is evaluated in the test data.

- **Data shown:** RNA-Seq gene expression (TCGA) from **12 cancers types, 20 patients per cancer.**

  (+ Further evaluation not shown: 1,270 RNA-Seq samples and 582 microarrays datasets.)
Promoter definition

Nucleotide and di-nucleotide compositions: \( \%CG = \frac{\#CG}{\text{length}-1} \)
The highest accuracy was obtained combining the 3 segments.
Our model achieved higher predictive accuracy with the promoters centered around the 2nd TSS, in agreement with Cheng et al. (2012).
Considering all (di-)nucleotides achieved better model performance.
The increase in performance when including TF motifs or DNA shapes is rather modest.
Plan

1. Model building
2. **Comparison with experimental data (Chip-Seq)**
3. Advanced model
4. Biological interpretation
DNA features vs. experimental data (ChIP-seq)

- Comparison with models integrating:
  - TF-binding signals with Chip-Seq (RACER, Y. Li and al. PLoS, 2014)
  - Open-chromatin signals (TEPIC, Schmidt F. et al. NAR, 2017)

- In both cases, the models were built using the same set of genes:
  (i) on the original data,
  (ii) on randomized predictive variables (gene centered shuffling: rand)
  (iii) on the maximum value of all predictive variables (gene centered maximum: max).
In cases (ii) and (iii), the links between the predictive variables and expression is broken and a regression model is expected to poorly perform as our model does (Left, light pink).
In cases (ii) and (iii), the links between the predictive variables and expression is broken and a regression model is expected to poorly perform as our model does (Left, light pink). ** ** **
Plan

1. Model building
2. Comparison with experimental data (Chip-Seq)
3. Advanced model
4. Biological interpretation
Gene expression regulation

Taha, Bessière et al.
Probing DNA instructions for expression regulation
Contribution of additional genomic regions

Nucleotide and di-nucleotide compositions: \( %CG = \frac{\#CG}{\text{length}-1} \) in 8 selected regions (20 variables per region)
Contribution of additional genomic regions

DNA regions ‘forward-like’ selection procedure

Our model: Nucleotide and di-nucleotide compositions in 8 selected regions (20 variables per region)
### Contribution of additional genomic regions

DNA regions ‘forward-like’ selection procedure

Our model: Nucleotide and di-nucleotide compositions in **8 selected regions** (20 variables per region)
Stable variables selection

- **Stability selection** (Meinshausen et al., 2010)

- Lasso inference is repeated 500 times where, for each iteration,
  (i) only 50% of the genes is used (uniformly sampled)
  (ii) a random weight (uniformly sampled in [0.5; 1]) is attributed to each predictive variable.

- A variable is considered as stable if selected in more than 70% of the iterations.

  (Functions stabpath and stabsel from the R package C060 for glmnet models.)
Stable variables selection

<table>
<thead>
<tr>
<th>Variable</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>TpG_CORE</td>
<td>0.0</td>
</tr>
<tr>
<td>CpA_DD</td>
<td>0.2</td>
</tr>
<tr>
<td>CpA_5UTR</td>
<td>0.4</td>
</tr>
<tr>
<td>GpT_INTR</td>
<td>0.6</td>
</tr>
<tr>
<td>CpA_CORE</td>
<td>0.8</td>
</tr>
<tr>
<td>CpG_CORE</td>
<td>1.0</td>
</tr>
<tr>
<td>ApT_DFR</td>
<td>0.0</td>
</tr>
<tr>
<td>GpA_INTR</td>
<td>0.2</td>
</tr>
<tr>
<td>ApT_CORE</td>
<td>0.4</td>
</tr>
<tr>
<td>GpT_5UTR</td>
<td>0.6</td>
</tr>
<tr>
<td>CpA_INTR</td>
<td>0.8</td>
</tr>
<tr>
<td>CpG_5UTR</td>
<td>1.0</td>
</tr>
<tr>
<td>GpT_3UTR</td>
<td>0.0</td>
</tr>
<tr>
<td>CpA_INTR</td>
<td>0.2</td>
</tr>
<tr>
<td>CpG_3UTR</td>
<td>0.4</td>
</tr>
<tr>
<td>GpT_3UTR</td>
<td>0.6</td>
</tr>
<tr>
<td>TpA_DDR</td>
<td>0.8</td>
</tr>
<tr>
<td>GpC_3UTR</td>
<td>1.0</td>
</tr>
<tr>
<td>A_3UTR</td>
<td>0.0</td>
</tr>
<tr>
<td>CpA_CDS</td>
<td>0.2</td>
</tr>
<tr>
<td>GpT_CDS</td>
<td>0.4</td>
</tr>
<tr>
<td>A_3UTR</td>
<td>0.6</td>
</tr>
<tr>
<td>CpA_CDS</td>
<td>0.8</td>
</tr>
<tr>
<td>GpT_CDS</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Proportion of samples in which each variable is selected with high consistency (> 70% stability)

(Average ∼ 16 variables per sample)
Plan

1. Model building
2. Comparison with experimental data (Chip-Seq)
3. Advanced model
4. Biological interpretation
A) DNA features associated with good predictions

- We characterized best predicted genes with regression trees (CART) which performs sequentially binary splits (minimizing RSS).
- Response variable is the prediction error of our linear model.
- (di-)nucleotide compositions are used as classifiers.
A) DNA features associated with good predictions

- Columns: samples gathered by cancer type, ranked by decreasing error
- Lines: the 3,680 groups of genes ranked by decreasing error
- Red and light blue: Top 25% well predicted groups of genes

⇝ Our model mainly fits certain classes of genes with specific genomic features
Groups of genes well predicted in all cancers (low prediction error) seems to correspond to ubiquitously expressed and housekeeping genes.

Functional enrichment for general and widespread biological processes:

<table>
<thead>
<tr>
<th>Gene ontology term</th>
<th>Count</th>
<th>Benjamini corrected P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cellular macromolecule metabolic process</td>
<td>612</td>
<td>1.8E-23</td>
</tr>
<tr>
<td>Cellular metabolic process</td>
<td>681</td>
<td>1.2E-16</td>
</tr>
<tr>
<td>Cellular protein metabolic process</td>
<td>390</td>
<td>2.8E-16</td>
</tr>
<tr>
<td>Macromolecule metabolic process</td>
<td>624</td>
<td>4.0E-16</td>
</tr>
<tr>
<td>Nucleic acid metabolic process</td>
<td>404</td>
<td>4.0E-16</td>
</tr>
</tbody>
</table>
Groups well predicted in only certain cancer types

- In contrast, groups well predicted in only certain cancers are associated to specific biological function.

  For instance, a regression tree learned in one PAAD sample identified a group of 1,531 genes, which has:

  - Low prediction error in LGG and PAAD but high error in LAML, LIHC and DLBC.
  - Functional enrichment for specific biological processes (brain).

<table>
<thead>
<tr>
<th>Gene ontology term</th>
<th>Count</th>
<th>Benjamini corrected P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive regulation of cellular process</td>
<td>528</td>
<td>7.0E-14</td>
</tr>
<tr>
<td>Nervous system development</td>
<td>284</td>
<td>1.3E-13</td>
</tr>
<tr>
<td>Positive regulation of macromolecule metabolic process</td>
<td>346</td>
<td>3.5E-12</td>
</tr>
<tr>
<td>Positive regulation of biological process</td>
<td>565</td>
<td>8.1E-12</td>
</tr>
<tr>
<td>Neurogenesis</td>
<td>200</td>
<td>5.9E-11</td>
</tr>
</tbody>
</table>
B) Link with the genome architecture

Do the groups of genes identified by the regression trees correspond to specific TADs?

Motivations

- Genes within the same TAD tend to be coordinately expressed (Nora et al. 2012, Fanucchi et al. 2013).
- Nucleotide composition along the genome can help define TADs (Jabbari and Bernardi, 2017)

Validation:

- We used the 373 TADs containing more than 10 genes.
- For each TAD and each (di-)nucleotide, we used a Kolmogorov-Smirnov test to compare the (di-)nucleotide distribution of the embedded genes with that of all other genes (multiple testing controlled with FDR).

→ 87% of the TADs are characterized by at least one specific nucleotide signature.
We next considered the 967 groups of genes whose expression is accurately predicted by our model (regression trees).

60% of the well predicted groups of genes (top 25% well predicted) were enriched for at least one TAD (p-value < 0.05, hyper-geometric test).
We confirm the existence of sequence-level instructions for gene expression by developing a method able to explain the expression of different genes using only DNA sequence.

Our model is as accurate as methods based on experimental data but its biological interpretation appears more straightforward.

We provide evidence that the genes nucleotide composition can be linked to co-regulations associated with genome 3D architecture and to associations of genes within TADs.
Ongoing work

- Further improve the model
  - Relax linearity assumption ?
  - Include variable interactions ?
    (+ Comparison with deep learning approaches)
  - Integrate TF binding motifs ?

- Get more biology
  - TADs
  - methylation
  - ...

Taha, Bessière et al. Probing DNA instructions for expression regulation


The team!