Applied Machine Learning

Project Proposal

Target

Genetic component of gene expression (as measured with RNA-Seq experiments) of individual genes in a specific cellular type. (see Background section)

Background (For non-biotech)

Differential gene expression

The differential level of expression of the genes in a cell, determined by molecular mechanisms that regulate the activation, the inactivation, and the regulation of the gene transcription from DNA to RNA, are known to be at the very basis of basically all biological processes in living systems.

Differential gene expression is critical in determining cellular differentiation, cellular response to stress and environmental stimuli, cellular matabolism, apoptosis (programmed cellular death), immune response, cellular proliferation, cancer, ... etc.

Quantifying gene expression

Here the term gene expression refers to the relative abundance of a given transcript - e.g. number of molecules of messenger RNA (mRNA) of a certain gene - in a population of cells of the same type (e.g. hepatocytes)

RPKM (Reads Per Kilobase of transcript per Million mapped reads) is the unit used to quantify the relative level of expression of the transcripts as detected in RNA Sequencing experiments

Regulation of gene expression

The current model that describes the activation of transcription of a gene in eukaryotes is quite complex.

The RNA polymerasis (RNApolII) is the protein that ultimately catalyses the transcription process, but for RNA polymerasis to bind the promoter region and to start the actual transcription of a region of DNA, a huge number of events has to occur; here we highlight the only most critical ones, that are those that in some way are taken in consideration when looking at the features in the machine learning workflow:

- a number of different **Transcription Factors (TFs)**, i.e. proteins that can bind the DNA in non-random points and influence the rate of transcription (making it more or less likely to occur) of various genes, have to bind the regulatory region in proximity of the transcribed region;
    - Note: in the very same cell type, a given TF may activate the gene A binding the regulatory region of gene A, but inactivating gene B binding the regulatory region of gene B
- ~~To be possible for the TFs and the RNApolII to bind the DNA in a certain region, the chromatin - i.e. the physical structure, that includes other molecules such as histones, in which the DNA is arranged in the cell nucleus, of that region has to be *accessible*. **Chromatin accessibility** is mainly cell type-specific and no individual chromatin accessibility data are available for me to use at the best of my knowledge. For~~

~~these reasons this potential feature will be neglected.~~

Total Binding Affinity (TBA)

Trancription Factors (TFs) bind the DNA on non-random short sequences of nucleotides (e.g. ATCAGGGT). Each TF has a binding affinity for the DNA that is a function of the sequence of nucleotides. The sequence preferences for each TF are well studied and described by Position Weight Matrices (PWM). For instance, this is a visual representation of the PWM of the human TF HOXA1:

(Note: it seems that HOXA1 binding sites are observed to have almost always an "A" in the 4th position and a "T" in the 5th position)

TBA is a score for quantifying the affinity of a TF (with a known PWM) for a given regulatory sequence, that may be 1500 base pairs long, for instance.

A regulatory region with a high TBA score for a given TF is expected to be a potential binding site of that TF, and the expression of the gene downstream to be regulated by that TF

Mutations in the regulatory sequence due to genetic variability in different individuals may change the affinity for a TF, and the change should be observed in the TBA score as well. In this case, the expression of the gene linked to the mutated regulatory region may be affected by a change in its level of expression

Target vector for the gene A

In each **individual (Ind_N)** The **gene A** is associated to a **level of gene expression (RPKM_GeneA)**, that we want to predict

This table should be generated for each different gene:

Individual	RPKM_GeneA
Ind_1	<float>
Ind_2	<float>
Ind_3	<float>
...	...
Ind_Y	<float>

Features matrix for the gene A

The Gene A is associated to a regulatory sequence, that may have mutations in different individuals. For each Transcription Factor X the TBA score is computed on that region: **TBA_ TF_X**

Transcription Factors (TF) are genes themselves and are associated to a level of expression in each individual **RPKM_TF_X**

This table should be generated for each different gene:

	TBA_TF_1	RPKM_TF_1	TBA_TF_2	RPKM_TF_2	...	...	TBA_TF_X	RPKM_TF_X
Ind_1	<float>	<float>	<float>	<float>	...	...	<float>	<float>
Ind_2	<float>	<float>	<float>	<float>	...	...	<float>	<float>
Ind_3	<float>	<float>	<float>	<float>	...	...	<float>	<float>
Ind_4	<float>	<float>	<float>	<float>	...	...	<float>	<float>
...	...	...	...	...	...	...	...	...
Ind_Y	<float>	<float>	<float>	<float>	...	...	<float>	<float>

Columns

The total number of columns should be around 1000 since around 400 human transcription factors will be considered, from the HOCOMOCO PWM dataset

- **TBA_TF_N**: TBA value for the TF_N on the regulatory region of the gene A
- **RPKM_TF_N**: transcript quantification (in RPKM) of the trancription factor TF_N
- ~~**H3K27ac**: chromatin accessibility level of the regulatory region of gene A (measured as the level of acetylation of the 27th Lysin if Histon H3). Not used (see Total Binding Affinity (TBA) section)~~

Rows

The total number of rows is 344 since genomic and transcriptomic data for 344 individuals is available in the GEUVADIS public dataset

- **Ind_N**: individual number N, from whom the sample of cells in which the DNA and the RNA was sequenced comes from

Expected Results

- Significant prediction of gene expression for at least a set of genes
- Selection of relevant features by the algorithm will give insights on biological mechanisms: each gene is mostly regulated by a specific little subset of transcription factors, i.e. few features should be critical for a given gene to make a prediction

Potential weakness

- The number of features is much higher than the number of examples. Not all the regulatory regions may be interested by mutations among individuals; basically, the dataset is not huge but I expect that for each gene only few features would be relevant and then included in the final hypothesis