Linking recombinant genes sequence to protein products

Javier González
Join work with Neil D. Lawrence, David James, Joseph Longworth

23, February 2015

The University Of Sheffield.

Sheffield Institute for Translational Neuroscience

The big picture

"Civilization advances by extending the number of important operations which we can perform without thinking of them."
(Alfred North Whitehead)

Project background

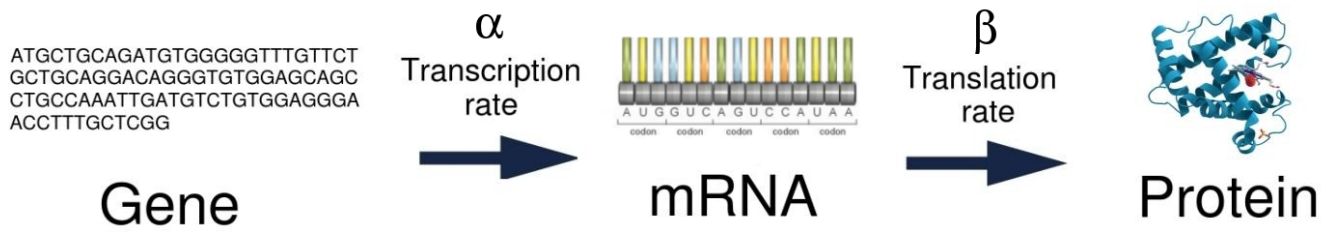
- Use cells as factories to produce compounds of drugs.

Research agenda

- Optimization of proteins production using hamster cells. Minimize the number of lab experiments!
- Development of new Machine Learning techniques.

Motivation

- Synthetic biology: re-design of elements of living systems for useful purposes



- Industrial interest: design of synthetic genes to engineer living cells to produce compounds of interest.
- Optimization goal: rewrite gene sequences to optimize protein production (increase transcription and translation rates).

The problem in a nutshell

Proteins: Fundamental bricks of the cell.
Genes: Information to synthesize proteins (1 protein $\rightarrow$ 1 gene)

Gene sequence
ATGCTGCAGATGTGGGGGTTTGTTCTCTATCTCTTCCTGAC
TTTGTTCTCTATCTCTTCCTGACTTTGTTCTCTATCTCTTC...

Considerations

- Different gene sequences $\rightarrow$ same protein.
- The sequence affects the synthesis efficiency.

Which is the most efficient sequence to produce a protein?

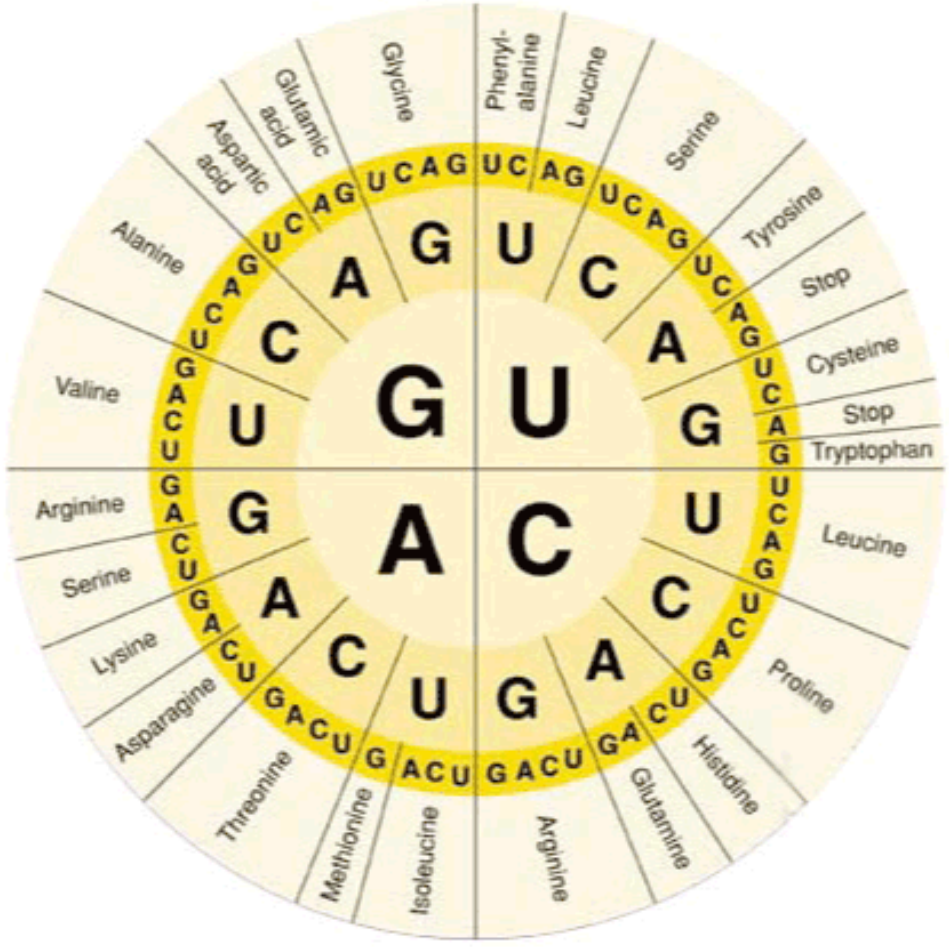
Redundancy of the genetic code

- Codon: Three consecutive bases: AAT, ACG, etc.
- Protein: sequence of amino acids.
- Different codons may encode the same aminoacid.
- ACA=ACU encodes for Threonine.

$$ATUUUGACA = ATUUUGACU$$

synonyms sequences. $\rightarrow$ same protein but different efficiency

‘Complexity’ of the genetic code



How to design a synthetic gene?

Very complex cell behavior!

Gene sequence (features) $\rightarrow$ protein production efficiency

Bayesian Optimization (BO)

do:

1. Model as an emulator of the cell behavior.
2. Obtain a set of gene design rules.
3. Design a new gene coherent with the design rules.
4. Repeat experiment (get new data).

until the gene is optimized (or the budget is over...)

1. Model as an emulator of the cell behavior

Model inputs:

- Set of gene sequences from which we extract some biologically relevant features: codon frequency, cai, gene length, etc.

Model outputs:

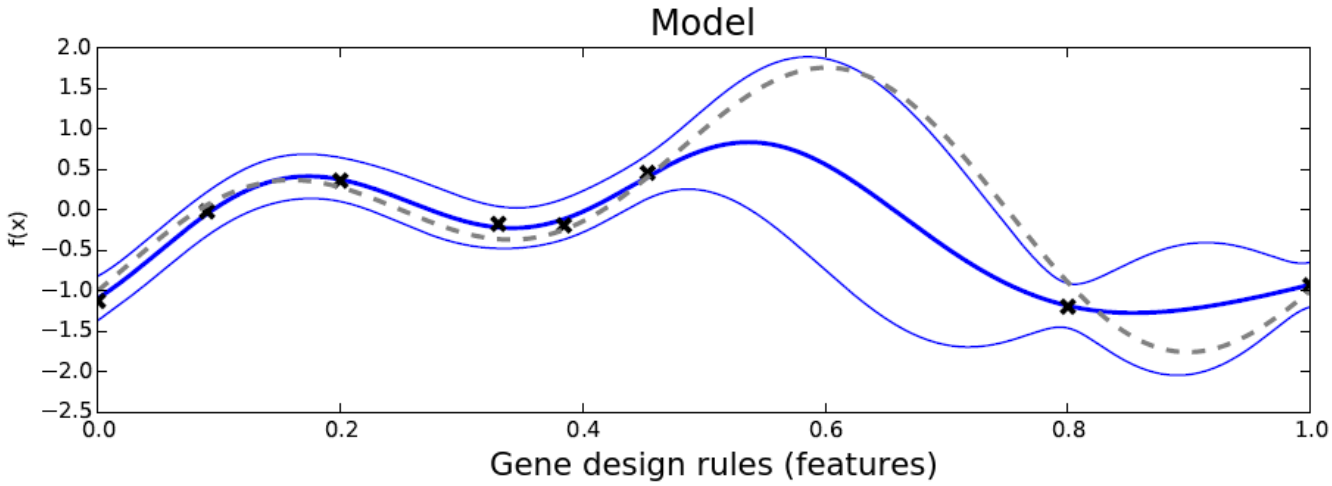
- Transcription and translation rates.

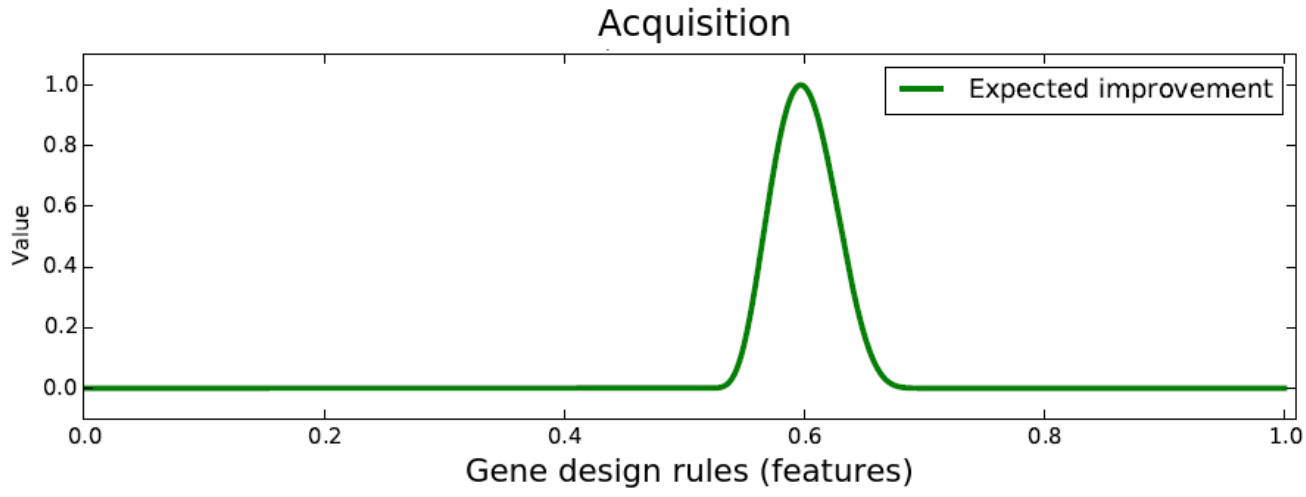
Model type:

- Multi-output Gaussian process model.

2. Obtain a set of gene design rules

The model is used to determine which are the best gene design rules $x^{\star}$.





3. Design a new gene coherent with the desing rules

1. Simulate a set of genes coherent with the one we want to design.

2. Extract the features from all the simulated genes.

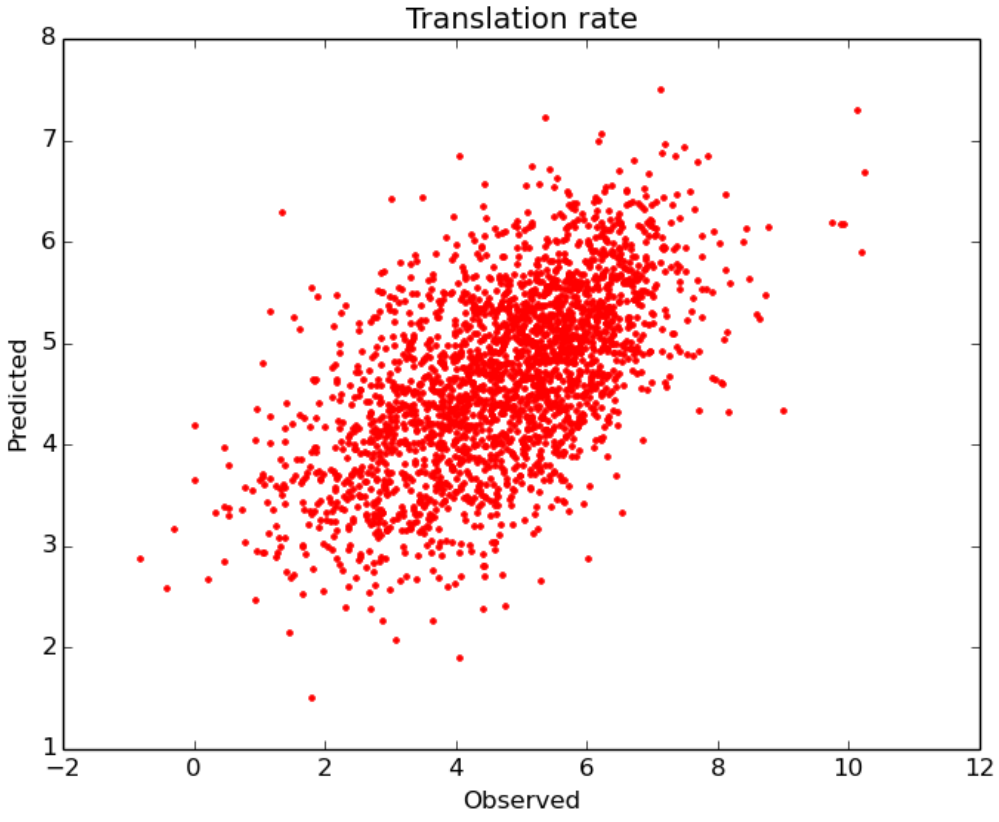
3. Rank the generated genes according to their similarity to the 'optimal' design rules.

4. Pick up the best one and test it in the lab.

Experiments

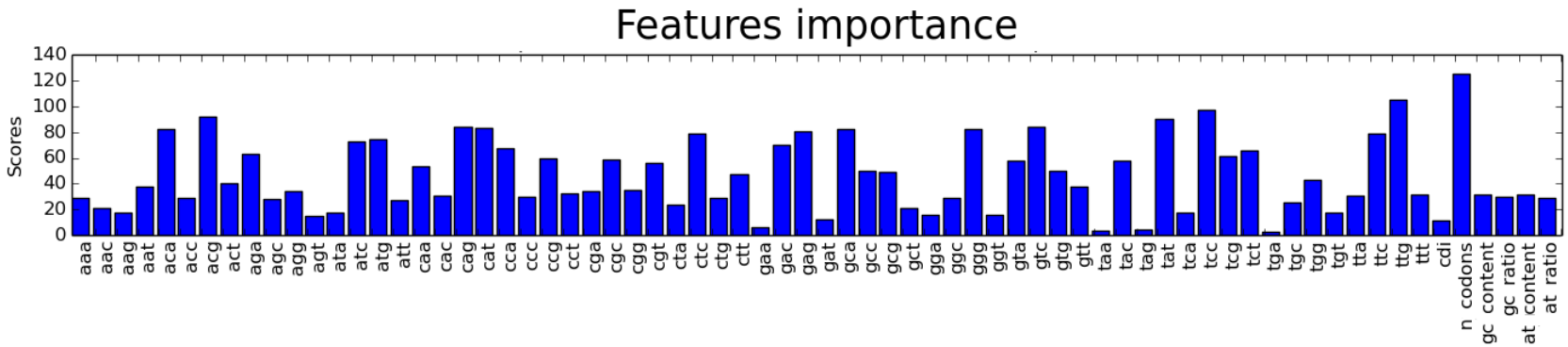
- Optimization gene designs in mammalian cells.
- Dataset in Schwanhausser et al. (2011) for 3810 genes rates. Associated sequences were extracted from http://www.ensembl.org.
- Features: frequency of appearance the 64 codons, length of the gene, GC-content, AT-content, GC-ratio and AT-ratio.
- Selection of 10 random difficult-to-express genes (average log ratio < 1.5).
- 1,000 random 'synonyms sequences' generated from each gene.

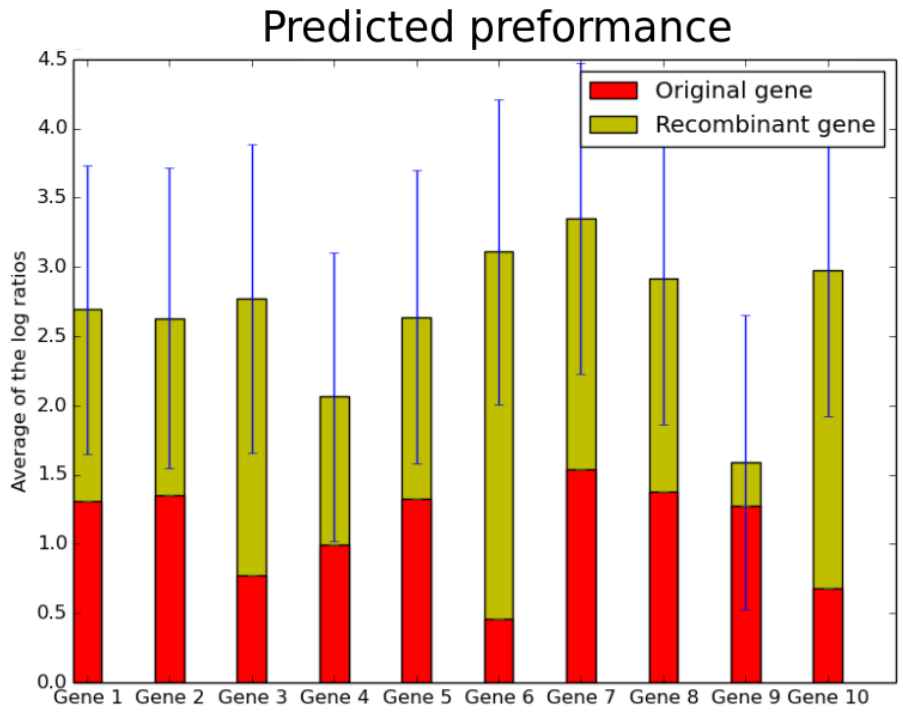
Prediction of the translation rates (in a test sample of 1500 genes)



correlation ≈ 0.6

Results for 10 low-expressed genes





Conclusions and future work

- Bayesian optimization is a promising technique to design synthetic genes. Reduces the number experiments!!
- Important aspects of the problem $\rightarrow$ to have a good surrogate model for the cell behavior.
- More features $\rightarrow$ better gene design $\rightarrow$ development of scalable Bayesian optimization methods able to work well in high dimensions.
- Alternative approach: focus on the optimization of the sequences. Combinatorial problem.