Bayesian Optimization for Synthetic Gene Design

Javier González

DDHL workshop, Nottingham, UK, 2015

The University Of Sheffield.

Sheffield Institute for Translational Neuroscience

At the University of Sheffield:

- Department of Computer Science.
- Department of Chemical and Biological Engineering.

Neil D. Lawrence, David James, Joseph Longworth, Paul Dobson, Josselin Noirel and Mark Dickman

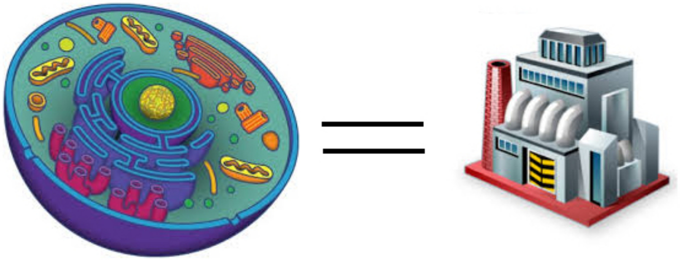
BRIC • BIOPROCESSING RESEARCH INDUSTRY CLUB

BBSRC
bioscience for the future

The big picture

- 8 of the 10 top-selling drugs in are biologics (monoclonal antibodies) used in rheumatology, dermatology, and various types of Cancer.
- Huge market of $73 billion just in Europe.
- Growing interest in the availability of biosimilars.

New drug production paradigm



- Use mammalian cells to make protein products.
- Control the ability of the cell-factory to use synthetic DNA.

Cornerstone of modern biotechnology: Design DNA code that will best enable the cell-factory to operate most efficiently.

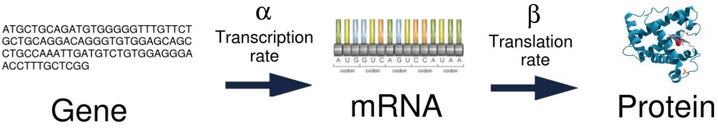
Synthetic genes!

How does a cell work?

A mammalian cell in numbers:

- approx. 20,000 genes able to produce 20,000 proteins.
- A few of them are of therapeutical interest.
- The average gene length is 7902 bases pairs (A,T G, C).
- Millions of molecules interactions.

Central dogma of systems biology

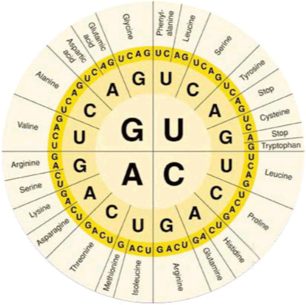

ATGCTGCAGATGTGGGGGTTTGTTCT
GCTGCAGGACAGGGTGTGGAGCAGC
CTGCCAAATTGATGTCTGTGGAGGGA
ACCTTTGCTCGG
Gene
α
Transcription rate
A U G G U C A G U C C A U A A
codon
codon
codon
codon
codon
mRNA
β
Translation rate
Protein

'Natural' genes are not optimized to maximize protein production.

Why can we rewrite the genetic code?

Considerations

- Different gene sequences may encode the same protein...
- ...but the sequence affects the synthesis efficiency.
- The codon usage is the key (codon = triplet of bases).



The genetic code is redundant:

$$UUG\,ACA = UUG\,ACU$$

Both genes encode the same protein.

Key question

Given a protein of interest, which is the recombinant gene sequence that will enable the cell to produce it in the most efficient way?

Key question

Given a protein of interest, which is the recombinant gene sequence that will enable the cell to produce it in the most efficient way?

- Average mamalian gene: 7000 nucleotides.

Key question

Given a protein of interest, which is the recombinant gene sequence that will enable the cell to produce it in the most efficient way?

- Average mamalian gene: 7000 nucleotides.
- Consider a gene with coding region of 900 nucleotides: 300 codons.

Key question

Given a protein of interest, which is the recombinant gene sequence that will enable the cell to produce it in the most efficient way?

- Average mamalian gene: 7000 nucleotides.
- Consider a gene with coding region of 900 nucleotides: 300 codons.
- Assume only pairs of synonymous codons.

Key question

Given a protein of interest, which is the recombinant gene sequence that will enable the cell to produce it in the most efficient way?

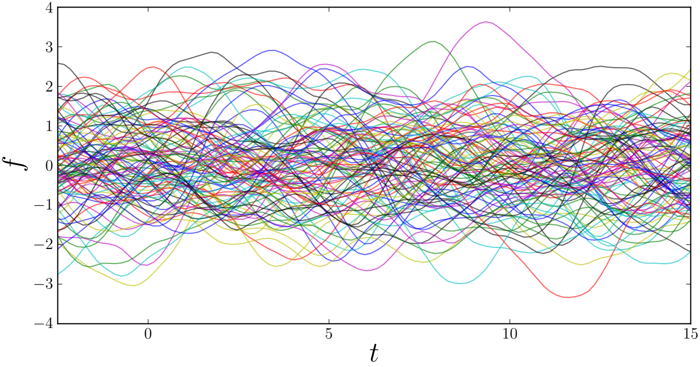
- Average mamalian gene: 7000 nucleotides.
- Consider a gene with coding region of 900 nucleotides: 300 codons.
- Assume only pairs of synonymous codons.
- $\approx 2^{300} \approx 2 \times 10^{90}$ possible recombinant gene alternatives (in the order of the number of atoms in the universe).

Machine Learning challenges

- Very complex cell behaviour. Limited prior knowledge.
- Multi-task optimization problem: increase cell efficiency, maintain cell survival, control protein and mRNA stability.
- Lab experiments are very expensive.
- Gene tests can be run in parallel.
- The design space is defined in terms of long string sequences.
  - Alternative: gene features, high dimensional problem.

Tools: Gaussian processes

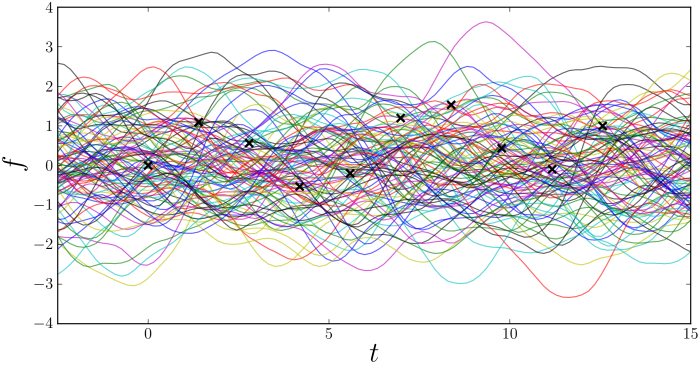
Gaussian Process: Probability density over functions, such that each linear finite dimensional restriction is multivariate Gaussian.



- Fully parametrized by a covariance function K.
- Close-form posterior under Gaussian likelihoods.

Tools: Gaussian processes

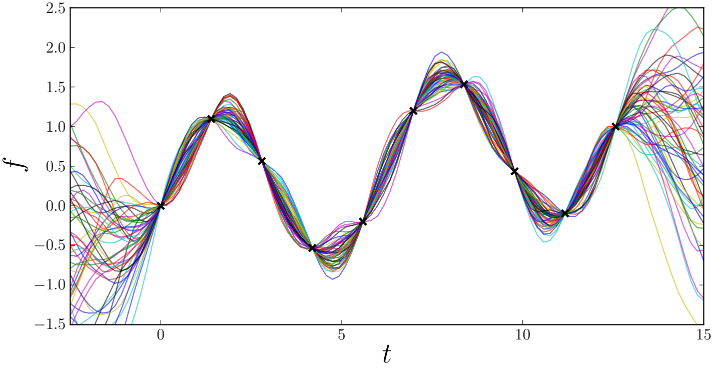
Gaussian Process: Probability density over functions, such that each linear finite dimensional restriction is multivariate Gaussian.



- Fully parametrized by a covariance function K.
- Close-form posterior under Gaussian likelihoods.

Tools: Gaussian process

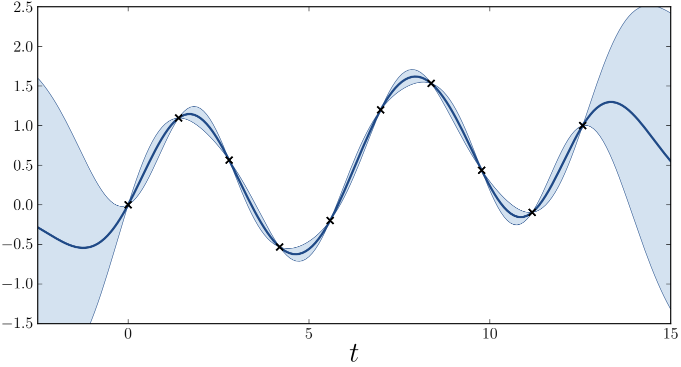
Gaussian Processes: Probability density over functions, such that each linear finite dimensional restriction is multivariate Gaussian.



- Fully parametrized by a covariance function K.
- Close-form posterior under Gaussian likelihoods.

Tools: Gaussian processes

Gaussian Process: Probability density over functions, such that each linear finite dimensional restriction is multivariate Gaussian.

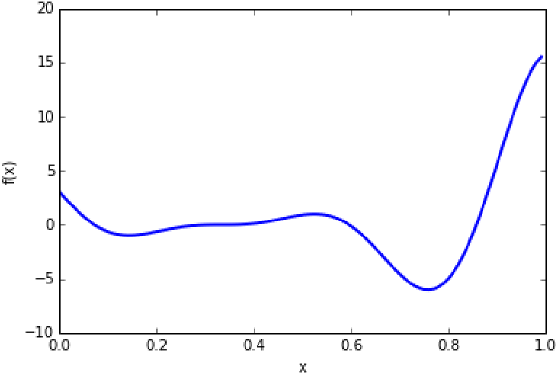


- Fully parametrized by a covariance function K.
- Close-form posterior under Gaussian likelihoods.

Tools: Bayesian Optimization

BO: Heuristic to reduce the number of evaluations in optimization problems [Mockus, 1978; Snoek et al., 2012].

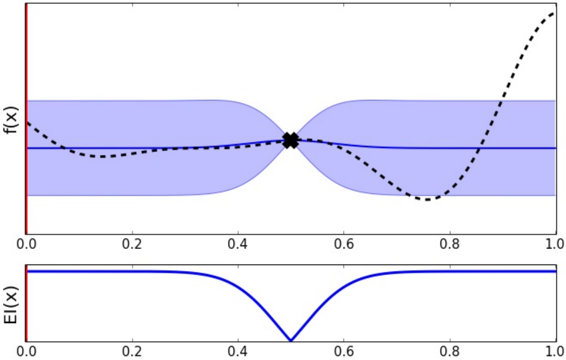
Example: $x^* = \arg\min_{[0,1]} f(x)$?



Tools: Bayesian Optimization

BO: Heuristic to reduce the number of evaluations in optimization problems [Mockus, 1978; Snoek et al., 2012].

Example: $x^* = \arg\min_{[0,1]} f(x)$?

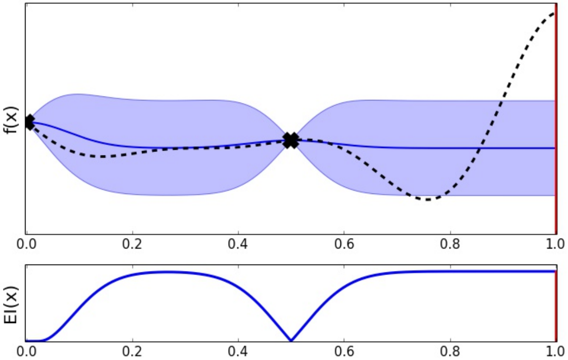


Expected Improvement: $E_x[\max(0, f(x_{min}) - f(x))]$

Tools: Bayesian Optimization

BO: Heuristic to reduce the number of evaluations in optimization problems [Mockus, 1978; Snoek et al., 2012].

Example: $x^* = \arg\min_{[0,1]} f(x)$?

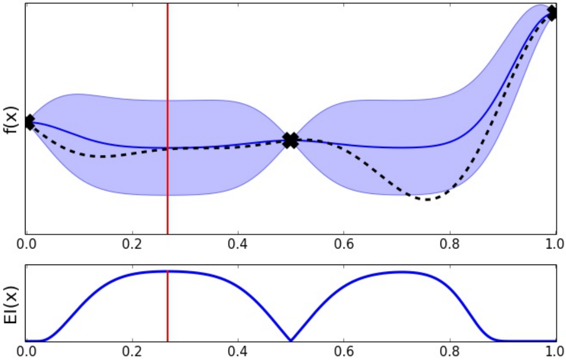


Expected Improvement: $E_x[\max(0, f(x_{min}) - f(x))]$

Tools: Bayesian Optimization

BO: Heuristic to reduce the number of evaluations in optimization problems [Mockus, 1978; Snoek et al., 2012].

Example: $x^* = \arg\min_{[0,1]} f(x)$?

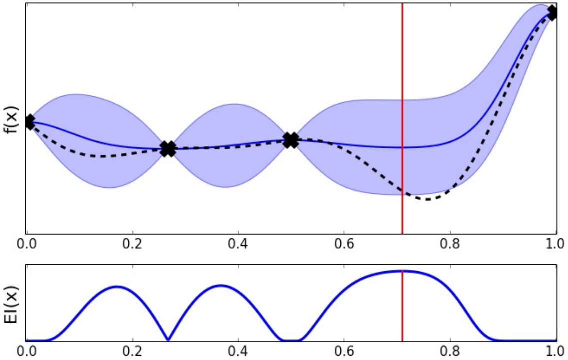


Expected Improvement: $E_x[\max(0, f(x_{min}) - f(x))]$

Tools: Bayesian Optimization

BO: Heuristic to reduce the number of evaluations in optimization problems [Mockus, 1978; Snoek et al., 2012].

Example: $x^* = \arg\min_{[0,1]} f(x)$?

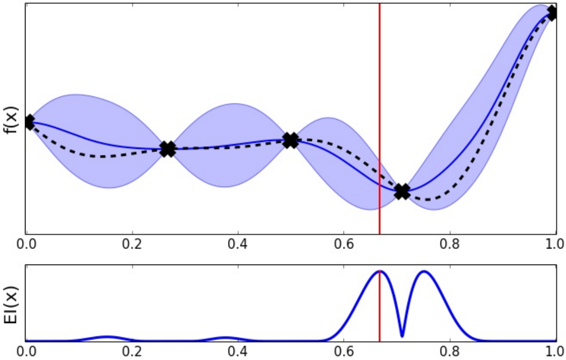


Expected Improvement: $E_x[\max(0, f(x_{min}) - f(x))]$

Tools: Bayesian Optimization

BO: Heuristic to reduce the number of evaluations in optimization problems [Mockus, 1978; Snoek et al., 2012].

Example: $x^* = \arg\min_{[0,1]} f(x)$?

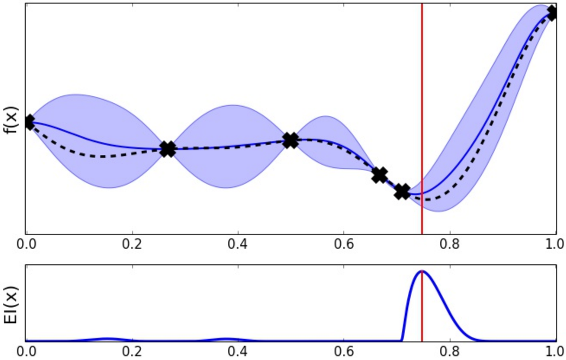


Expected Improvement: $E_x[\max(0, f(x_{min}) - f(x))]$

Tools: Bayesian Optimization

BO: Heuristic to reduce the number of evaluations in optimization problems [Mockus, 1978; Snoek et al., 2012].

Example: $x^* = \arg\min_{[0,1]} f(x)$?

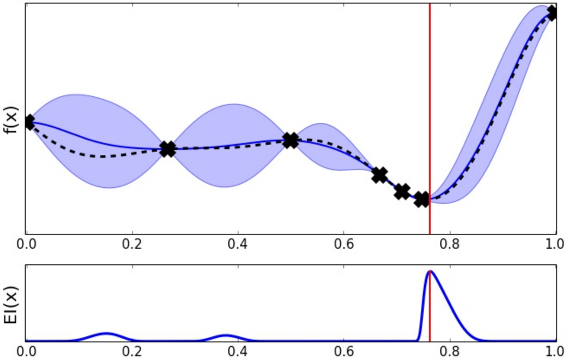


Expected Improvement: $E_x[\max(0, f(x_{min}) - f(x))]$

Tools: Bayesian Optimization

BO: Heuristic to reduce the number of evaluations in optimization problems [Mockus, 1978; Snoek et al., 2012].

Example: $x^* = \arg\min_{[0,1]} f(x)$?

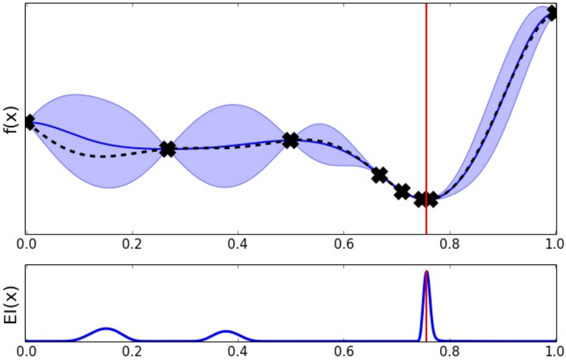


Expected Improvement: $E_x[\max(0, f(x_{min}) - f(x))]$

Tools: Bayesian Optimization

BO: Heuristic to reduce the number of evaluations in optimization problems [Mockus, 1978; Snoek et al., 2012].

Example: $x^* = \arg\min_{[0,1]} f(x)$?

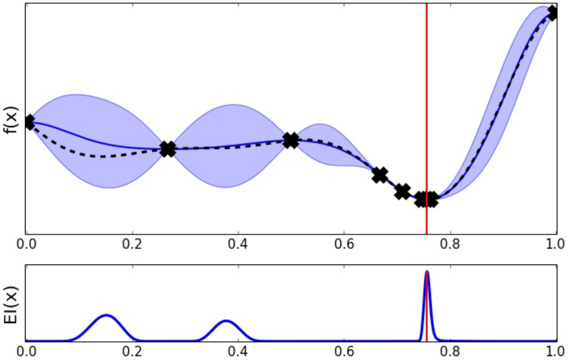


Expected Improvement: $E_x[\max(0, f(x_{min}) - f(x))]$

Tools: Bayesian Optimization

BO: Heuristic to reduce the number of evaluations in optimization problems [Mockus, 1978; Snoek et al., 2012].

Example: $x^* = \arg\min_{[0,1]} f(x)$?



Expected Improvement: $E_x[\max(0, f(x_{min}) - f(x))]$

How to design a synthetic gene?

A good model is crucial

Gene sequence features → protein production efficiency.

Bayesian Optimization principles for gene design

do:

1. Build a GP model as an emulator of the cell behavior.
2. Obtain a set of gene design rules (features optimization).
3. Design one/many new gene/s coherent with the design rules.
4. Test genes in the lab (get new data).

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

Model as an emulator of the cell behavior

Model inputs

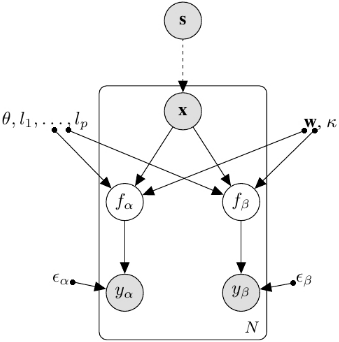
Features ($\mathbf{x}_i$) extracted gene sequences ($\mathbf{s}_i$): codon frequency, cai, gene length, folding energy, etc.

Model outputs

Translation and trasncriptions rates $\mathbf{f} := (f_\alpha, f_\beta)$.

Model type

Multi-output Gaussian process $\mathbf{f} \approx \mathcal{GP}(\mathbf{m}, \mathbf{K})$ where $\mathbf{K}$ is a corregionalization covariance for the two-output model (+ SE with ARD).



The correlation in the outputs help!

Obtaining optimal gene design rules

Maximize the averaged Expected improvement for both outputs [Swersky et al. 2013]

$$\alpha(\mathbf{x}) = \bar{\sigma}(\mathbf{x})(-u\Phi(-u) + \phi(u))$$

where $u = (y_{max} - \bar{m}(\mathbf{x}))/\bar{\sigma}(x)$ and

$$\bar{m}(\mathbf{x}) = \frac{1}{2}\sum_{l=\alpha,\beta}\mathbf{f}_*(\mathbf{x}),\ \bar{\sigma}^2(\mathbf{x}) = \frac{1}{2^2}\sum_{l,l'=\alpha,\beta}(\mathbf{K}_*(\mathbf{x},\mathbf{x}))_{l,l'}.$$

A batch method is used when several experiments can be run in parallel

Designing new genes coherent with the optimal design rules

Simulating-matching approach:

1. Simulate genes ‘coherent’ with the target (same aminoacids).
2. Extract features.
3. Rank synthetic genes according to their similarity with the ‘optimal’ design rules.

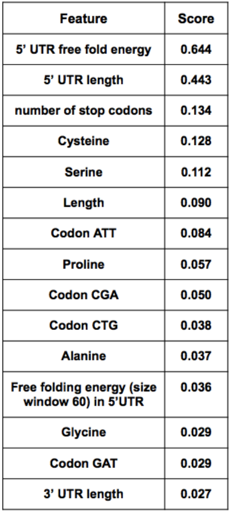
Ranking criterion: $eval(\mathbf{s}|\mathbf{x}^\star) = \sum_{j=1}^{p} w_j |\mathbf{x}_j - \mathbf{x}_j^\star|$

- $\mathbf{x}^\star$: optimal gene design rules.
- $\mathbf{s}$, $\mathbf{x}_j$ generated ‘synonyms sequence’ and its features.
- w_j: weights of the p features (inverse lengthscales of the model covariance).

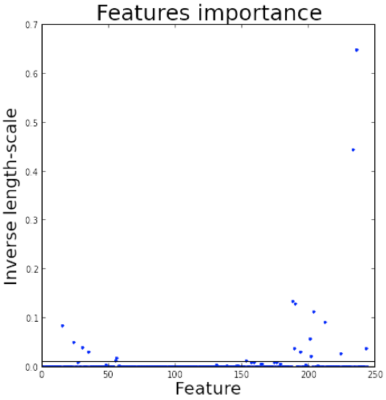
Experiments

- Optimization gene designs in mammalian cells.
- Dataset in Schwanhausser et al. (2011) for 3810 genes rates. Sequences were extracted from http:wet-labpic/www.ensembl.org.
- 250 features involving 5'UTR, 3'UTR and coding region.
- Gaussian process with ARD and coregionalized outputs.
- Selection of 10 random difficult-to-express genes (average log ratio < 1.5).
- 10,000 random 'synonyms sequences' generated from each gene.

Features importance



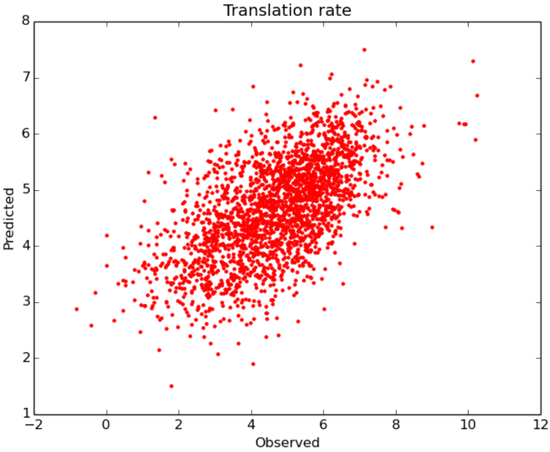
Feature	Score
5’ UTR free fold energy	0.644
5’ UTR length	0.443
number of stop codons	0.134
Cysteine	0.128
Serine	0.112
Length	0.090
Codon ATT	0.084
Proline	0.057
Codon CGA	0.050
Codon CTG	0.038
Alanine	0.037
Free folding energy (size window 60) in 5’UTR	0.036
Glycine	0.029
Codon GAT	0.029
3’ UTR length	0.027



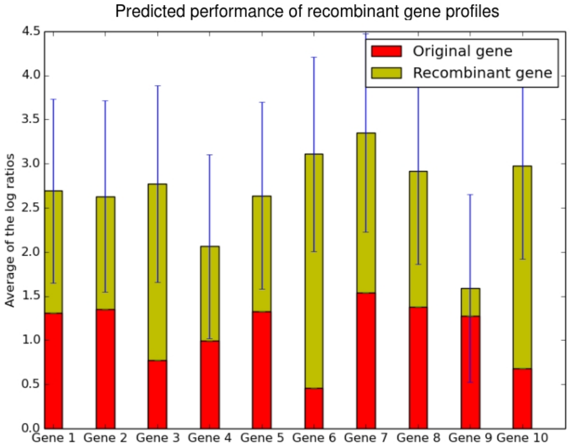
A few number of features are relevant

Predictions

The model is able to predict translation rates:



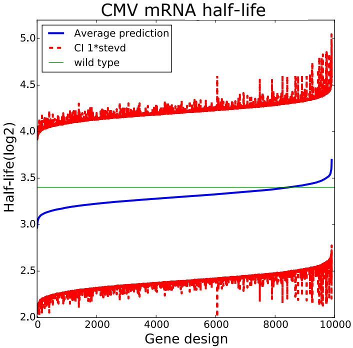
Results for 10 low-expressed genes

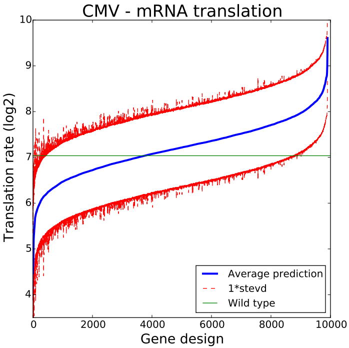


Results from simulation: currently testing the results in the lab!

Results for the CMV-pp65 Recombinant Protein

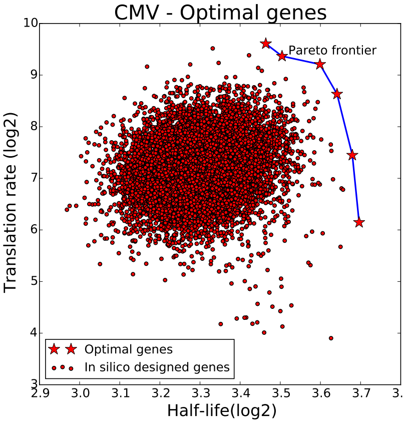
Alternative model: translation rates + mRNA half life.





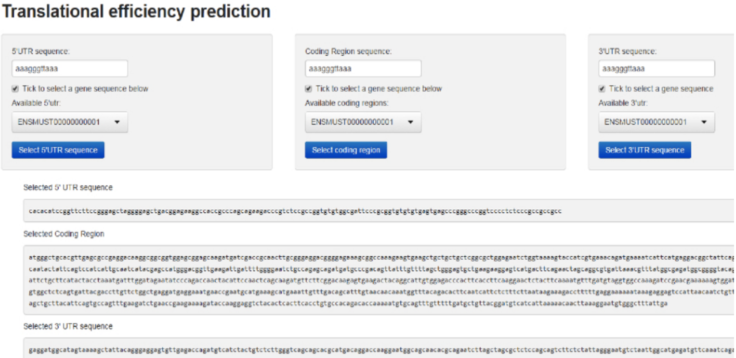
Predicted results for the 10,000 simulated sequences.

Results for the CMV-pp65 Recombinant Protein



Multi-objective optimization problem.

Prediction/design web tool



Translational efficiency prediction

5'UTR sequence:

aaagggttaaa

☑ Tick to select a gene sequence below

Available 5'utr:

ENSMUST00000000001

Coding Region sequence:

aaagggttaaa

☑ Tick to select a gene sequence below

Available coding regions:

ENSMUST00000000001

3'UTR sequence:

aaagggttaaa

☑ Tick to select a gene sequence

Available 3'utr:

ENSMUST00000000001

Selected 5' UTR sequence

```
cacacatccggttcttccgggagctaggggagctgacggagaaggccaccgcccagcagaagacccgtctccgccggtgtgtggcgattcccgcggtgtgtgtgagtgagcccgggcccggtcccctctcccgccgccgcc
```

Selected Coding Region

```
atgggctgcacgttgagcgccgaggacaaggcggcggtggagcggagcaagatgatcgaccgcaacttgcgggaggacggggagaaagcggccaaagaagtgaagctgctgctgctcggcgctggagaatctggtaaaagtaccatcgtgaaacagatgaaaatcattcatgaggacggctattcag
caatactattcagtccatcattgcaatcatacgagccatgggacggttgaagattgattttggggaatctgccagagcagatgatgcccgacagttatttgttttagctgggagtgctgaagaaggagtcatgacttcagaactagcaggcgtgattaaacgtttatggcgagatggcggggtacag
attctgcttcatactacctaaatgatttggatagaatatcccagaccaactacattccaactcagcaagatgttcttcggacaagagtgaagactacaggcattgtggagacccacttcaccttcaaggaactctacttcaaaatgtttgatgtaggtggccaaagatccgaacgaaaaagtggat
gtggctctcagtgattacgaccttgttctggctgaggatgaggaaatgaaccgaatgcatgaaagcatgaaattgtttgacagcatttgtaacaacaaatggtttacagacacttcaatcattctctttcttaataagaaagaccttttgaggaaaaaataaagaggagtccattaacaatctgtt
agctgcttacattcagtgccagtttgaagatctgaaccgaagaaaagataccaaggaggtctacactcacttcacctgtgccacagacaccaaaaatgtgcagtttgtttttgatgctgttacggatgtcatcattaaaaacaacttaaaggaatgtgggctttattga
```

Selected 3' UTR sequence

```
gaggatggcatagtaaaagctattacagggaggagtgttgagaccagatgtcatctactgtctcttgggtcagcagcacgcatgacaggaccaaggaatggcagcaacacgcagaatcttagctagcgctctccagcagtcttctctattagggaatgtctaattggcatgagatgttcaaatcag
```

Web app for gene design based on

- GPy (https://github.com/SheffieldML/GPy).
- GPyOpt (https://github.com/SheffieldML/GPyOpt).

Final remarks

- Bayesian optimization is a promising technique to design synthetic genes: reduces drastically the number of needed experiments.

- Very important aspect of the problem → to have a good surrogate model for the cell behavior.

- Currently, working out a model with more outputs, such as the protein stability and cell survival.

- Alternative approach: focus on the direct optimization of the sequences. Combinatorial problem.