Synthetic Biology, 2018, 3(1): ysy010

doi: 10.1093/synbio/ysy010
Advance Access Publication Date: 26 June 2018
Research article

Hybrid semiparametric systems for quantitative sequence-activity modeling of synthetic biological parts

Rui M.C. Portela[1], Moritz von Stosch[2], and Rui Oliveira[1,*]

[1]REQUIMTE/LAQV, Departamento de Química, Faculdade de Ciências e Tecnologia Universidade Nova de Lisboa, Caparica, Portugal and [2]CEAM Faculty of Science, Agriculture and Engineering, Newcastle University, Newcastle upon Tyne, UK

*Corresponding author: E-mail: rmo@fct.unl.pt

Abstract

Predicting the activity of modified biological parts is difficult due to the typically large size of nucleotide sequences, resulting in combinatorial designs that suffer from the “curse of dimensionality” problem. Mechanistic design methods are often limited by knowledge availability. Empirical methods typically require large data sets, which are difficult and/or costly to obtain. In this study, we explore for the first time the combination of both approaches within a formal hybrid semiparametric framework in an attempt to overcome the limitations of the current approaches. Protein translation as a function of the 5’ untranslated region sequence in *Escherichia coli* is taken as case study. Thermodynamic modeling, partial least squares (PLS) and hybrid parallel combinations thereof are compared for different data sets and data partitioning scenarios.
The results suggest a significant and systematic reduction of both calibration and prediction errors by the hybrid approach in comparison to standalone thermodynamic or PLS modeling. Although with different magnitudes, improvements are observed irrespective of sample size and partitioning method. All in all the results suggest an increase of predictive power by the hybrid method potentially leading to a more efficient design of biological parts.

Key words: quantitative sequence-activity modeling; hybrid semiparametric systems; standard biological parts; ribosome binding site (RBS); *Escherichia coli*.

1. Introduction

Mathematical modeling is a fundamental tool in systems and synthetic biology for better understanding biological systems and to improve design efficiency (1, 2). A class of problems deals with the design of nucleotide sequences of standard biological parts such as promoters, riboswitches, ribosome binding sites (RBSs) and other DNA/RNA devices. Due to the large space of potential nucleotide sequences, experimental screening of the whole design space is impractical. As an illustrative example, the 5’ untranslated region (UTR) in *Saccharomyces cerevisiae* was the target design in a previous study by Dvir *et al.* (3) A large library of mutants was generated by randomly mutating the 10 Bp that precede the start codon. Even using high-throughput techniques, only 0.2% out of 10^6 possible sequences, were experimentally screened. Rational design, aided by mathematical models, is thus essential to saving time and resources.

Mechanistic modeling is the method of choice for biological part design (4) with several successful examples published mainly for *Escherichia coli*, the model system with more mechanistic insight to date. Brewster *et al.* (5) developed a thermodynamic transcription initiation model focused on the −10 and −35 *E. coli* promoter regions and its affinity to RNA polymerase. Synthetic promoters designed with the aid of this model were

Submitted: 7 March 2018; **Received (in revised form):** 21 May 2018; **Accepted:** 11 June 2018

shown to increase protein expression by 3-fold in comparison to natural promoters. Salis *et al.* (6) proposed a protein translation model as a function of the 5'UTR sequence in *E. coli* assuming translation initiation as the limiting step (RBS calculator model v1.0). The free Gibbs energy associated with the formation of the mRNA-ribosome complex (determined from five molecular interactions) is the key parameter that controls the amount of protein expressed. Na *et al.* (7) proposed an alternative kinetic model focusing on three molecular interactions only and using the ordinary differential equations formalism to describe the transitions between ribosome binding states. In another study, Amman *et al.* (8) included in their translation initiation model the interactions between small non-coding RNAs and the reporter protein mRNA. Borujeni *et al.* (9) detailed the computation of the free Gibbs energy associated with the standby sequence, which further improved Salis *et al.* (6) model (RBS calculator model v2.0). Later on, Borujeni and Salis (10) concluded that folded RNA structures may not have enough time to fold inside the cell, creating a non-equilibrium effect termed as "Ribosome Drafting" (RBS calculator model v2.1). In such cases, the thermodynamic modeling framework fails to deliver accurate predictions (10).

Alternatively to mechanistic modeling, empirical methods have also been applied with two main objectives: (i) classification of nucleotide sequences and (ii) regression analysis of biological activity as function of the respective nucleotide sequence. González-Díaz *et al.* (11) used Markov molecular negentropies to describe the secondary structure of putative RNA molecules and used such predictions to identify mycobacterial promoters. Tavares *et al.* (12) performed a comparative study on the performance of 31 machine learning methods (hidden Markov models and different topologies of neural networks and decision trees) to classify *E. coli* promoters. Li *et al.* (13) used a mixture of Gaussian models to predict translation initiation sites in yeasts. An integrated Bayesian model was used to identify and predict several features of transcription factor binding sites (like number, position and composition) in several yeasts promoters (14). Artificial neural networks were used for both classification and regression problems. In Zuo and Li (15), an encoding method based on DNA helical parameters was adopted to predict DNA curvature and transcription rate in *E. coli*. Jonsson *et al.* (16) used partial least squares (PLS) with binary encoding to design two synthetic *E. coli* promoters. Liang *et al.* (17) compared the performance of support vector machines and PLS to predict the transcription rate of *E. coli* promoters. Ran and Higgs (18) developed a statistical test, based on maximum likelihood and codon adaptation index, to assess the significance and the strength of codon bias on transcription elongation speed and accuracy.

In previous studies, sequence-activity modeling either follows a parametric paradigm, where models have a fixed structure inspired by knowledge, or follow a non-parametric approach, where model structure is derived exclusively from data. In this article, we explore the combination of both approaches in hybrid semiparametric systems for sequence-activity modeling. The main advantage of the semiparametric over the parametric or non-parametric frameworks lies in that it broadens the knowledge base to solve a complex problem. In a recent review paper, several areas of application of hybrid modeling have been outlined, ranging from chemical, biological to mechanical engineering (19). Several semiparametric systems biology studies have been published following the constraints-based formalism, namely hybrid metabolic flux analysis (20, 21) and hybrid metabolic pathway analysis (22, 23). Others have addressed dynamic modeling of biological systems either following the differential equations formalism (24, 25) or time series analysis (26, 27). Despite some progress at the systems biology front, applications to synthetic biology are still largely absent in the literature. In designing biological parts, it is unlikely that all relevant processes can be fully described by a mechanistic (parametric) approach. Purely empirical (non-parametric) modeling approaches are often limited by the availability of sufficient experimental data. Opting for the one or the other framework will invariably promote reductionism. On the contrary, the "complementary" use of both types of resources provides more comprehensive descriptions of the biological system at hand. To illustrate this concept we use the 5'UTR sequence in *E. coli* as case study. The starting point is the RBS calculator model v1.0 published by Salis *et al.* (6) and respective data set. This version of the model is ideal to showcase hybrid modeling because the data are very limited (only 132 mRNA sequences) and the thermodynamic model (TM) still has room for improvement. Afterwards, the hybrid approach is also applied to the larger data set of RBS model calculator v2.1 (10).

2. Materials and methods

2.1 RNA sequences and protein expression data

The data set of RBS calculator model v1.0 (6) was adopted for model development and benchmarking. The data set contains 132 modifications of the 5'UTR sequence and respective green fluorescence levels obtained in transformed *E. coli* DH10B strains. The data were divided into a model identification partition and a test partition. The former served to identify model structure and respective parameter values. The latter served to assess the model predictive power. Three different data partitioning scenarios were studied:

Partition R: Random selection of 67% of sequences for model identification and 33% of sequences for model testing. The sequences were randomly selected from the uniform distribution. The procedure is repeated 100 times yielding 100 different models to eliminate data sampling bias.

Partition E33: Heuristic selection of 67% of sequences with lowest protein expression for model identification and 33% of sequences with highest protein expression for model testing.

Partition E67: Heuristic selection of 33% of sequences with lowest protein expression for model identification and 67% of sequences with highest protein expression for model testing.

2.2 Thermodynamic modeling

The equilibrium TM proposed by Salis *et al.* (6) is the first module of the hybrid structure (Figure 1). It assumes initiation as the limiting step in the protein translation process. The key thermodynamic parameter is ΔG_{TOT}, representing the difference in Gibbs free energy between the initial mRNA folded state and the final 30S pre-initiation complex. The ΔG_{TOT} accounts for five terms:

$$\Delta G_{TOT} = \Delta G_{mRNA:rRIB} + \Delta G_{START} + \Delta G_{SPACING} - \Delta G_{STANDBY} - \Delta G_{mRNA} \quad (1)$$

ΔG_{mRNA} is the mRNA Gibbs free energy when it is not interacting with any other molecule. It may be viewed as the energy required to unfold it, so that it becomes accessible to the rRNA. It is calculated using a portion of the mRNA sequence surrounding the start codon. After unfolding, mRNA hybridizes with

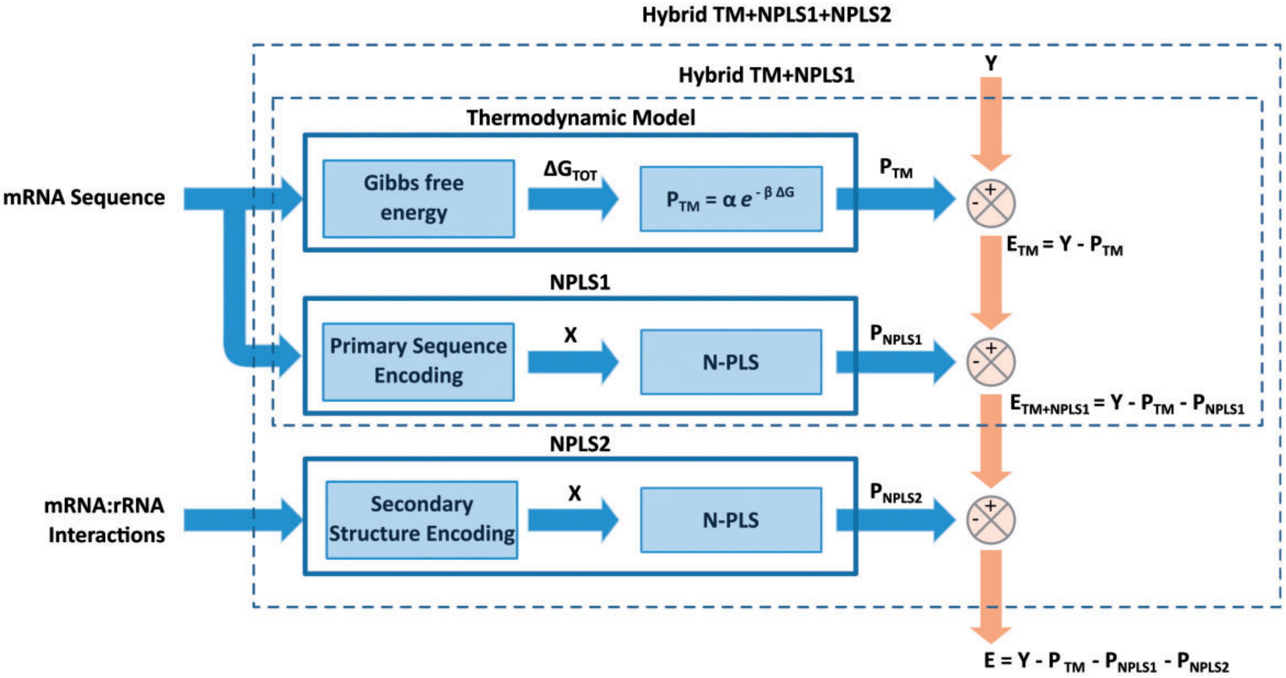

Figure 1. Parallel hybrid model structure describing protein expression (**Y**) as function of mRNA sequence. The first module is the TM that predicts protein expression as function of ΔG_{TOT}. The second module (NPLS1) is an N-PLS model that runs in tandem with the TM and extracts information from the TM residuals as function of mRNA primary structure. The third module (NPLS2) is an N-PLS model that runs in tandem with TM+NPLS1 and extracts information from the TM+NPLS1 residuals as function of the possible mRNA:rRNA interactions.

rRNA ($\Delta G_{mRNA:rRIB}$). To calculate $\Delta G_{mRNA:rRIB}$, all possible interactions between the mRNA and rRNA were first computed. The interaction that minimizes the sum of $\Delta G_{mRNA:rRIB}$ with $\Delta G_{SPACING}$ was chosen. The $\Delta G_{SPACING}$ can be seen as an empirical penalty (relationship estimated experimentally (6)) to be applied when the distance between the mRNA-rRNA interaction and start codon is not optimal (too far away or too close). ΔG_{START} accounts for the interaction between the mRNA sequence at the start codon and the respective tRNA (calculated using these two sequences with three nucleotides each). $\Delta G_{STANDBY}$ is the Gibbs free energy needed to unfold any mRNA secondary structure generated after the rRNA hybridization that blocks the protein synthesis initiation (this term is calculated by subtracting the energies of two states: one allowing the positions surrounding the selected mRNA-rRNA interaction to have a secondary structure and another preventing it).

The calculation of the Gibbs free energies (ΔG_{START}, $\Delta G_{mRNA:rRIB}$, $\Delta G_{STANDBY}$ and ΔG_{mRNA}) is direct (neither additional fitting nor parameters are needed). To this end, we used the same tool as in the original study (6): NUPACK with Mfold3.0 RNA energy parameters (28–30).

Finally, the expressed protein level (P_{TM}) is a function of the respective mRNA secondary structures, represented by ΔG_{TOT}, as follows:

$$P_{TM} = \alpha t e^{-\beta \Delta G_{TOT}} \qquad (2)$$

with α an empirical calibration parameter, t the cultivation time and β the Boltzmann factor that accounts for translation-independent parameters, such as the DNA copy number, the promoter's transcription rate and the mRNA stability.

The identification of this model was performed by linear regression of the natural logarithm of the measured protein expression ($\ln(P_{MES})$) against ΔG_{TOT} over the model identification data, with P_{MES} the measured reporter protein (RFP1) fluorescence. The MATLAB function "fit" was adopted implementing a linear least squares algorithm. The slope corresponds to the Boltzmann constant, β, while the intercept corresponds to $\ln(\alpha\, t)$.

2.3 N-PLS modeling

N-way partial least squares (N-PLS) is a well-known multivariate regression method with data factorization, in that a target matrix, **Y**, is linearly regressed against an input (regressor) matrix **X** of many possibly collinear variables (31). The most used method is the two-way PLS. N-PLS is an extension of the two-way PLS by taking **X** with $N > 2$ dimensions, with N the number of dimensions of **X**. The **X** and **Y** matrices are decomposed in *Fac* latent variables. In each decomposition step, a scores matrix (**t**) and $N-1$ weight matrices are calculated. In the case of 3-way PLS, the decomposition of a 3 D **X** ($I \times J \times K$) gives a scores vector **t**, two weight vectors $\mathbf{w}^J$ (for dimension 2) and $\mathbf{w}^K$ (for dimension 3) and a residuals matrix, **E**:

$$\mathrm{X}_{ijk} = \mathrm{t}_i\, \mathrm{w}_J^J \mathrm{w}_K^K + \mathrm{E} \qquad (3)$$

The indexes i, j and k denote the position in dimensions 1, 2 and 3, respectively. The decomposition is performed in the sense of maximizing the covariance between **X** and **Y**, as follows:

$$\max_{w^J w^K}\left(\sum\nolimits_{i=1}^{I} \mathrm{t}_i \mathrm{y}_i \middle| \mathrm{t} = \sum\nolimits_{j=1}^{J} \sum\nolimits_{k=1}^{I} \mathrm{x}_{ijk} \mathrm{w}_j^J \mathrm{w}_k^K \right) \qquad (4)$$

Modeling multi-dimensional data sets by bilinear PLS implies that multi-dimension input matrix **X** is unfolded into a 2D representation. When comparing both approaches, N-PLS presents clear advantages in terms of input decomposition stabilization, since fewer parameters are needed, resulting in a more robust, parsimonious and interpretable final model. Bilinear PLS is more flexible, usually performing better in the calibration partition, but being prone to overfitting when the number of input variables is too large (31), which is clearly the case of DNA sequences. For this reason, we opted in this work for the N-PLS MATLAB implementation described in (32). Two N-PLS models were developed as described below.

Nucleotide sequences encoding

N-PLS requires the input/output data to be numeric rather than symbolic. Six different encoding methods to translate the symbolic nucleotide sequences into a numerical representation were compared. A numerical representation, consisting of a vector of numerical states, was assigned to each nucleotide as follows:

Encoding 1: Adenine (0, −1), Cytosine (−1, 0), Guanine (1, 0), Uracil (0, 1) and blank space (0, 0);

Encoding 2: Adenine (−1, 0), Cytosine (0, 1), Guanine (1, 0), Uracil (0, −1) and blank space (0, 0);

Encoding 3: Adenine (1, 0), Cytosine (−1, 0), Guanine (0, −1), Uracil (0, 1) and blank space (0, 0);

Encoding 4: Adenine ($\sin(\pi/6)$, $-\sin(\pi/3)$), Cytosine ($\sin(\pi/6)$, $\sin(\pi/3)$), Guanine ($\sin(\pi/3)$, $-\sin(\pi/6)$), Uracil ($\sin(\pi/3)$, $\sin(\pi/6)$) and blank space (0, 0);

Encoding 5: Adenine (1, 0, 0, 0), Cytosine (0, 0, 0, 1), Guanine (0, 1, 0, 0), Uracil (0, 0, 1, 0) and blank space (0, 0, 0, 0);

Encoding 6: Adenine (−3.9505, 4.0764, −1.1507, 1.24226), Cytosine (4.3677, 1.0541, 1.5173, 3.2084), Guanine (−2.7552, −4.8467, 1.1540, 1.4321), Uracil (1.9163, −1.1601, −4.9190, −1.7917) and blank space (0, 0, 0, 0);

A detailed description and examples of applications can be found elsewhere: *Encoding* 1 to 4 (33), *Encoding* 5 (16) and *Encoding* 6 (17).

NPLS1: primary structure N-PLS model

NPLS1 is a 3-way N-PLS describing protein titer as function of primary mRNA sequence. The mRNA sequences were trimmed to 70 Bp (35 Bp upstream and downstream of the start codon, i.e. the same sequences used to calculate ΔG_{mRNA}). Since some of the mRNA molecules are shorter, they were first aligned by their start codon and then filled with blank spaces up to 75 Bp. Afterwards, one of the previously described encoding methods was applied. The encoding resulted into a 3D **X** matrix ($np \times nb \times ne$), with $np = 132$ the number of mRNA sequences, $nb = 70$ the maximum sequence length (in base pairs) and $ne = 4$ or $ne = 2$ (depending on the encoding method) the number of values representing a single nucleotide. **X** was then autoscaled (subtracting the mean and dividing by the standard deviation) column wise. The target vector **Y** was the protein expression data (measured reporter protein fluorescence). **Y** was transformed by applying the natural logarithm to obtain more normal distributed values in **Y**. This transformation is in agreement with eq. (2), where the N-PLS is predicting a free energy-like quantity. The **Y** matrix was then autoscaled. The normalized **X** and **Y** were subject to N-PLS regression using the MATLAB implementation described in (32). The optimal number of latent variables was determined by the leave-one-out method (34). The number of NPLS1 parameters, *npar*, is equal to the optimal number of latent variables, Fac_{NPLS1}.

NPLS2: mRNA:rRNA interactions N-PLS model

NPLS2 describes protein expression as function of the mRNA standby sequence. The mRNA standby sequences (used before to calculate the $\Delta G_{\text{STANDBY}}$), comprising all base pairs upstream of the mRNA-rRNA interaction locus, are taken as indirect measure of the mRNA-rRNA interaction formed. All possible mRNA-rRNA interactions were computed using the *subopt* function of NUPACK (28–30), in order to determine the standby sequences for a given mRNA molecule. These sequences were organized in a 3D **X** matrix ($np \times nb \times (ns \times ne)$) with $np = 132$ the number of mRNA molecules, $nb = 20$ the maximum standby sequence length (in base pairs), and $ns \times ne$ (third dimension) the number of standby sequences (ns) multiplied by the encoding length ($ne = 2$ or $ne = 4$ depending on the method). It should be noted that the different mRNA molecules generate a different number of possible mRNA-rRNA interactions and respective standby sequences, e.g. an mRNA molecule with a consensus Shine–Dalgarno sequence will bind strongly to the rRNA and generate fewer interactions. On the other hand, a degenerated binding sequence allows many different mRNA-rRNA interactions. The **X** matrix was autoscaled column wise. The protein expression data **Y** was as before transformed by applying the natural logarithm and then autoscaled. The normalized **X** and **Y** were subject to N-PLS regression using MATLAB implementation described in (32). The optimal number of latent variables was determined by the leave-one-out method (34). The number of NPLS2 parameters, *npar*, is equal to the optimal number of latent variables, Fac_{NPLS2}.

2.4 Hybrid semiparametric modeling

The structure of the hybrid model is represented in Figure 1, consisting of three parallel modules. The first module is the TM that predicts protein expression as function of ΔG_{TOT}. The second module is an N-PLS model (NPLS1) that runs in tandem with the TM and extracts information from the TM residuals as function of mRNA primary structure. The third module is an N-PLS model (NPLS2) that runs in tandem with TM+NPLS1 and extracts information from the TM+NPLS1 residuals as function of mRNA-rRNA interactions. Therefore, the hybrid model decomposes the target (measured) protein vector **Y** in four terms:

$$\mathrm{Y} = \mathrm{P}_{TM} + \mathrm{P}_{NPLS1} + \mathrm{P}_{NPLS2} + \mathrm{E} \tag{5}$$

The first three terms represent the contribution of the three modules (TM, NPLS1 and NPLS2, respectively) to the prediction of **Y**. The vector **E** is the final hybrid model residuals. All terms in eq. 5 are normalized in the same way, i.e. by applying natural logarithm and then by autoscaling. The hybrid model identification was performed in three consecutive steps:

Step 1: Identification of the TM module. The TM module has priority to describe observations and is thus the first to be fitted to the data. The method to identify the TM module is exactly the same as the standalone TM (see Section 2.1). In the end of this step, $\mathbf{P}_{\text{TM}}$ is identified as function of $\Delta G_{\text{TOT.}}$

Step 2: Identification of the TM+NPLS1 structure. Firstly, the TM module residuals, $\mathbf{E}_{\text{TM}}$, are calculated:

$$\mathrm{E}_{TM} = \mathrm{Y} - \mathrm{P}_{TM}. \tag{6}$$

Then, NPLS1 is set to identify patterns from the TM residuals. The method is the same as described above for standalone NPLS1 except that the target output is $\mathbf{E}_{\text{TM}}$ rather than **Y**. Also,

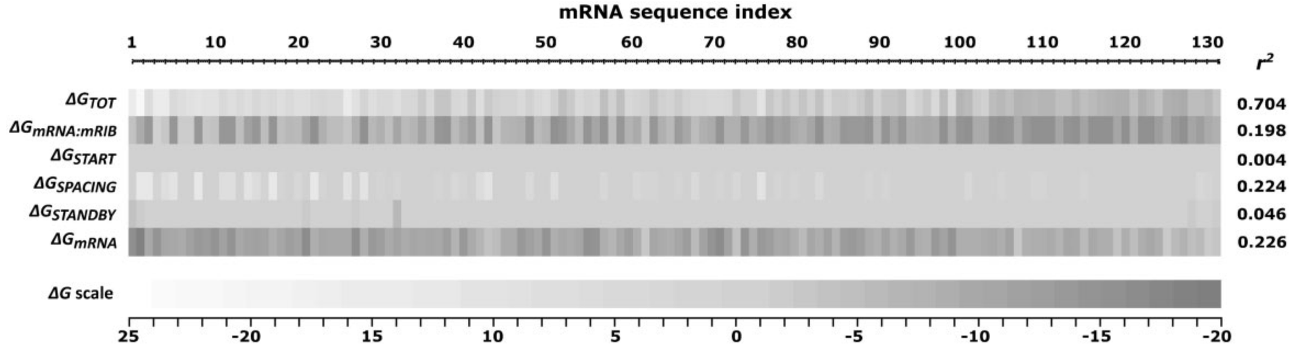

Figure 2. Heat map representing the free Gibbs energy for each of the 132 RNA sequences sorted from high to low protein fluorescence values. The columns refer two ΔG_{TOT} and to the individual $\Delta G_{mRNA:rRIB}$, ΔG_{START}, $\Delta G_{SPACING}$, $\Delta G_{STANDBY}$ and ΔG_{mRNA}. The correlation coefficient (r^2) on the right side refers to the correlation of measured protein fluorescence in relation to calculated free Gibbs energy. The three steps with highest correlation are $\Delta G_{mRNA:rRIB}$ ($r^2 = 0.20$), $\Delta G_{SPACING}$ ($r^2 = 0.22$) and ΔG_{mRNA} ($r^2 = 0.23$).

the optimal number of latent variables was not determined by leave-one-out. Rather, the model identification data set was divided into a calibration subset (67% of data points) and a validation subset (33% of points). The validation subset comprised the data points with highest TM residuals (i.e. highest values in $\mathbf{E}_{TM}$). This ensures the selection of the optimal number of latent variables that maximizes predictive power of the TM residuals. At the end of this step, NPLS1 calculates $\mathbf{P}_{NPLS1}$ as function of mRNA primary structure. The hybrid TM+NPLS1 output is given by $\mathbf{P}_{TM} + \mathbf{P}_{NPLS1.}$

Step 3: Identification of the TM+NPLS1+NPLS2 structure. Firstly, the TM+NPLS1 residuals are calculated:

$$E_{TM+NPLS1} = Y - P_{TM} - P_{NPLS1} \tag{7}$$

Then NPLS2 is identified following the same method previously described for standalone NPLS2 except that the target output is $\mathbf{E}_{TM+NPLS1}$ instead of $\mathbf{Y}$. The optimal number of latent variables was determined as in Step 2. At the end of this step, NPLS2 calculates P_{NPLS2} as function of mRNA secondary structure. The hybrid TM+NPLS1+NPLS2 output is given by $\mathbf{P}_{TM} + \mathbf{P}_{NPLS1} + \mathbf{P}_{NPLS2}$.

2.5 Model performance criteria

Three different metrics were employed for model performance assessment, namely the mean squared error (MSE) (Eq. 8), explained variance (Var., %), (Eq. 9) and the Akaike Information Criterion with second order bias correction (AIC$_c$), (Eq. 10):

$$MSE = \frac{1}{n}E^T E \tag{8}$$

$$Var(\%) = 100\left(1 - \frac{E^T E}{Y^T Y}\right) \tag{9}$$

$$AIC_c = n\ln(MSE) + 2k + \frac{2k(k+1)}{n-k-1} \tag{10}$$

with n the number of data points, $\mathbf{E}$ a vector of model residuals, k the number of model parameters given by:

$$k = 2 + Fac_{NPLS1} + Fac_{NPLS2} \tag{11}$$

AIC accounts for an overparameterization penalty and is commonly used to discriminate between empirical model candidates and to select a parsimonious model (34).

3 Results and discussion

3.1 Standalone TM

Determination of Gibbs free energy and model fitting

Figure 2 represents the calculated free Gibbs energy parameters for each of the 132 mRNA sequences, sorted from low to high reporter protein fluorescence values. As previously shown by Salis *et al.* (6), measured reporter protein fluorescence is correlated with ΔG_{TOT} ($r^2 = 0.70$). ΔG_{TOT} is the Gibbs free energy variation between the folded mRNA and the assembled 30S preinitiation complex, accounting for five terms: $\Delta G_{mRNA:rRIB}$, ΔG_{START}, $\Delta G_{SPACING}$, $\Delta G_{STANDBY}$ and ΔG_{mRNA}. The correlation with individual ΔG terms is however much lower than with ΔG_{TOT}. The three individual ΔG values with highest correlation are $\Delta G_{mRNA:rRIB}$ ($r^2 = 0.20$), $\Delta G_{SPACING}$ ($r^2 = 0.22$) and ΔG_{mRNA} ($r^2 = 0.23$).

The TM (Eq. 2) was fitted to the calculated ΔG_{TOT} and measured fluorescence data, adopting the uniform data partitioning strategy (partition R), i.e. 67% of data points are randomly selected for fitting (Eq. 2), with the remaining 33% of data points used to assess predictive power. The procedure is repeated 100 times to eliminate data sampling bias. The results are shown in Table 1 (first row). The average Boltzmann constant among the 100 different trials was 0.37 ± 0.034 mol/kcal, which is slightly lower than the value reported by Salis *et al.* (6) of 0.45 ± 0.05 mol/kcal for two different data partitions. The average MSE was 0.29 for the identification data set and 0.31 for the test data set, showing that the prediction accuracy is comparable to the calibration accuracy. Nevertheless, the model systematically underpredicts the highest protein expression sequences, e.g. −36.57% for the top 5% of protein expression sequences and −61.17% for the top 1% of protein expression sequences (Table 1, first row).

Effect of data sparsity on predictive power

Data sparsity is common in sequence-activity modeling because the design space is very large and typically only a small number of sequences are experimentally screened. Here, the length of the design sequence is 35 Bp (mRNA sequence upstream of start

Table 1. Comparison of standalone thermodynamic and N-PLS models for three different data partitioning scenarios

Partition	Models	Identification				Test			Relative error	
		% Var	MSE	AIC_c	MSE/MSE_{TM}	% Var	MSE	MSE/MSE_{TM}	Top 5%	Top 1%
R	TM[a]	70.59	0.29		1	68.47	0.31	1	−36.57	−61.17
	NPLS1[b]	71.20	0.28	−94.74	0.97	29.02	0.72	2.32	−45.00	−21.82
	NPLS2[c]	29.43	0.71	−24.15	2.45	10.58	0.84	2.71	−80.94	−134.60
E33	TM[d]	68.14	0.30		1	56.59	0.49	1	−62.06	−81.47
	NPLS1[e]	63.30	0.34	−87.85	1.13	51.32	0.55	1.12	−56.62	3.53
	NPLS2[f]	15.5	0.78	−19.29	2.61	30.43	0.78	1.60	−85.66	−129.12
E66	TM[g]	85.86	0.25		1	<0	0.98	1	−98.32	−112.83
	NPLS1[h]	90.23	0.17	−71.03	0.68	<0	1.48	1.45	−39.61	21.50
	NPLS2[i]	10.50	1.59	22.41	6.36	7.46	0.56	0.57	−97.22	−101.83

Identification refers to performance metrics in the identification data partition. Prediction refers to performance metrics in the test data partition. Relative errors refer to the relative absolute deviation of model prediction and measurement for the top 5% and 1% protein expression sequences.

[a] $\beta = 0.37$.

[b] With encoding 5 and *Fac* = 3.

[c] With encoding 4 and *Fac* = 3.

[d] $\beta = 0.30$.

[e] With encoding 3 and *Fac* = 3.

[f] With encoding 6 and *Fac* = 1.

[g] $\beta = 0.25$.

[h] With encoding 5 and *Fac* = 3.

[i] With encoding 6 and *Fac* = 1.

codon) with 4^{35} possible combinations, of which only 132 sequences were experimentally screened. To assess the ability of the TM to extrapolate the high activity mRNA sequences, the E33 and E67 data partitioning scenarios were studied. The overall results are shown in Table 1 (fourth and seventh rows).

In the case of partition E33 (extrapolation of the 33% best sequences), the MSE of the identification data set is 0.30, but the MSE of the test data set increased to 0.49. The top 5% sequences are systematically underpredicted by −62.06%. In the case of partition E67 (extrapolating the 67% best sequences), the results degrade much further. The average MSE of the identification data set decreases to 0.25 while that of the test data increases to 0.98 and the top 5% sequences are systematically underpredicted by −98.32%. The model is clearly overfitting the identification data set and failing to predict the test data set.

Figure 3A plots predicted over measured protein expression for the data partition E33. Figure 3B and C shows the residuals distribution for the identification and test data sets, respectively. It may be observed, according to the Shapiro–Wilk normality test, that the residuals are normal for the data identification partition but this no longer holds for the test data partition. Moreover, it may be confirmed (Figure 3C) that model predictions are largely biased in the test partition (−0.62 mean and 0.32 standard deviation) in the sense of underprediction. These results suggest that data sparsity has a large negative impact in the ability of the TM to describe high expression sequences. Given the structural simplicity of the model, this can only mean that model assumptions do not fully represent the real system.

3.2 Standalone N-PLS regression

Standalone N-PLS regression was compared to the TM. Firstly, N-PLS regression of protein fluorescence as function of primary mRNA sequence (NPLS1 model) was studied. The same three data partitioning scenarios as for the TM were applied. The mRNA encoding method is an important factor, with encodings 1, 3 and 5 producing significantly better results than encodings 2, 4 and 6. The overall results are shown in Table 1 for the best encoding (second, fifth and eighth rows).

The second row of Table 1 summarizes the results for partition R. N-PLS describes the identification partition with average MSE of 0.28, very similar to the TM (MSE of 0.29). The description of the test partition is, however, much worse for the NPLS1 model (The MSE_{NPLS1}/MSE_{TM} ratio is $2.32 \gg 1$). In the case of partition E33 (Table 1, fifth row), NPLS1 shows a slightly worse but comparable performance to the TM in terms of data fitting flexibility (MSE ratio of 1.13 in the identification partition) and also predictive power (MSE ratio of 1.12 in the test partition). In the case of partition E67 (Table 1 eighth row), the NPLS1 model significantly improves the fitting power (MSE ratio of 0.68 in the identification partition) at cost of much higher prediction error (MSE ratio of 1.45 in the test partition). This result is typical of non-parametric identification in general (using N-PLS or other techniques). When non-parametric models are calibrated with less data there is a tendency for calibration error to decrease. The lower calibration error often reflects data overfitting resulting in a less representative model with lower predictive power.

A similar analysis was performed with NPLS2 model, whereby protein fluorescence is described as function of the standby sequence upstream of the mRNA-rRNA interaction locus. The results are shown in Table 1 (third, sixth and ninth rows) for the three different data partitioning scenarios. The previous observations regarding the effect of data partitioning and encoding methods for NPLS1 are generically valid for the NPLS2 model. The key result is the much lower explained variances (or much higher MSE) for the NPLS2 model in relation to the NPLS1 and to the TM. This is not surprising since the input to the NPLS2 model is restricted to mRNA-rRNA interactions, thus incomplete.

In the TM, the information content of the standby sequence is given by the $\Delta G_{STANDBY}$ term, which is zero for a large number of sequences (Figure 2). However, the explained variance of the identification and testing partitions are comparable. This suggests that

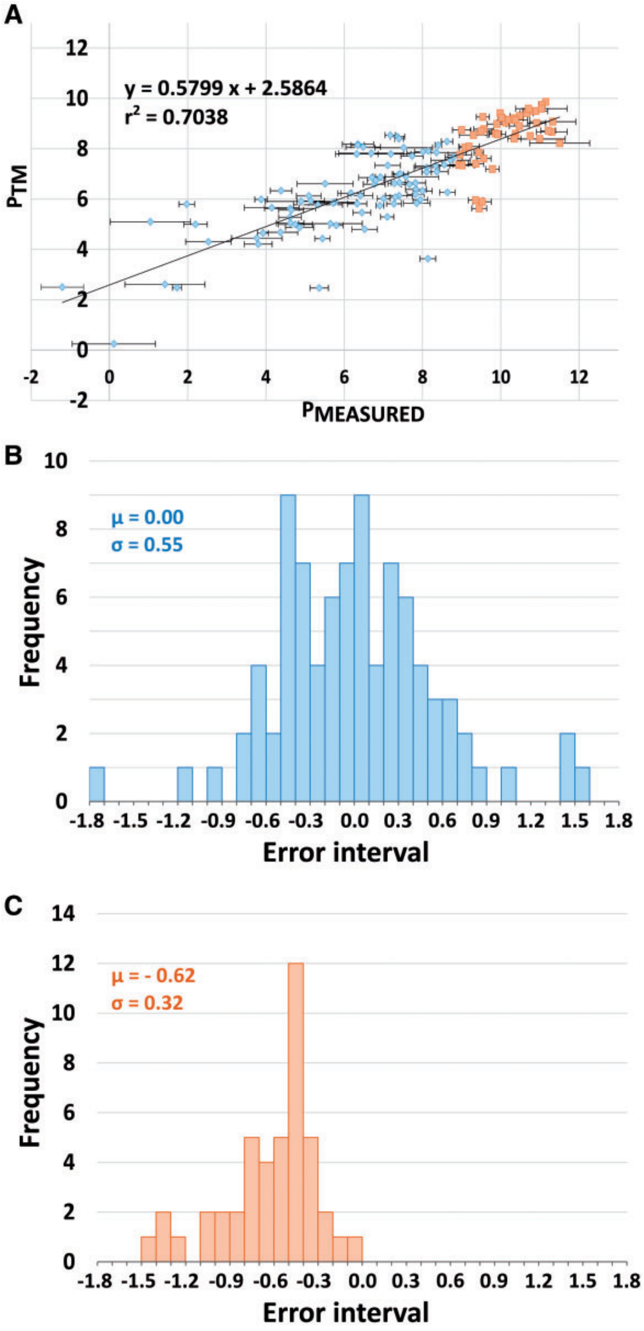

Figure 3. TM results for partition E33. In total, 67% of mRNA sequences with lowest protein fluorescence are used for model calibration (identification data set) while 33% of mRNA sequences with highest protein fluorescence are used for model predictive power assessment (test data set). (**A**) Predicted over measured protein fluorescence. Data are normalized with natural logarithm only. Blue diamonds are calibration points. Orange squares are test points. (**B**) TM residuals histogram for the identification data set. (**C**) TM residuals histogram for the test data set. (B and C) Data points are normalized.

N-PLS is extracting meaningful, though partial, information from the standby sequence that is relevant for describing reporter protein fluorescence. Interestingly, this result has parallels to the study by Borujeni *et al.* (6, 9). They detailed the computation of the free Gibbs energy associated with the standby sequence, which further improved Salis *et al.* (6) model predictions.

All in all, these results suggest the N-PLS method to be rather sensitive to data partitioning in a data sparsity context. The encoding method is an important factor but no clear rule could be identified regarding the best approach. Rather it must be studied case by case. As general trend, N-PLS tends to be more flexible in data calibration but clearly inferior to the TM in terms of extrapolation. Thus, for the particular problem at hand we conclude the TM to be a more powerful methodology than N-PLS when compared with standalone methods.

3.3 Hybrid thermodynamic-N-PLS models

The design of a hybrid model structure must consider the different types of knowledge available (19). For the present problem,

Table 2. Comparison of TM, hybrid TM+NPLS1 and hybrid TM+NPLS1+NPLS2 models for data partition scenario E33

Models	Identification				Test			Relative error	
	% Var	MSE	AIC$_c$	MSE/MSE$_{TM}$	% Var	MSE	MSE/MSE$_{TM}$	Top 5%	Top 1%
TM[a]	68.14	0.30		1	56.59	0.49	1	−62.06	−81.47
Hybrid TM+NPLS1[b]									
NPLS1 module[b]	70.83				52.27				
TM+NPLS1[b]	90.71	0.09	−197.35	0.29	79.28	0.23	0.47	−27.73	−33.76
Hybrid TM+NPLS1+NPLS2[c]									
NPLS2 module[c]	6.52				4.27				
TM+NPLS1+NPLS2[c]	91.31	0.08	−201.20	0.26	80.16	0.22	0.44	−26.97	−37.39

Identification refers to performance metrics in the identification data partition. Prediction refers to performance metrics in the test data partition. Relative errors refer to the relative absolute deviation of model prediction and measurement for the top 5% and 1% protein expression sequences.

[a] $\beta = 0.30$.

[b] With encoding 4 and *Fac* = 6.

[c] With encoding 5 and *Fac* = 1.

there are two main sources of knowledge: (i) *a priori* knowledge regarding the thermodynamics of the mRNA and ribosome complex formation and (ii) a sequence-activity data set that fully reflects all mechanisms involved in protein translation, many of which still lacking mechanistic interpretation. A key rule in hybrid modeling is that reliable mechanistic knowledge has priority over heuristic or empirical knowledge (35). We have thus adopted a hybrid structure that gives priority to the TM (Figure 1). Therefore, the hybrid model may be seen as an improvement of a core TM. Firstly, we have studied the combination of the TM and NPLS1 in parallel (hybrid TM+NPLS1 structure). Then, we have studied the inclusion of the NPLS2 module in parallel (hybrid TM+NPLS1+NPLS2 structure).

Hybrid TM+NPLS1 model

In this hybrid structure, NPLS1 runs in tandem with the TM resorting to the same input information (i.e. mRNA sequence) and extracting information from the TM residuals. Therefore, the job of NPLS1 is to improve the TM by considering primary structure information that might not be adequately represented by the Gibbs free energy framework.

The procedure to identify the TM module as part of the hybrid model is the same as for the standalone TM (Identification Step 1 described in Section 2.4). Thus the results of this first identification step were previously discussed and summarized in Table 1 and Figures 2 and 3.

The results of NPLS1 identification in tandem with the TM (Identification Step 2 described in Section 2.4) are summarized in Tables 2 and 3 for partitions E33 and E67, respectively. In the case of partition E33, NPLS1 was able to explain 70.8% of TM residuals in the model identification data set and 52.3% in the test data set (second row of Table 2). These results clearly indicate NPLS1 succeeded to extract a significant amount of information from TM residuals. TM residuals are due to experimental noise and mechanisms not adequately described by the TM. Given the high variance explained in both the identification and test data sets, the information extracted by NPLS1 is likely to be related to unknown mechanisms rather than to random noise. In the case of E67, the improvement is also clear but less expressive (second row of Table 3). For this reason, in what follows we focus the analysis on partition E33 results.

The hybrid TM+NPLS1 output is calculated with the contributions of the TM and NPLS1 modules together, obtained by summing the output of both modules ($\mathbf{P}_{TM} + \mathbf{P}_{NPLS1}$). This procedure improved the description of the model identification data set, with a significant decrease in the MSE ratio to 0.29. Even more impressively, the test data set MSE ratio decrease to 0.47, which means that the prediction error is approximately 50% of the TM (third row of Table 2).

Figure 4A–C plots predicted over measured protein fluorescence data for the hybrid TM+NPLS1 and respective residuals distribution. Comparing with the standalone TM (Figure 3A–C), it may be seen that the dispersion of model identification residuals decreases 1.7-fold for the hybrid model (Figure 3B, $\sigma = 0.32$) when compared with the TM (Figure 3B, $\sigma = 0.55$). In the case of the test partition, it strikes the almost 3-fold reduction in model bias ($\mu = -0.62$ for the TM, Figure 3C, $\mu = 0.23$ for the hybrid, Figure 4C). Moreover, according to the Shapiro–Wilk normality test, the residuals of the test partition are normal for the hybrid model while they are not for the TM. This means a more consistent model representation of observations across the full measured space. It means in particular for the present problem that the prediction bias of high performing sequences is practically eliminated by the hybrid approach.

Hybrid TM+NPLS1+NPLS2 model

In this structure, NPLS2 runs in tandem with the TM+NPLS1 having as input information the standby sequence upstream the mRNA-rRNA interactions loci, on the basis of which it extracts information from the TM+NPLS1 residuals. The job of NPLS2 is thus to improve the TM+NPLS1 model by considering mRNA-rRNA interactions, which are not accounted for neither by the TM nor by the NPLS1 model.

The results of this identification step (Identification Step 3 previously described in Section 2.4) are summarized in Tables 2 and 3 (fourth and fifth rows). The inclusion of the NPLS2 module does not significantly improve the hybrid model performance. In the case of partition E33 (Table 2), the MSE ratio is 0.26 in the identification partition and 0.44 in the test partition. In the case of partition E67 (Table 3), the MSE ratio is 0.60 in the identification partition and 0.94 in the test partition. Again focusing on partition E33, we calculated the *AICc* values to discriminate the more parsimonious model among the two hybrid model candidates. The *AICc* values are −197.35 (TM+NPLS1) and −201.20 (TM+NPLS1+NPLS2) for the identification data set. According to this criterion, the hybrid TM+NPLS1+NPLS2 model is more parsimonious than the TM+NPLS1 model, but the difference is marginal thus inconclusive.

Effect of sample size and partitioning method

This article is focused on maximizing predictive power. The key idea is to show that a model developed from a small number of

Table 3. Comparison of TM, hybrid TM+NPLS1 and hybrid TM+NPLS1+NPLS2 models for data partition scenario E67

Models	Identification				Test			Relative error	
	% Var	MSE	AIC$_c$	MSE/MSE$_{TM}$	% Var	MSE	MSE/MSE$_{TM}$	Top 5%	Top 1%
TM[a]	85.86	0.25		1	<0	0.98	1	−98.32	−112.83
Hybrid TM+NPLS1[b]									
NPLS1 module[b]	28.63				4.88				
TM+NPLS1[b]	89.91	0.18	−69.61	0.71	<0	0.93	0.94	−67.00	−70.38
Hybrid TM+NPLS1+NPLS2[c]									
NPLS2 module[c]	18.97				1.34				
TM+NPLS1+NPLS2[c]	91.82	0.15	−74.87	0.60	<0	0.92	0.94	−56.49	−69.75

Identification refers to performance metrics in the identification data partition. Prediction refers to performance metrics in the test data partition. Relative errors refer to the relative absolute deviation of model prediction and measurement for the top 5% and 1% protein expression sequences.

[a] $\beta = 0.25$.

[b] With encoding 4 and *Fac* 1.

[c] With encoding 5 and *Fac* 2.

experimentally validated sequences is able to predict better performing sequences. For this purpose, the sparse data set of RBS calculator model 1.0 (with just 132 sequences) together with the data partitioning E33 and E67 are ideal test cases. To eliminate the possibility of the good results being obtained by chance, the partition R scenario (100 runs with 33% sequences randomly sampled for testing) was also studied. The obtained MSE for the TM (RBS calculator model 1.0) was 0.29 ± 0.03 (identification partition) and 0.30 ± 0.07 (testing partition). The results for the hybrid TM+NPLS1 model were 0.23 ± 0.03 (identification partition) and 0.26 ± 0.06 (testing partition). Thus an average MSE reduction is observed of 22% and 15% for identification and testing, respectively.

We have also applied the same random partitioning approach for the larger data set with 485 sequences of RBS calculator model 2.1 (10). In this case, the identification method of the hybrid model is the same as previously described except that the ΔG data published by (10) was directly used for the TM fitting (Identification Step 1 described in Section 2.4). As before, the procedure is repeated 100 times with random sampling of 33% sequences for testing giving rise to 100 different models. The obtained MSE for the TM (RBS calculator model 2.1) was 0.35 ± 0.01 (identification partition) and 0.35 ± 0.03 (testing partition). The results for the hybrid TM+NPLS1 model were 0.18 ± 0.01 (identification partition) and 0.24 ± 0.03 (testing partition). Thus an average MSE reduction is observed of 49% and 32% for identification and testing, respectively. These results suggest a significant and systematic reduction of modeling errors by the hybrid approach in relation to the standalone TM independently of the sample size and data partitioning. Moreover, improvements are more substantial for the RBS model calculator model 2.1. For this data set, Borujeni and Salis (10) hypothesized that folded RNA structures may not have enough time to fold inside the cell, creating a non-equilibrium effect termed as “Ribosome Drafting”. In such cases, thermodynamic modeling fails to deliver accurate predictions. In the hybrid model, however, the NPLS1 module seems to effectively correct the TM residuals associated to the Ribosome Drafting effect. This example illustrates well the advantages of the hybrid approach. Mechanistic models are given priority to describe the process but key mechanisms might be missing (in this case a kinetic effect). Empirical models are set to “learn” from the mechanistic model errors, thereby compensating for the missing mechanisms.

4. Concluding remarks

Mechanistic modeling based on well-established thermodynamic/kinetic principles is recognized as the most insightful approach to design biological parts, but many times sufficient knowledge for developing a coherent mechanistic model is lacking. On the other hand, sequence-activity data sparsity, which is very common in these problems, hinders the development of empirical models with sufficient predictive power for design. In this study, it is shown that combining both approaches in the form of hybrid semiparametric models may result in improved interpolation and extrapolation properties when compared to the “standalone” methodologies, thus paving the way for more efficient designs.

It is shown in particular how previously published *E. coli* RBS thermodynamic models (6, 10) can be improved by adopting a parallel hybrid construct where a N-way PLS model extracts information from the TM residuals. The patterns captured by the N-way PLS model represent the knowledge lacking in the TM. “Learning” form the TM errors is key to improve the predictive power of the full construct. The results suggest a systematic improvement by the hybrid method in relation to the standalone modeling methods irrespective of the sample size and partitioning method.

In this study, N-way PLS was adopted in the hybrid structures but several other empirical techniques could have been used. Besides (N)PLS, different forms of artificial neural networks, support vector machines, fuzzy systems, splines and many other techniques have been employed embedded in hybrid structures (see review by von Stosch *et al.* (19)). The comparison of approaches is not always concordant in the literature and seems to be problem dependent. For sequence-activity modeling problems, given the data sparsity difficulty, it is important to choose a method that performs data factorization and regression simultaneously. We have investigated multilinear regression, principle components regression, support vector machines, one-way PLS and multi-way PLS and found that the multi-way PLS provided the best results for the case study addressed here. Moreover, the choice of the method also depends if the final objective is maximizing predictive power or a better understanding of the underlying mechanisms. Neural networks and PLS are typically adopted when the goals is maximizing predictive power. PLS is frequently used for process interpretation, namely to identify the most influential input

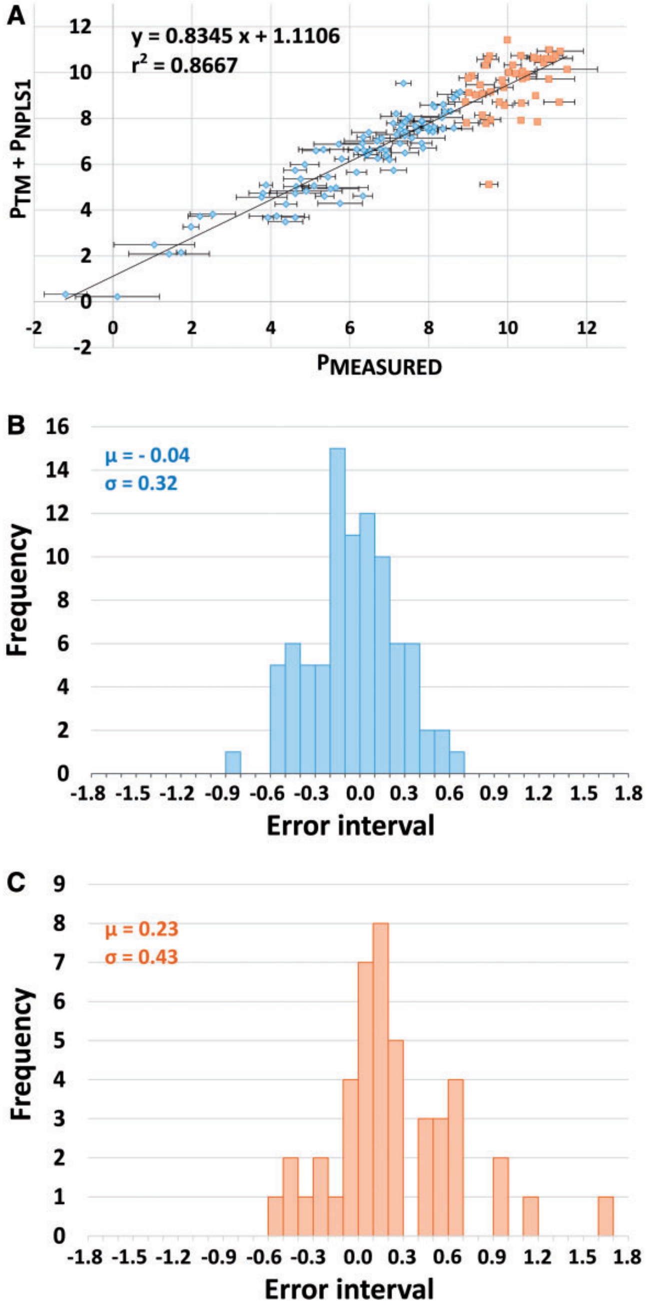

Figure 4. Hybrid model TM+NPLS1 modeling results for partition E33. In total, 67% of mRNA sequences with lowest protein fluorescence are used for model calibration (identification data set) while 33% of mRNA sequences with highest protein fluorescence are used for model predictive power assessment (test data set). (**A**) Predicted over measured protein fluorescence. Data are normalized with natural logarithm only. Blue diamonds are calibration points. Orange squares are test points. (**B**) Hybrid model residuals histogram for the identification data set. (**C**) Hybrid model residuals histogram for the test data set. (B and C) Data points are normalized.

variables for the target output. There is, however, some controversy regarding the efficiency of such analysis (36) particularly for sparse data sets. For the present study, it would be interesting to identify the sequence positions associated with higher TM residuals. We have applied the variable importance in projection (VIP) analysis and observed high variability of results depending on data sampling (results not shown). Such high variability precludes a reliable interpretation.

Different ways of combining mechanistic/empirical methods into hybrid models have been reported for a wide range of engineering problems (19). Similar design principles could be applied to a wide range of synthetic biology problems for which a sufficiently predictive model is still lacking. Parallel hybrid structures, such as the ones in the present study, are applicable to problems where a full mechanistic model is available but lacking predictive power. Serial hybrid structures are applicable to problems where only some parts are understood mechanistically with the remaining parts being modeled by empirical methods. Differential equations models of biological parts (21) with unrealistic kinetics and/or equilibrium assumptions can be tackled by simultaneously serial and parallel hybrid structures (33). In general, eukaryotic organisms are less understood than prokaryotic organisms from a mechanistic point of view. For instance, the nucleosome occupancy has been one of the key features to be included in a model for transcription initiation in *S. cerevisiae* (37). Hybrid semiparametric modeling becomes a natural candidate to improve sequence-activity models for such complex eukaryotic organisms. All in all, this study represents a first step toward the demonstration of the potential of hybrid modeling in synthetic biology, which could in principle be replicated to many different problems in the future.

Funding

This work was supported by the Portuguese Fundação para a Ciência e Tecnologia [PhD grant SFRH/BD/51577/2011 to R.M.C.P.].

Conflict of interest statement. None declared.

References

1. Chandran,D., Copeland,W.B., Sleight,S.C. and Sauro,H.M. (2008) Mathematical modeling and synthetic biology. *Drug. Discov. Today Dis. Model.*, 5, 299–309.
2. Marchisio,M.A. and Stelling,J. (2008) Computational design of synthetic gene circuits with composable parts. *Bioinformatics*, 24, 1903–1910.
3. Dvir,S., Velten,L., Sharon,E., Zeevi,D., Carey,L.B., Weinberger,A. and Segal,E. (2013) Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast. *Proc. Natl. Acad. Sci. U S A.*, 110, E2792–E2801.
4. Drubin,D.A., Way,J.C. and Silver,P.A. (2007) Designing biological systems. *Genes Dev.*, 21, 242–254.
5. Brewster,R.C., Jones,D.L. and Phillips,R. (2012) Tuning promoter strength through RNA polymerase binding site design in *Escherichia coli*. *PLoS Comput. Biol.*, 8, e1002811.
6. Salis,H.M., Mirsky,E.A. and Voigt,C.A. (2009) Automated design of synthetic ribosome binding sites to precisely control protein expression. *Nat. Biotechnol.*, 27, 946–950.
7. Na,D., Lee,S. and Lee,D. (2010) Mathematical modeling of translation initiation for the estimation of its efficiency to computationally design mRNA sequences with desired expression levels in prokaryotes. *BMC Syst. Biol.*, 4, 71.
8. Amman,F., Flamm,C. and Hofacker,I. (2012) Modelling translation initiation under the influence of sRNA. *Int. J. Mol. Sci.*, 13, 16223–16240.
9. Borujeni,A.E., Channarasappa,A.S. and Salis,H.M. (2014) Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res.*, 42, 2646–2659.
10. ——— and Salis,H.M. (2016) Translation initiation is controlled by RNA folding kinetics via a ribosome drafting mechanism. *J. Am. Chem. Soc.*, 138, 7016–7023.
11. González-Díaz,H., Pérez-Bello,A., Cruz-Monteagudo,M., González-Díaz,Y., Santana,L. and Uriarte,E. (2007) Chemometrics for QSAR with low sequence homology: mycobacterial promoter sequences recognition with 2D-RNA entropies. *Chemom. Intell. Lab. Syst.*, 85, 20–26.
12. Tavares,L.G., Lopes,H.S. and Erig Lima,C.R. (2008) A comparative study of machine learning methods for detecting promoters in bacterial DNA sequences. In: Huang,D.S., Wunsch,D.C., Levine,D.S. and Jo,K.H. (eds) *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence. ICIC 2008. Lecture Notes in Computer Science*, Vol. 5227. Springer, Berlin, Heidelberg.
13. Li,G., Leong,T. and Zhang,L. (2004) *Translation Initiation Sites Prediction with Mixture Gaussian Models*. pp. 338–349.
14. Li,S.M., Wakefield,J. and Self,S. (2008) A transdimensional Bayesian model for pattern recognition in DNA sequences. *Biostatistics*, 9, 668–685.
15. Zuo,Y.C. and Li,Q.Z. (2010) The hidden physical codes for modulating the prokaryotic transcription initiation. *Phys. Stat. Mech. Appl.*, 389, 4217–4233.
16. Jonsson,J., Norberg,T., Carlsson,L., Gustafsson,C. and Wold,S. (1993) Quantitative Sequence-Activity Models (QSAM)-tools for sequence design. *Nucleic Acids Res* 21, 733–739.
17. Liang,G. and Li,Z. (2007) Scores of generalized base properties for quantitative sequence-activity modelings for *E. coli* promoters based on support vector machine. *J. Mol. Graph. Model.*, 26, 269–281.
18. Ran,W. and Higgs,P.G. (2012) Contributions of speed and accuracy to translational selection in bacteria. *PLoS One*, 7, e51652.
19. von Stosch,M., Oliveira,R., Peres,J. and Feyo de Azevedo,S. (2014) Hybrid semi-parametric modeling in process systems engineering: past, present and future. *Comput. Chem. Eng.*, 60, 86–101.
20. Carinhas,N., Bernal,V., Teixeira,A.P., Carrondo,M.J.T., Alves,P.M. and Oliveira,R. (2011) Hybrid metabolic flux analysis: combining stoichiometric and statistical constraints to model the formation of complex recombinant products. *BMC Syst. Biol.*, 5, 34.
21. Isidro,I.A., Portela,R.M., Clemente,J.J., Cunha,A.E. and Oliveira,R. (2016) Hybrid metabolic flux analysis and recombinant protein prediction in Pichia pastoris X-33 cultures expressing a single-chain antibody fragment. *Bioprocess Biosyst. Eng.*, 39, 1351–1363.
22. Teixeira,A.P., Dias,J.M.L., Carinhas,N., Sousa,M., Clemente,J.J., Cunha,A.E., von Stosch,M., Alves,P.M., Carrondo,M.J.T. and Oliveira,R. (2011) Cell functional enviromics: unravelling the function of environmental factors. *BMC Syst. Biol.*, 5, 92.
23. Folch-Fortuny,A., Marques,R., Isidro,I.A., Oliveira,R. and Ferrer,A. (2016) Principal elementary mode analysis (PEMA). *Mol. Biosyst.*, 12, 737–746.
24. Costa,R.S., Machado,D., Rocha,I. and Ferreira,E.C. (2010) Hybrid dynamic modeling of Escherichia coli central metabolic network combining Michaelis-Menten and approximate kinetic equations. *Biosystems*, 100, 150–157.
25. von Stosch,M., Peres,J., de Azevedo,S.F. and Oliveira,R. (2010) Modelling biochemical networks with intrinsic time delays: a hybrid semi-parametric approach (2010). *BMC Syst. Biol.*, 4, 131.
26. Berry,T. and Harlim,J. (2016) Semiparametric modeling: correcting low-dimensional model error in parametric models. *J. Comput. Phys.*, 308, 305–321.
27. Hamilton,F., Lloyd,A.L. and Flores,K.B. (2017) Hybrid modeling and prediction of dynamical systems. *PLoS Comput. Biol.*, 13, e1005655.
28. Xia,T., SantaLucia,J., Burkard,M.E., Kierzek,R., Schroeder,S.J., Jiao,X., Cox,C. and Turner,D.H. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. *Biochemistry*, 37, 14719–14735.
29. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288, 911–940.
30. Dirks,R.M., Bois,J.S., Schaeffer,J.M., Winfree,E. and Pierce,N.A. (2007) Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev.*, 49, 65–88.

31. Bro,R. (1996) Multiway calibration. Multilinear PLS. *J. Chemom.*, 10, 47–61.
32. Andersson,C.A. and Bro,R. (2000) The N-way toolbox for MATLAB. *Chemom. Intell. Lab. Syst.*, 52, 1–4.
33. Nandy,A. (2006) Mathematical descriptors of DNA sequences: development and applications. *ARKIVOC*, 2006, 211–238.
34. Li,B., Morris,J. and Martin,E.B. (2002) Model selection for partial least squares regression. *Chemom. Intell. Lab. Syst.*, 64, 79–89.
37. Curran,K.A., Crook,N.C., Karim,A.S., Gupta,A., Wagman,A.M. and Alper,H.S. (2014) Design of synthetic yeast promoters via tuning of nucleosome architecture. *Nat. Commun.*, 5, 8.
35. Von Stosch,M., Oliveria,R., Peres,J. and De Azevedo,S.F. (2012) Hybrid modeling framework for process analytical technology: application to Bordetella pertussis cultures. *Biotechnol. Prog.*, 28, 284–291.
36. Kvalheim,O.M. (2010) Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots. *J. Chemom.*, 24, 496–504.