Rui Miguel Correia Portela
Mestre em Biotecnologia

Towards the rational design of standard biological parts for Synthetic Biology: applications to microbial systems

Dissertação para obtenção do Grau de Doutor em Bioengenharia

Orientador: Rui Manuel Freitas Oliveira,
Professor Associado, FCT-UNL
Co-orientador: Anton Glieder,
Professor, IMB-TU Graz

Júri:
Presidente: Prof. Doutor Pedro Manuel Corrêa C. Barahona,
Professor Catedrático, FCT-UNL
Arguentes: Prof. Doutora Lígia Raquel Marona Rodrigues,
Professora Auxiliar, EE-UM
Doutora Susana de Almeida M. Vinga Martins,
Investigadora Principal, IST-UL
Vogal: Prof. Doutora Paula Maria T. M. B. Gonçalves,
Professora Auxiliar, FCT-UNL

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Novembro de 2016

Towards the rational design of standard biological parts for Synthetic Biology: applications to microbial systems

Acknowledgments

As always, this work would have never been possible without the contribution and support of many people, both at professional and personal level.

Firstly, I would like to thank my supervisor Prof. Dr. Rui Oliveira for supporting my Ph. D. application, for giving me the opportunity to work with him again, for his supervision and for encouraging and establishing the collaboration with TU Graz.

Also, I would like to express my gratitude to Prof. Dr. Anton Glieder for accepting me in his lab, for his supervision and for providing all the conditions for a successful Ph. D. thesis.

I would like to thank people that I met throughout these five years. I will try to keep it roughly chronological and as short as possible. As a result, I might leave some people out of this list. So, I take this chance to acknowledge all the people who, in any way, contributed to this work but are not listed.

Starting the list itself, I begin with a sincere “thank you” to the people involved in the first year courses and lab rotations, in particular:

- To Prof. Dr. Armindo Salvador and the members of his group (especially Rui and David) for accepting me as a lab rotation student, for the scientific challenge that was proposed (I learned a lot during this period) and, of course, for all the stimulating and exhausting discussions.

- To Navjot, Rohit and Sara, you made special what was probably the best part of the Ph. D.: sharing that awful flat, the boat ride, the shirt that I had to throw away and the trips back there.

- To Tânia P. and Sofia, it was great to have classes with you again (although with Sofia and the BEB guys it was the first time sharing the classroom, it didn’t feel like that). And not to forget Tânia L., even without attending the course at this point, it was great to have someone close to in the beginning.

- To Gianluca, with whom I shared the last lab rotation (it could have been better, *e. g.* Italy-Spain, but that’s life). Thank you for the helpful discussions about the final project.

Then, the next phase came around, and I returned to SBE-FCT. Of course, I would like to thank all the members of SBE, especially to João (for his advices).

After few months, the experimental part of this project started in Graz. At first, I manage to avoid speaking German and English at the personal level with the Iberian group. Particularly, I would like to thank Inma, perhaps we didn’t meet in the best moment (it’s always a matter of timing), but it was still great to get to know an amazing Spanish that can actually speak Portuguese. One of a kind!

Then, with all the first semester goodbyes came the international group, I don’t need to say much… “all the nights that we don’t remember with all the friends that we’ll not forget”: Camilo, Diogo, Lorenzo, Liad and Kristina.

At the professional level, the Glieder group received me so well and made my stay so pleasant and fulfilling that I had to extend it… twice! I have to thank Elisa, Flo, Julia and all the

others members of the group for all the enjoyable lunch talks and social events. A really good work environment, especially when the experimental results were not that great. In particular, I would like to acknowledge the role of:

– Christian and Amneris (for the Croatian classes and) for letting us focus on the scientific work.

– Andrea Mellitzer for always (really always! How can you do it?) being in a good mood and with something to talk about. It was a shame that you were running away… to the 4th floor and later to Vienna.

– Astrid, I don't really have space here to repeat the email, even so, thanks again!

– I will end this part with the most important acknowledgement: to Thomas! As said to Astrid, I don't have space to state everything, but even so: without you my stay in Graz would have been completely different at all levels! Thanks for all the help with the project, all the discussions and experimental guidance! Even the great welcome letter… Then, as you say, "in essence 'all the wing man fails (and the one really good win!!)'". I would add all the non-wing-man moments outside the lab, always fun ("never forget the bamboo…" still makes me laugh)!

For the last phase, back to SBE-FCT, I would like to thank Moritz, Cristiana and Mauro for all the side projects and dinners. It was also great to discuss science with you, Moritz. Hope we can still do it from time to time!

A special word to Irina. You made the last months in Graz and the last goodbye irreplaceable. We ended up sharing all the ups and downs of the following year. For it, and all that is left unsaid: thanks!

Gostaria de agradecer ao Marino pelo recíproco resumo da década passada e por mostrar que algumas coisas nunca mudam (estou certo de este ser um dos elogios mais almejados). Obviamente, muito há a dizer, mas o espaço é curto. Saliento o esmiuçar da problemática associada a inúmeros resumos, apresentações, discussões, relatórios, portfólios e sessões de estudo, na mítica e gloriosa sala de BCM ("amanhã às 8h"), nas unidades curriculares partilhadas para chegar ao busílis da questão, o final do 3º ciclo. Há ainda a salientar os momentos humorísticos diários, e do quotidiano, bem como o acompanhar das vicissitudes extracurriculares. Termino com o expressar da vontade de estabelecer várias colaborações futuras e um "Obrigado, com o desejo que continuemos a 'legislelar bué', para além dos projectos comuns, por mais umas boas e muitas (não longas) décadas".[1]

Por fim, um obrigado à minha família (pais e tios) pelo apoio firme e constante durante todo este tempo, especialmente os últimos dez anos.

This Ph. D. thesis was undertaken with financial support from the Fundação para a Ciência e a Tecnologia through scholarship SFRH/BD/51577/2011.

[1] *Escrito em português por ter sido pedido expressamente.*

Resumo

A Biologia Sintética assenta em Partes Standard Biológicas (SBPs) modulares aptas a serem montadas em diferentes circuitos genéticos sintéticos capazes de realizar funções não observadas na natureza. A construção de tais circuitos sintéticos requer, frequentemente, o ajuste fino da expressão dos genes envolvidos de forma a equilibrar e optimizar os níveis de proteínas regulatórias ou enzimas metabólicas.

A presente tese explora novas metodologias conducentes ao melhoramento do desenvolvimento de SBPs, estando dividida em três partes principais. Na primeira parte, apresentam-se os resultados da diversificação racional de uma SBP natural (promotor *AOX1* de *Pichia pastoris*). Neste estudo é adoptado uma estratégia de mutagénese de alta resolução que consiste em substituir sistematicamente três nucleótidos ao longo do promotor. Os resultados obtidos comprovam a robustez do promotor em estudo com poucas mutações (em regiões relevantes) capazes de alterar significativamente a sua expressão. Na segunda parte, apresenta-se um método, testado experimentalmente, de geração de SPBs *de novo*, com o objectivo de desenvolver centros de promotores sintéticos em *P. pastoris.* Estes mostraram-se funcionais em diferentes contextos. Por último, apresenta-se uma nova estratégia *in silico* para desenhar SBPs. Com esta estratégia, demonstra-se que o actual modelo mecanístico do Sitio de Ligação do Ribossoma (RBS) de *Escherichia coli* é passível de ser melhorado pela adição de informação proveniente de um procedimento de aprendizagem automática, de forma híbrida. Estes resultados representam a primeira aplicação bem-sucedida de modelação híbrida em Biologia Sintética.

De forma geral, esta tese aborda os principais problemas que precedem a implementação de um circuito genético sintético. Esta tese sustenta a concepção de SBPs sintéticos *de novo*, funcionais em diferentes contextos genéticos, e métodos de desenho baseados em modelação híbrida que melhoram o actual modelo de desenho de SBPs. A relevância desta estratégia irá certamente aumentar progressivamente à medida que mais dados de validação de SBPs estejam disponíveis no futuro.

Palavras-chave

Pichia pastoris · *Escherichia coli* · Biologia Sintética · Sistemas Híbridos · Mínimos Quadrados Parciais · Partes Standard Biológicas · Promotores · Região 5´ Não Traduzida

Abstract

Synthetic Biology relies on modular Standard Biological Parts (SBPs) that can be flexibly assembled in synthetic genetic circuits able to perform complex unnatural functions. The construction of such synthetic circuits frequently requires the fine-tuning of gene expression to balance and optimize protein levels of regulators or metabolic enzymes.

This thesis develops novel methodologies that improve the design of SBPs. It is divided in three main parts. Firstly, the rational diversification of a natural SBP, the *Pichia pastoris* promoter (*AOX1*), is studied by performing a high resolution mutagenesis. The method replaces systematically three nucleotides along the promoter. The results show that this promoter is remarkably robust, since only few mutations (on relevant regions) cause significant changes in expression. Next, a *de novo* SBP design method was experimentally tested for the development of synthetic core promoters for *P. pastoris*. The designed synthetic core promoters are shown to be functional in different contexts. Lastly, a new *in silico* strategy to design SBPs is presented. It is shown that the prediction capabilities of the current state-of-the-art *Escherichia coli* Ribosome Binding Site (RBS) mechanistic model can be improved by adding information using a machine learning procedure, in a hybrid manner. This represents the first successful application of hybrid modeling in Synthetic Biology.

Overall, this Ph. D. thesis targets the main problems that occur prior to the synthetic genetic circuit implementation: this thesis supports the *de novo* design of synthetic SBPs, functional in different genetic contexts, and a hybrid modeling design method that improve the current state of the art of SBP design. This strategy will certainly increase in relevance as SBPs validation data become increasingly available in future.

Keywords

Pichia pastoris · *Escherichia coli* · Synthetic Biology · Hybrid Systems · Partial Least Squares · Standard Biological Parts · Promoters · 5' Untranslated Regions

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

α	Thermodynamic model empirical calibration parameter
β	Thermodynamic model Boltzmann factor
ΔG	Gibbs free energy (kcal/mol)
ΔG_{mRNA}	Gibbs free energy referent to the most stable mRNA secondary structure
$\Delta G_{mRNA:rRIB}$,	Gibbs free energy referent to the interaction between the mRNA and rRNA molecules that minimizes the sum of $\Delta G_{mRNA:rRIB}$ with $\Delta G_{SPACING}$
$\Delta G_{SPACING}$	Gibbs free energy empirical penalty applied when the distance between the mRNA-rRNA interaction and start codon is not optimal
$\Delta G_{STANDBY}$	Gibbs free energy needed to unfold any mRNA secondary structure generated in the area of the rRNA hybridization after its establishment
ΔG_{START}	Gibbs free energy referent to the tRNA and start codon interaction
σ^{70}	Primary sigma factor protein needed for transcription initiation of housekeeping genes in bacteria
5'UTR	5' Untranslated Region
AAA	Mutation consisting of adenine triplets
AIC	Akaike Information Criterion
ANOVA	One-way Analysis-Of-Variance
AOX1	*Alcohol Oxidase* 1 gene
BMD1 culture medium	Buffered Minimal culture Medium with Dextrose
BMM10 culture medium	Buffered Minimal culture Medium with Methanol (10 time concentrated)
BMM2 culture medium	Buffered Minimal culture Medium with Methanol (2 time concentrated)
CCC	Mutation consisting of cytosine triplets
CRM	*cis* Regulatory Module
E	**X** residuals matrix (on 4.3.4. N-PLS models) and final hybrid model **Y** residuals (on 4.3.5. Hybrid semiparametric model)
eGFP	enhanced Green Fluorescence Protein
$\mathbf{E}_{TM}$	Thermodynamic model **Y** residuals
$\mathbf{E}_{TM+NPLS1}$	First hybrid model (TM+NPLS1) **Y** residuals
Fac	Number of PLS latent variables
Fac_{NPLS1}	Number of NPLS1 latent variables
Fac_{NPLS2}	Number of NPLS2 latent variables
Group A	Synthetic core promoters group created based on data from *Saccharomyces cerevisiae*, namely TATA box and motifs frequency and position
Group M	Synthetic core promoters group created based on data from *Saccharomyces cerevisiae*, namely motifs frequency and position
Group P	Synthetic core promoters group created based on data from *Saccharomyces cerevisiae*, namely nucleotide frequency distribution

Group T	Synthetic core promoters group created based on data from *Saccharomyces cerevisiae*, namely TATA box position
i	Index to identify the different **X** entries in the 1st dimension (input sequences in this case). Varies between 1 and I (*np* in our case)
I	Number of entries on **X** 1st dimension (equal to *np* in our case)
iGEM	International Genetically Engineered Machine
j	Index to identify the different **X** entries in the 2nd dimension (sequence base pairs in this case). Varies between 1 and J (*nb* in our case)
J	Number of entries on **X** 2nd dimension (equal to *nb* in our case)
k	Index to identify the different **X** entries in the 3rd dimension (encoding size in this case). Varies between 1 and K (*ne* in our case)
K	Number of entries on **X** 3rd dimension (equal to *ne* in our case)
LB culture media	Luria-Bertani culture media
mRFP1	Red Fluorescent Protein 1 mRNA
MSE	Mean Squared Error
MUT	Constitutive promoters (*AOX1*, *GAL*, *GPD1* and *ADH1*) with mutated TATA box (part of the TATA box replaced with cytosine triplets)
N	Dimensions of **X** (can take the value of 2 or 3 in this thesis)
nb	Maximum sequence length (in base pairs) in **X** (2nd dimension)
ne	Number of values representing a single nucleotide (depending on the encoding used) in **X** (3rd dimension)
np	Number of mRNA sequences in **X** (1st dimension)
npar	Number of model parameters
N-PLS	N-way Partial Least Squares
NPLS1	N-way Partial Least Squares model using mRNA primary structure as input
NPLS2	N-way Partial Least Squares model using mRNA:rRNA interactions as input
ns	Number of standby sequences in **X** (3rd dimension)
NUPACK	Nucleic Acid Package
OD600	Optical Density at 600nm
P_{AOX1}	*AOX1* promoter. A similar nomenclature is used for other genes
$P_{AOX1\text{-}R}$	*AOX1 cis* Regulatory Module. A similar nomenclature is used for other genes
Partition E33	Partition heuristically selected with 67% of sequences with lowest protein expression for model identification and the remaining for model testing
Partition E67	Partition heuristically selected with 33% of sequences with lowest protein expression for model identification and the remaining for model testing
Partition R	Partition randomly selected with 67% of sequences for model identification and the remaining for model testing
PCR	Polymerase Chain Reaction
PLS	Partial Least Squares

P_{MES}	Measured reporter protein (RFP1) fluorescence
P_{NPLS1}	Predicted reporter protein (RFP1) fluorescence by the NPLS1 model (N-way Partial Least Squares model using mRNA primary structure as input)
P_{NPLS2}	Predicted reporter protein (RFP1) fluorescence by the NPLS2 model (N-way Partial Least Squares model using mRNA:rRNA interactions as input)
P_{TM}	Predicted reporter protein (RFP1) fluorescence by the thermodynamic model
$P_{TM+NPLS1}$	Predicted reporter protein (RFP1) fluorescence by the hybrid model composed by the thermodynamic model coupled to the NPLS1 model (N-way Partial Least Squares model using mRNA primary structure as input)
QSAM	Quantitative Sequence-Activity Model
r^2	Correlation coefficient
RBS	Ribosomal Binding Sites
RFP1	Red Fluorescent Protein 1 expressed protein
RNAPII	RNA Polymerase II
RSBP	Registry of Standard Biological Parts
SBP	Standard Biological Part
SVM	Support Vector Machine
t - scores 4.3.4	**X** residuals matrix (on 4.3.4. N-PLS models) and final hybrid model **Y** residuals (on 4.3.5. Hybrid semiparametric model)
t - time 4.3.2	Thermodynamic model cultivation time (Eq. 4.2) and N-way Partial Least Squares **X** scores vector (Eq. 4.3)
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TM	Thermodynamic Model
TM+NPLS1	Hybrid model consisting of Thermodynamic Model coupled with the NPLS1 model (N-way Partial Least Squares model using mRNA primary structure as input)
TM+NPLS1+NPLS2	Hybrid model consisting of Thermodynamic Model coupled with the NPLS1 model (N-way Partial Least Squares model using mRNA primary structure as input) and NPLS2 model (N-way Partial Least Squares model using mRNA:rRNA interactions as input)
TSS	Transcription Start Site
UAS	Upstream Activating Sequence
URS	Upstream Repressing Sequence
Var. (%)	Percentage of Explained Variance
$\boldsymbol{w}$	Vector of **X** weights for dimension 2 ($\boldsymbol{w}^J$) and dimension 3 ($\boldsymbol{w}^K$)
WT	Wild type
X	Input Partial Least Squares matrix
x	Entries of **X** matrix (denoted by the index $\mathbf{x}_{ijk}$)
Y	Output Partial Least Squares matrix or vector

y	Entries of **Y** matrix (denoted by the index $\mathbf{y}_i$)
YPD culture media	Yeast extract Peptone Dextrose culture media
YPGal	Yeast extract Peptone Galactose culture media

Chapter 1

Introduction

1.1. A brief introduction to “Synthetic Biology”

There were many scientific discoveries and technological developments that enabled the emergence of a new field, around the turn of the millennium, that aimed to implement rigorous engineering principles in biology. That field is Synthetic Biology. There are many proposed definitions, but perhaps the most accepted one defines Synthetic Biology as the field that makes use “of molecular biology tools and techniques to forward-engineer cellular behavior” (1). However, to fully understand what this definition means, an historical perspective is needed.

1.1.1. Historical background

From a molecular biology perspective, the most relevant discoveries that compose the foundations of Synthetic Biology can be traced back to 1868. In this year, DNA was isolated and characterized as the molecule responsible for the transmission of hereditary traits by Friedrich Miescher (2). Other landmarks were the discovery of DNA structure by Watson and Crick in 1953 (3) and the first description of a genetic regulatory circuit by Monod and Jacob in 1961 (4). Soon after the characterization of this circuit, it was hypothesized that it would be possible to create new regulatory systems using molecular parts alone (5), however the technologies and knowledge needed to implement such idea were not yet available.

Later on, with the development of Polymerase Chain Reaction (PCR) (6), the discovery of restriction enzymes and the development of the recombinant DNA technology, molecular cloning and genetic engineering (7), it was possible to develop rudimentary circuits but not as complex as the ones found in nature (7). This limitation was mainly due to the lack of flexible DNA manipulation methodologies, absence of high throughput technologies and nonexistence of genomic level information (1). Such enabling technologies were developed and fine-tuned in parallel to the establishment of a field closely related to Synthetic Biology: Systems Biology (8).

Many important concepts to Synthetic Biology were discovered within Systems Biology using a top-down approach (9). Namely, the understanding of the systems level organization of biological networks and the realization that there was some parallelism between them and systems engineering (9). All of this knowledge was then used as foundations for a bottom-up approach. One of the aims of this approach was to forward engineer new regulatory networks in cells with the use of molecular parts (10). It became then possible to study the natural networks organization and to develop synthetic networks with interesting applications in all fields of biotechnology (11).

Synthetic Biology has faced many obstacles. Perhaps the main one has been the inability to make Standard Biological Parts (SBPs) perform in an equivalent manner, independently of the local genetic context (12). In a first stage, the attempted solution consisted on the SBPs standardization, and the establishment of registries (*e. g.* International Genetically Engineered Machine, iGEM, and Registry of Standard Biological Parts, RSBP). This lead to the establishment of several organized SBP repositories that the community could rely on, reducing the part diversity and providing a SBP performance indication (12, 13). Additionally, a community effort was made to

use similar protocols across research laboratories (*e. g.* openwetware.org) and to create easily interchangeable software (using, Systems Biology and Synthetic Biology Open Language) (14, 15).

However, the SBP performance was shown not to be valid in all circumstances. For instance, among other problems, the SBP performance varies depending on the local genetic context (12). Consequently, with few exceptions (Ribosome Binding Site (RBS) designer (16)), there were difficulties predicting the behavior of SBPs. Only the progress achieved with novel and more flexible molecular cloning and genome editing procedures (*e. g.* Gibson assembly (17) and CRISP-Cas9 (18), respectively) and the development of high-throughput screening methods enabled faster design-test-redesign cycles (*e. g.* using a microfluidics device in a direct evolution study (19)). This, together with the decline of sequencing (20) and gene synthesis costs (21), allowed the development of proof-of-concept synthetic genetic circuits (22, 23), in a first phase, and the implementation of applied genetic circuits (11), latter on. Furthermore, the implemented synthetic genetic circuits grew in complexity once the technical problems were overcomed (10).

The main genetic circuits landmarks were the implementation of a toggle switch (22), a repressilator (23) and, soon after, the negative-feedback loop as a way to reduce gene expression noise in synthetic gene networks (24). These pioneer circuits were constructed with the minimum number of SBPs that are needed to create a circuit (at least two responsive and interacting genes (22)). They establish the first successful attempts to have a synthetic circuit, with predictable and dynamic biological response from an initial external signal (1). It was also proven that such signals can have a biological origin when the first cell to cell communication circuit was presented (based on quorum sensing (25)). Since there was a limitation on the complexity of synthetic genetic circuits (10, 12), the ability to synthetically control cell to cell communication enabled the development of more complex circuits. This was particularly valuable in cases where circuits could be decomposed in modular functions and a microbial consortium is suitable to solve the problem at hand (26).

Later on, more complex circuits were studied (27). The next phase of Synthetic Biology involved the development of circuits that implemented logic functions (such as *or*, *not* and *nor*). Such functions would be the basis of a second generation of circuits that were able to integrate and properly handle the input signals and generate a appropriate output (10, 28).

Finally, with all the developed circuits, and gathered knowledge, several applications were presented (11). Two prominent examples were the implementation of a metabolic pathway to produce a precursor of an antimalarial drug, artemisin (29, 30), and the development of an engineered bacteria to invade and kill cancer cells (31).

In this way, if we come back to the initially stated Synthetic Biology definition (the field that makes use "(…) of molecular biology tools and techniques to forward-engineer cellular behavior" (1)), it becomes apparent that the referred molecular biology tools and techniques are not limited to the traditional genetic engineering procedures. Rather, the community thrives to develop new ways to improve molecular cloning (32), reduce the impact of biological noise on genetic circuits (24)

and, overall, construct of more complex and efficient genetic circuits (10) and biological functions by “forward-engineer cellular behavior” (33) using Synthetic Biology tools.

1.1.2. Organisms for Synthetic Biology

A host organism used in Synthetic Biology in which the SBPs are assembled to create genetic circuits that perform desired complex behaviors is called a chassis (34). In the same way as for genetic engineering and others fields related with molecular biology, the most commonly used organism is *Escherichia coli* (35). Much of the foundational work in the field was carried out in this organism, namely the first circuits and logic gates (22, 23, 27). Some devices (set of assembled parts that carry out specific functions (34)) and Synthetic Biology applications have been developed in other more complex organisms, like yeasts and mammalian cell lines (for a review please refer to (36, 37)). Here we will focus on microbial Synthetic Biology (35).

E. coli was not only used as host organism for genetic circuits gates (22, 23, 27), but also as a source of SBPs used in the firsts circuits (38). Examples of these first modules include: switches (22), oscillators (23) and circuits that implement logic functions (27). Later on, cell to cell communication circuits and more complex circuits (25), like pulse generators (39), or even light sensitive circuits (*e. g.* edge detector (40)) were presented.

In the case of yeasts, and in a similar way to *E. coli*, the most used organism in Synthetic Biology is the most used yeast in other molecular biology related fields: *Saccharomyces cerevisiae* (35). However, most of the circuits applied in *S. cerevisiae* were preceded by equivalent ones in *E. coli*. Given that *E. coli* is one of the easiest organisms to control in the laboratory (41), most of the proof-of-concept circuits were made with it. Only later, if it posed an interesting problem or had enough advantages, it was implemented in more complex organisms. For instance, the metabolic pathway to produce a precursor of an antimalarial drug, artemisin, was first developed in *E. coli* (42), and only then in *S. cerevisiae* (43), in order to achieve higher productivities. A similar profile can be seen for the genetic circuits implemented in *S. cerevisiae*, that is, the circuits implemented in this yeast were usually preceded by similar circuits developed in bacteria (*e. g.* cell to cell communication circuits (25, 44)).

An exception to what was previously stated are the fundamental studies that are specific for eukaryotes and some applications that take advantage some feature only present in eukaryotes. Some examples are the use of epigenetics and chromatin regulation (45) to facilitate the implementation of a new function (for instance by fine-tuning gene expression modulating nucleosome positioning (46, 47)). Additionally, while performing such studies, fundamental knowledge can also be gained, and be used as basis for higher eukaryotes research (45). Alternatively, the intended device or circuit might take advantage of a eukaryotic property, like the internal membrane system, for instance, to improve the productivity of one molecule of interest (29, 30, 43).

In the last years, the scientific community has become increasingly aware that such laboratory workhorses might be the most suitable for fundamental studies (whose conclusions

might be valid for other organisms) and for implementing proof-of-concept circuits, but, when it comes to industrial applications, those organisms might be suboptimal (48). At the industrial level, there are other difficulties that need to be addressed. For instance, for a particular application, it might happen that another organism is better suited for industrial scale-up or downstream processing (48).

To target this problem, as Voigt stated (41), the number of labs that are considering alternative organisms to develop Synthetic Biology applications is growing. Since the choice of a new chassis organism for a project in this area is frequently limited to the available and characterized SBPs (35), we considered an yeast with numerous industrial applications but with a lack of specific Synthetic Biology tools, *Pichia pastoris*.

Overall, this yeast presents numerous industrial advantages in comparison to other yeast species. Namely, *P. pastoris* is a methylotrophic yeast that can use methanol as sole carbon and energy source. It can grow up to $200gL^{-1}$ of dry weight without producing high amounts of byproducts (like ethanol or acetic acid). From the protein production point of view, this yeast can secrete the target protein and still obtain high yields (up to $30gL^{-1}$). Also, it has the relevant advantage of having a glycosylation pattern more similar to the one found in humans than the pattern obtained with *S. cerevisiae* (49, 50). Additionally, the *Alcohol Oxidase* 1 (*AOX1*) promoter has been the most used promoter for these applications, since it is inducible and tightly regulated. For these reasons, in this thesis, one of the main focus is in the development of SBPs for *P. pastoris*.

1.1.3. Standard Biological Parts

A SBP is an individual component of the overall gene expression machinery (34). Specifically, these SBPs are DNA or RNA sequences (51–53). Originally, they were restricted to promoters, ribosome binding sites, terminators, translation terminator sites, protein coding regions and upstream regulatory regions (38). However, with the development of Synthetic Biology, more and more parts were discovered and applied. That is the case of SBPs based on RNA (aptamers, ribozymes and small RNAs) (51, 52). Such SBPs are then assembled in devices that are able to perform a specific function. To do so, they interact with each other to create another layer of complexity, forming systems (10).

The three main design principles that were imported from engineering and incorporated in this field were abstraction, modularization and standardization (38). They aimed to facilitate the construction of new genetic circuits that coded a new biological function. One of the ways to do so is by characterize and standardize the SBPs available. In this way, in theory, it should be possible to interchange them, in a modular manner, and select the one most suited to the problem at hand without a SPB significant performance change (13). Such SBPs should be as orthogonal as possible to the natural system so that interferences with each other are minimal (38). Additionally, with the increase of complexity of the genetic circuits created, it is impossible to keep track of all the DNA and RNA sequences, SBPs, devices and its interactions. Thus, there is the need for

computational tools that assist in this design process and translate a complex network in an intangible high level representation of the genetic circuits (34, 38).

With the development of Synthetic Biology, many control points for genetic circuits were presented, namely transcription, translation and RNA based (52). These increased the control and flexibility with which genetic circuits can be designed and controlled. However, there is the need to, within each group, have available a wide diversity of SBPs that are sequence diversified, have different properties (*e. g.* respond to different molecules) and have different performances. Hence, there is the need to design new SBPs.

1.1.4. Methods for designing Standard Biological Parts

One of the concepts that Synthetic Biology imported from engineering is abstraction (38). With the increase in systems complexity, it is unlikely that someone can manage all the information underling it (*e. g.* from DNA sequence, to molecular interactions and system behavior). Abstraction helps to deal with this complexity by dividing the system into layers, simplifying the problem. However, for it to be useful, a modification in one layer of abstraction must have predictable effects on the remaining ones (38). As mentioned previously, among other problems in Synthetic Biology, there is a SBPs performance context dependency (12). Thus, there is the need for models to predict SBP and circuits performance accurately (34) and truly synthetic SBPs that are as orthogonal and modular as possible (54).

There are several methods that were used to develop new SBPs. These parts can be the natural DNA sequences, can be derived from them (in a rational or random manner) or can be design from scratch.

So far, most of the developed circuits and SBPs, as well as the advances in their design and predictive models, were derived from known natural circuits and parts. Additionally, each device was limited to few interacting elements (34). There were cases in which such elements were adapted from the host organism (34), and others that were imported from other organisms (for instance, invasion machinery from *Yersinia pseudotuberculosis* (31) and quorum sensors from *Vibrio fischeri* (25)).

Subsequently to the SBPs identification, it is possible to create a library of derived DNA or RNA sequences that cover a wider range of the original SBP performance. This enables the posterior selection of an appropriate SBP depending on the circuit objective. There are two main procedures to do so: random sequence modifications (55–57) or rational approaches (58–60).

The rational design of SBPs is limited to the mechanistic knowledge available, thus, there are fewer studies with this objective. One exception is the RBS designer in *E. coli* (16) and subsequent derived models (for instance (61)). More recently, the rational fine-tuning of synthetic promoters in yeast has been proven (62).

1.2. Thesis outline

1.2.1. Objectives

In this thesis, the main objective is to develop novel methodologies to create SBPs for Synthetic Biology. Different approaches were performed to achieve such goal, since each of them targets a specific problem.

Two main topics were particularly addressed:

- Design of synthetic core promoters for yeasts using rational design methods. Such sequences were both implemented and assessed in *S. cerevisiae* and *P. pastoris*.
- Development of Quantitative Sequence-Activity Model (QSAM) based on hybrid semiparametric mathematical formalism targeting more efficient design of SBPs.

The first part of the thesis (Chapters 2 and 3) focus on the development of such SBPs for *P. pastoris*, while in chapter 4 we target the development of SBPs in *E. coli.* The phases of the SBPs development that we aim to improve in this thesis are: *de novo* design of SBPs, together with their modularity and orthogonality test, interpretation of their most important features, fine-tuning of SBPs and development of new approaches for the improvement of current SBPs design tools.

1.2.2. Thesis structure

This thesis is divided in five chapters structured as follows:

In Chapter 1 (current chapter), a general introduction to the topics addressed in the thesis is presented. Firstly, a basic introduction to the main topic, Synthetic Biology, its historical background and main landmarks are overviewed. Then, a more detailed description on the strategies employed to accomplish such goals is presented. Namely, the description of the main chassis used and the principles of development of SBPs are described. This part of the thesis aims to provide scientific context (and main references) for the development of Chapters 2, 3 and 4. Lastly, the thesis general objective is presented together with a brief outline of each chapter.

In Chapters 2 and 3 we focused on developing SBPs for *P. pastoris*. In the first study (Chapter 2), we developed a high resolution mutagenesis analysis of the most used *P. pastoris* promoter (*AOX1*). Here, we performed a systematic three nucleotide replacement study that covered, adjacently, the whole 200bp long *AOX1* core promoter and 5' Untranslated Region (5'UTR) with the minimum number of variants. Such variants were characterized and allowed the identification of the most important regions on this promoter (that were most affected by the tested mutations), and, potentially, these main conclusions can be transferred to any core promoter in yeast. The main results show that the *AOX1* core promoter is remarkably robust and only specific areas of the core promoter show a higher sensitivity to the tested mutations. Namely, the protein binding regions (*e. g.* TATA box) and other regions with biological importance (*e. g.* downstream of the Transcription Start Site – TSS) show a significant difference in expression when compared to the respective wild type.

With the main conclusions of Chapter 2 (high core promoter robustness to mutations), we implemented a strategy to develop synthetic core promoters in *P. pastoris*. This strategy (Chapter 3) used a data set of *S. cerevisiae* core promoters' main features (nucleotide frequency, important motifs, TATA box location (63)). With this information, we generated synthetic sequences (with no significant match with the genomic sequences of *P. pastoris*) to be tested as core promoters. In the first screening, when the synthetic core promoters were controlled by the *AOX1 cis* Regulatory Module (CRM), its activity spanned a 200-fold range (0.3% to 70.6% of the wild type *AOX1* level). Then the best performing ones were tested in different conditions (in different yeasts and controlled by different CRMs). Inducible CRM constructs showed significantly higher activity than constitutive CRMs, reaching up to 176% of the respective natural core promoter. Additionally, comparing the activity of the synthetic core promoters in different conditions, it was possible to identify high correlations only for CRMs within the same organism. Overall, this approach enabled a successful design (that was experimentally tested) and confirms, to some extent, the synthetic core promoters' modularity.

In Chapter 4, we explored the utilization of hybrid modeling as Quantitative Sequence--Activity Modeling (QSAM) tool for the *in silico* design of SBPs. Due to the fact that there is few information about the *P. pastoris* molecular mechanisms, especially compared with other unicellular model organisms (*S. cerevisiae* and *E. coli*), the development of mechanistic or hybrid models for this yeast is unfeasible at this stage. For this reason, it was used a data set of *E. coli* RBSs, respective protein expression data and state of the art mechanistic RBS design model (16). By coupling this mechanistic model with a statistical model, it became possible to develop a hybrid model that, in particular conditions, performs better than the standalone mechanistic model and machine learning algorithm.

In Chapter 5, we review the main accomplishments achieved in this thesis and summarily explore the main future perspectives that this thesis can lead to. Overall this thesis explored different methodologies that aim to solve the different problems related with the design and fine--tuning of SBPs.

1.3. References

1. Cameron,D.E., Bashor,C.J. and Collins,J.J. (2014) A brief history of synthetic biology. *Nat. Rev. Microbiol.*, **12**, 381–390.

2. Dahm,R. (2008) Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Hum. Genet.*, **122**, 565–581.

3. Watson,J.D. and Crick,F.H.C. (1953) Molecular structure of nucleic acids. *Nature*, **171**, 737–738.

4. Monod,J. and Jacob,F. (1961) Teleonomic Mechanisms in Cellular Metabolism, Growth, and Differentiation. *Cold Spring Harb. Symp. Quant. Biol.*, **26**, 389–401.

5. Jacob,F. and Monod,J. (1961) On the Regulation of Gene Activity. *Cold Spring Harb. Symp. Quant. Biol.*, **26**, 193–211.

6. Mullis,K.B. (1987) Process for amplifying nucleic acid sequences. US Patent 4683195.

7. Gavanji,S. (2013) Application of Recombinant DNA Technology – A review. *Applied Science Reports,* **2**, 29–31.

8. Ideker,T., Galitski,T. and Hood,L. (2001) A new approach to decode life: Systems Biology. *Annu. Rev. Genomics Hum. Genet.,* **2**, 343–372.

9. Bruggeman,F.J. and Westerhoff,H. V (2006) The nature of systems biology. *Trends Microbiol.,* **15**, 45–50.

10. Purnick,P.E.M. and Weiss,R. (2009) The second wave of synthetic biology: from modules to systems. *Nat. Rev. Mol. Cell Biol.*, **10**, 410–422.

11. Khalil,A.S. and Collins,J.J. (2010) Synthetic biology: applications come of age. *Nat. Rev. Gen.*, **11**, 367–379.

12. Kwok,R. (2010) Five hard truths for synthetic biology. *Nature,* **463**, 288–290.

13. Canton,B., Labno,A. and Endy,D. (2008) Refinement and standardization of synthetic biological parts and devices. *Nat. Biotech.,* **26**, 787–793.

14. Galdzicki,M., Clancy,K.P., Oberortner,E., Pocock,M., Quinn,J.Y., Rodriguez,C.A., Roehner,N., M, Wilson,M.L., Wilson,Y., *et al.* (2014) The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology. *Nat. Biotechnol.*, **32**, 545–550.

15. Hucka,M., Finney,A., Sauro,H.M., Bolouri,H., Doyle,J., Kitano,H., Bornstein,B.J., Bray,D., Cuellar,A.A., Dronov,S., *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics,* **19**, 524–531.

16. Salis,H.M., Mirsky,E. and Voigt,C. (2010) Automated Design of Synthetic Ribosome Binding Sites to Precisely Control Protein Expression. *Nat Biotechnol*, **27**, 946–950.

17. Gibson,D.G., Young,L., Chuang,R., Venter,J.C., Iii,A., Smith,H.O. and America,N. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 12–16.

18. Ran,F.A., Hsu,P.D., Lin,C. and Gootenberg,J.S. (2013) Resource Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell*, **154**, 1380–1389.

19. Agrestia,J.J., Antipov,E., Abate,A.R., Ahn,K., Rowat,A.C., Barete,J.-C., Marquez,M., Klibanov,A.M., Griffiths,A.D. and Weitza,D.A. (2010) Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proc. Natl. Acad. Sci. U. S. A*, **107**, 4004–4009.

20. Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotech.* **26**, 1135–1145.

21. Lashkari,D.A., Hunicke-Smith,S.P., Norgren,R.M., Davis,R.W. and Brennan,T. (1995) An automated multiplex oligonucleotide synthesizer: development of high-throughput, low-cost DNA synthesis. *Proc. Natl. Acad. Sci. U. S. A,* **92**, 7912–7915.

22. Gardner,T.S., Cantor,C.R. and Collins,J.J. (2000). Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, **403**, 339–342.

23. Elowitz,M.B. and Leibler,S. (1999) A synthetic oscillatory network of transcriptional regulators, *Nature*, **403**, 335–338.

24. Becskei,A. and Serrano,L., (2000) Engineering stability in gene networks by autoregulation. *Nature*, **405**, 590–593.

25. Balagadde,F.K., Song,H., Ozaki,J., Collins,C.H., Barnet,M., Arnold,F.H. and Quake,S.R. (2008) A synthetic *Escherichia coli* predator – prey ecosystem. *Mol. Syst. Biol.* **4**, 1–9.

26. Brenner,K., You,L. and Arnold,F.H. (2008) Engineering microbial consortia : a new frontier in synthetic biology. *Trends in Biotechno.,* **9**, 483–489.

27. Tamsir,A., Tabor,J.J. and Voigt,C.A. (2011) Robust multicellular computing using genetically encoded NOR gates and chemical 'wires'. *Nature*, **469**, 212–215.

28. Heinemann,M. and Panke,S. (2006) Systems biology Synthetic biology — putting engineering into biology. *Bioinformatics,* **22**, 2790–2799.

29. Paddon,C.J. and Keasling,J.D. (2014) Semi-synthetic artemisinin : a model for the use of synthetic biology in pharmaceutical development. *Nature Rev. Microbio.*, **12**, 355–367.

30. Paddon,C.J., Westfall,P.J., Pitera,D.J., Benjamin,K., Fisher,K., McPhee,D., Leavell,M.D., Tai, Main, Eng,D., *et al.* (2013) High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature*, **496**, 528–532.

31. Anderson,J.C., Clarke,E.J., Arkin,A.P., Voigt,C.A. and Berkeley,E.O.L. (2006) Environmentally controlled invasion of cancer cells by engineered bacteria. *J Mol Biol.,* **4**, 619–627.

32. Casini,A., Storch,M., Baldwin,G.S. and Ellis,T. (2015) Bricks and blueprints: methods and standards for DNA assembly. *Nat. Rev. Mol. Cell Biol.*, **9**, 568–576.

33. Vecchio,D. Del (2014) Modularity, context dependence, and insulation in engineered biological circuits. *Trends in Biotechnol.*, **33**, 111–119.

34. Koide,T., Pang,W.L. and Baliga,N.S. (2009) The role of predictive modelling in rationally re-engineering biological systems. *Nat. Rev. Microbiol.*, **7**, 297–305.

35. Hofer,U. (2014) Milestones in synthetic (micro)biology. *Net. Rev. Microb.,* **12**, 309.

36. Ausla,S. and Fussenegger,M. (2013) From gene switches to mammalian designer cells: present and future prospects. *Trends in Biotechnol.,* **31**, 155–168.

37. Karlsson,M. and Weber,W. (2012) Therapeutic synthetic gene networks. *Curr. Opin. Biotechnol.*, **23**, 703–711.

38. Endy,D. (2005) Foundations for engineering biology. *Nature*, **438**, 449–453.

39. Basu,S., Mehreja,R., Thiberge,S., Chen,M. and Weiss,R. (2004) Spatiotemporal control of gene expression with pulse-generating networks. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 6355–6360.

40. Tabor,J.J., Salis,H.M., Simpson,Z.B., Chevalier,A.A., Levskaya,A., Marcotte,E.M., Voigt,C.A. and Ellington,A.D. (2009) A Synthetic Genetic Edge Detection Program. *Cell*, **137**, 1272–1281.

41. Eisenstein,M. (2016) Living Factories of the Future. *Nature*, **531**, 401–403.

42. Martin,V.J.J., Pitera,D.J., Withers,S.T., Newman,J.D. and Keasling,J.D. (2003) Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. *Nature Biotech.***21**, 796–802.

43. Ro,D., Paradise,E.M., Ouellet,M., Fisher,K.J., Newman,K.L., Ndungu,J.M., Ho,K.A., Eachus,R.A., Ham,T.S., Kirby,J., *et al.* (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature,* **440**, 3–6.

44. Chen,M. and Weiss,R. (2005) Artificial cell-cell communication in yeast *Saccharomyces cerevisiae* using signaling elements from *Arabidopsis thaliana*. *Nature Biotech.,* **23**, 1551–1555.

45. Keung,A.J., Joung,J.K., Khalil,A.S. and Collins,J.J. (2015) Chromatin regulation at the frontier of synthetic biology. *Nat. Rev. Genetics*,**16**, 159-171.

46. 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**, 1–8.

47. Lam,F.H., Steger,D.J. and O'Shea,E.K. (2008) Chromatin decouples promoter threshold from dynamic range. *Nature*, **453**, 246–250.

48. Nikel,P.I., Martínez-garcía,E. and Lorenzo,V. De (2014) Biotechnological domestication of pseudomonads using synthetic biology. *Nat. Publ. Gr.*, **12**, 368–379.

49. Çelik,E. (2012) Production of recombinant proteins by yeast cells. *Biotechnol Adv.* **30**, 1108–1118.

50. Vogl,T., Hartner,F.S. and Glieder,A. (2013) New opportunities by synthetic biology for biopharmaceutical production in *Pichia pastoris*. *Curr. Opin. Biotechnol.*, **24**, 1094–1101.

51. Isaacs,F.J., Dwyer,D.J. and Collins,J.J. (2006) RNA synthetic biology. *Nature Biotechnology*. **24**, 545–554.

52. Qi,L.S. and Arkin,A.P. (2014) A versatile framework for microbial engineering using synthetic non-coding RNAs. *Nat. Publ. Gr.*, **12**, 341–354.

53. Cheng,A.A. and Lu,T.K. (2012) Synthetic Biology : an emerging engineering discipline. *Annu Rev Biomed Eng.*, **14**, 155–178.

54. Redden,H., Morse,N. and Alper,H.S. (2015) The synthetic biology toolbox for tuning gene expression in yeast. **15**, 1–10.

55. Berg,L., Strand,T.A., Valla,S. and Brautaset,T. (2013) Combinatorial mutagenesis and selection to understand and improve yeast promoters. *Biomed Res. Int.*, **2013**, 1–9.

56. Alper,H., Fischer,C., Nevoigt,E. and Stephanopoulos,G. (2005) Tuning genetic control through promoter engineering. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 12678–83.

57. Redden,H. and Alper,H.S. (2015) The development and characterization of synthetic minimal yeast promoters. *Nat. Commun.*, **7810**, 1–9.

58. Blazeck,J., Garg,R., Reed,B. and Alper,H.S. (2012) Controlling promoter strength and regulation in *Saccharomyces cerevisiae* using synthetic hybrid promoters. *Biotechnol. Bioeng.*, **109**, 2884–2895.

59. Hartner,F.S., Ruth,C., Langenegger,D., Johnson,S.N., Hyka,P., Lin-Cereghino,G.P., Lin-Cereghino,J., Kovar,K., Cregg,J.M. and Glieder,A. (2008) Promoter library designed for fine-tuned gene expression in *Pichia pastoris*. *Nucleic Acids Res.*, **36**, e76.

60. Mey,M. De, Maertens,J., Lequeux,G.J., Soetaert,W.K. and Vandamme,E.J. (2007) Construction and model-based analysis of a promoter library for *E . coli* : an indispensable tool for metabolic engineering. **14**, 1–14.

61. 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.*, **71**, 1-16.

62. Curran,K., 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**, 4002.

63. Lubliner,S., Keren,L. and Segal,E. (2013) Sequence features of yeast and human core promoters that are predictive of maximal promoter activity. *Nucleic Acids Res.*, **41**, 5569–5581.

Chapter 2

High resolution systematic mutations study of the *Pichia pastoris* *Alcohol Oxidase 1* (*AOX1*) core promoter

2.1. Abstract

Eukaryotic RNA Polymerase II (RNAPII) promoters have two parts: an upstream regulatory region (bound by specific Transcription Factors – TFs) and a core promoter (required for the binding of general TFs and RNAPII). Unravelling the core promoter sequence-function relationship is essential for the engineering of transcription initiation. This knowledge can then be used to fine--tune expression for Synthetic Biology and metabolic engineering applications.

Here we have performed systematic, high resolution nucleotide replacement studies of the core promoter and 5' Untranslated Region (5'UTR) of the exceptionally strong and tightly methanol regulated *Pichia pastoris Alcohol Oxidase 1* (*AOX1*) promoter at unprecedented resolution. Adjacent triplets of the 200bp long core promoter were mutated at a time by changing the wild type sequence into cytosine or adenine triplets. The resulting 130 variants were characterized with a reporter gene. Surprisingly, the *AOX1* core promoter tolerated the vast majority of the tested adjacent mutations with non-significant expression changes. Only mutations in the TATA box motif, regions downstream of the transcription start site or next to the start codon in the 5'UTR had a significant effect on the promoter strength. Regulation (repression/induction profiles) of the promoter variants were unaffected by most core promoter mutations. Only changes in the transition to upstream activating sequences led to a derepressed expression pattern. This suggests the existence of repressor binding sites for near the TATA box. Overall, our findings indicate that yeast core promoters show a high tolerance towards point mutations, supporting regulatory models of degenerate motifs or redundant design.

Keywords

Pichia pastoris · *Alcohol Oxidase 1* (*AOX1*) · Gene expression · Transcriptional fine-tuning · Core promoter library

2.2. Introduction

Eukaryotic promoters are generally constituted by two parts: an Upstream Regulatory Sequence (URS, also termed *cis* Regulatory Modules, CRMs (1–3)) and a core promoter. CRMs contain Transcription Factor Binding Sites (TFBSs), that are bound by specific Transcription Factors (TFs), conferring, for example, cell-cycle or carbon source dependent regulation (4–6). In contrast, the core promoter controls transcription initiation, as RNA Polymerase II (RNAPII) and general TFs bind to this region (5, 6). Gaining insights on core promoter sequence-function relationship is key for understanding transcription initiation and for generating core promoters' variants to fine-tune expression for Synthetic Biology and metabolic engineering applications (7–9).

In higher eukaryotes, synthetic core promoters have been designed based on commonly occurring motifs such as TATA box, Inr, DPE and MTE (10). In lower eukaryotes, namely yeasts, the only clearly conserved motif in core promoters is the TATA box (11). Proof-of-principle of

synthetic core promoter design has been established in yeasts (7, 8, 12). Studies in *Saccharomyces cerevisiae* and *Pichia pastoris* have been reported resorting to different methodologies, namely random mutagenesis (13), indels (insertions and deletions) of the 5' Untranslated Region (5'UTR) (14) and rationally designed core promoters based on one feature (*e. g.* nucleosome occupancy (7)) or one objective (*e. g.* minimal size (12)). Especially in *S. cerevisiae*, more fundamental core promoter studies were performed. Lubliner *et al*. (15) used a genome scale bioinformatic analysis to examine sequence features that correlate with core promoter strength. In another study (16), the core promoter sequence was modified in an attempt to infer factors at the basis of high promoter strength. It was also shown that the sequence in the vicinity of protein start codon is key to define the protein expression rate (17).

S. cerevisiae has historically been the experimental model system for studying transcriptional regulation (18). However, *P. pastoris*, one of the most commonly used expression hosts for heterologous protein production (19, 20), also offers exceptionally strong and tightly regulated promoters (21, 22). The most commonly used promoter in this yeast is the *Alcohol Oxidase 1* promoter (P_{AOX1}). It is tightly repressed in the presence of a carbon source (no traceable mRNA molecules in the presence of glucose) and highly inducible in the presence of methanol as only carbon source (21). Due to its importance and regulatory features, the P_{AOX1} CRM has been extensively studied (23–31). In contrast, there are few studies on the core promoter (8, 13) and the 5'UTR (14). Moreover, in *P. pastoris*, the best synthetic core promoter variant reached only 10% of the wild type *P. pastoris* P_{AOX1} (8), Here, we have thoroughly studied the *P. pastoris* P_{AOX1} core promoter by a high resolution mutagenesis approach.

2.3. Materials and methods

2.3.1. Strains, plasmids and cloning

The *P. pastoris* CBS7435 wild type strain was used for the expression studies. An *E. coli* TOP10 F' strain was used to perform the cloning work, which was based on the *P. pastoris/E. coli* shuttle vector pPpT4_SB-truncated*AOX1*-eGFP previously established for core promoter studies (8). The plasmid genbank file is available in the Supplementary File 2.1 (digital version) and the respective map is shown in Fig. 2.4. Initially, different versions of the truncated plasmid, with additional 50bp, 100bp and 150bp of the wild type *AOX1* core promoter and 5'UTR, were created to allow an easier insertion of mutations far from the Green Fluorescence Protein (eGFP) start codon. To this end, the truncated region was amplified with the three combinations of forward (p*AOX1*_Syn_dBamHI_SwaI-forward) and reverse (pTrunk*AOX1*-50-back, pTrunk*AOX1*-100-back and pTrunk*AOX1*-150-back) primers (all primer sequences are provided in Table 2.1). The original truncated plasmid was digested with SwaI and NheI and gel purified. Then, each one of the amplified truncated fragments was cloned by Gibson assembly into the previously digested plasmid and verified by Sanger sequencing (MicroSynth, Balgach, Switzerland).

The following procedure was applied to mutate every position of the *AOX1* core promoter. Three adjacent nucleotides of the *AOX1* core promoter were mutated for each construct, changing the wild type sequence into cytosine (CCC) and adenine (AAA) triplets. We used cytosine as a representative for a base forming three hydrogen bonds and adenine as representative for a base forming two hydrogen bonds. With these target mutations, we were able to generate transitions (C↔T and A↔G) and transversions (A↔C, A↔T, C↔G, and T↔G) in every position with thymine or guanine, while positions where adenine or cytosine were present were subject to transversions only. The core promoter was delimited to 200bp upstream of the start codon, as previously described (8, 32). A similar approach, as compared to the one used to create the different truncated plasmids, was used to insert the mutations. At first, the *AOX1* wild type promoter was PCR amplified with each combination of forward (p*AOX1*_Syn_dBamHI_SwaI-forward) and reverse primers (A#I, A#II, A#III, A#IV, C#I, C#II, C#III and C#IV). The name of the reverse primer stands for the nucleotide present in the mutation (adenine or cytosine), the position of the nucleotide to be mutated (starting from the last nucleotide present on the backbone, namely, the eGFP start codon for I primers, the -50bp for the II primer group, -100bp for the III primer group and -150bp for the IV primer group) and the truncated plasmid that would be used as backbone. In this case, I is the original truncated plasmid, II, III and IV are the truncated plasmids with 50bp, 100bp and 150bp of the wild type *AOX1* core promoter, respectively (Fig. 2.1). Additionally, the primers of each group share identical overhangs to be used in the Gibson assembly. Each of the plasmids was digested with SwaI and NheI, purified and ligated by Gibson assembly to each of the respective *AOX1* promoter fragments and verified by sequencing.

2.3.2. Transformation and cultivations conditions

P. pastoris cells were transformed with low amounts of the SwaI-linearized plasmid (1µg of DNA) using a condensed protocol (33). This amount of expression cassette was used to avoid multi copy integration and reduce the variability between transformants (8). Twenty eight transformants of each construct were screened using a previously reported method (8, 34). Briefly, cells were grown for 60h on 250µl BMD1 and subsequently induced with methanol (250µl BMM2 [1% methanol] after 60h and 50µl BMM10 [5% methanol] after 72h). The transformants were screened for uniformity and three representative transformants from the linear range of the landscape were selected for rescreening. Lastly, one transformant per construct was used for comparison with the other variants under the same growth conditions on separated 96 deep well plates. Biological replicates from three-fold cultivations of the same transformant were used to calculate the mean and standard deviations values, which are shown in Fig. 2.2 B-C and Fig. 2.3. These values represent the eGFP fluorescence values normalized per Optical Density at 600nm (OD600), where the background measurements of diluted medium were subtracted. eGFP fluorescence (excitation at 488nm and emission at 507nm) and absorption at OD600 were measured in micro titer plate, 48h after the first induction (Fig. 2.2 B-C) and before induction (Fig. 2.3).

2.4. Results and discussion

In this study, we analyzed the influence of point mutations on expression strength at high resolution over the whole core promoter sequence. Three adjacent nucleotides of the *AOX1* core promoter were mutated for each construct, changing the wild type sequence into cytosine (CCC) and adenine (AAA) triplets (Fig. 2.2 A). The core promoter region is known to be A/T rich (15), and replacements with three cytosines may disturb such A/T rich sequences more than adenine triplets. Adenine triplets were selected in addition, to assure that every position was changed at least once and to assess the effect of having a purine or a pyrimidine in every position.

To cover the whole 200bp long core promoter region (from start codon) with both mutations, 130 promoter variants were created (66xCCC and 64xAAA, excluding the cases where adenine or cytosine triplets were found in the natural wild type *AOX1* core promoter). We used three nucleotides to decrease the number of variants to test (at single nucleotide resolution, the same study would have required 400 variants and the measured effect, when a single position is altered, should be lower than the one observed here). The variants were fused to the P_{AOX1} CRM, cloned upstream of an eGFP reporter gene and transformed into *P. pastoris*. The mutations cover the whole promoter as a sliding window, in such a way that the last unchanged nucleotide in one strain is the first one in the next strain (Fig. 2.2 A). In this way, the whole core promoter was covered with the minimum number of constructs, representing, to the best of our knowledge, the highest resolution study of yeast core promoters.

One-way Analysis-Of-Variance (ANOVA) was performed to compare the wild type *AOX1* core promoter strength with each core promoter variant. Three groups were formed: nonsignificant expression change (*p-value*>0.05), significant expression change (*p-value*<0.05) and highly significant expression change (*p-value*<0.01).

Concerning the cytosine mutations, Fig. 2.2 B, most of the screened variants did not have a significant effect (48 strains – Fig. 2.2 B light blue) on reporter protein fluorescence when compared to the wild type core promoter (Fig. 2.2 B green). Nine variants showed a significantly different expression (*p-value*<0.05 – Fig. 2.2 B orange), and the remaining nine showed highly significant changes when compared to the wild type (*p-value*<0.01 – Fig. 2.2 B brown). Adenine mutations caused fewer highly significant changes (4 strains with *p*-value<0.01 – Fig. 2.2 C brown), while 15 strains showed significant changes when compared with the wild type (*p-value*<0.05 – Fig. 2.2 C orange). The remaining 45 adenine strains, showed no significant expression change (Fig. 2.2 C light blue). In Fig. 2.2 B and C, dark blue bars represent unmeasured strains, since the wild type sequence was the same as the mutated one (three naturally occurring cytosines or adenines).

Adenine and cytosine triplets' mutations causing significant reporter fluorescence changes can be divided in four groups: 1) transition region between core promoter and CRM, 2) TATA box, 3) downstream of the transcription start site and 4) near to start codon. The first cluster is represented in particular by mutations from -184bp to -196bp in the specific case of cytosine triplets

(C-184 and C-196), which show an increase in expression, notably also under derepressed conditions (Fig. 2.3 A). This increase may be explained by a mutation in a repressor binding site.

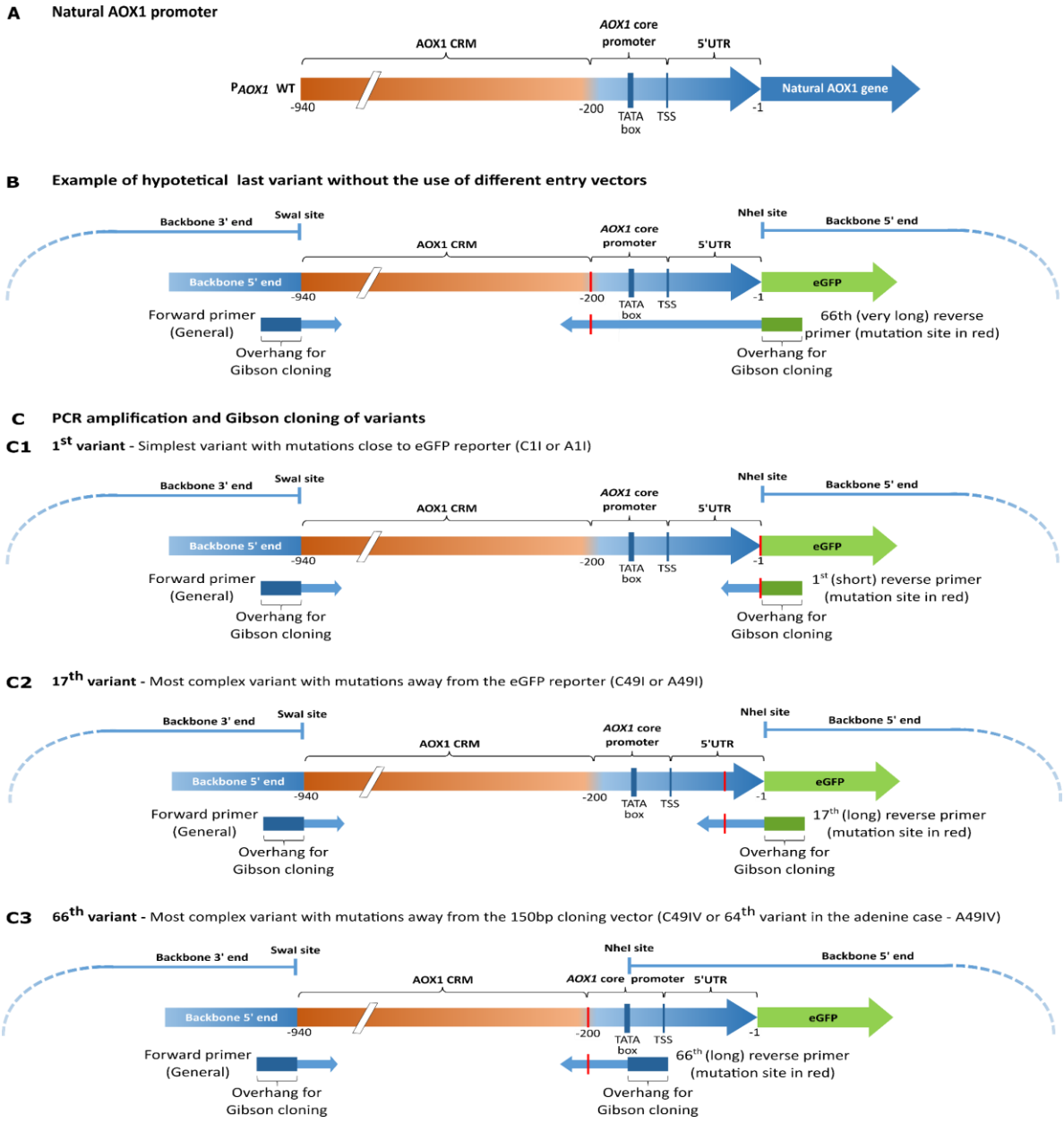

Fig. 2.1 – **Primer design rationale. A –** Sketch of the natural *AOX1* promoter (*cis* Regulatory Module – CRM – core promoter and 5' Untranslated Region – 5'UTR) controlling the expression of the *AOX1* gene. **B –** Example of the hypothetical last variant without the use of different entry vectors. In this case, the reverse primer would be very long, about 250bp. **C –** Three examples of primers used to add the adenine and cytosine mutations. The first primer (**C1**) has the mutation (red) next to the eGFP coding sequence. Its overhang for the Gibson cloning is the eGFP sequence (green). The longest primer used with the first entry vector (**C2**) has the mutation location away from the start codon. The subsequent primer (not shown) is a shorter primer to be used with the second entry vector, therefore it has an overhang sequence in the *AOX1* 5'UTR. The last primer (**C3**) is the longest primer used with the last entry vector, with the overhang sequence starting next to the *AOX1* TATA box and the mutation (red) in the end of the *AOX1* core promoter.

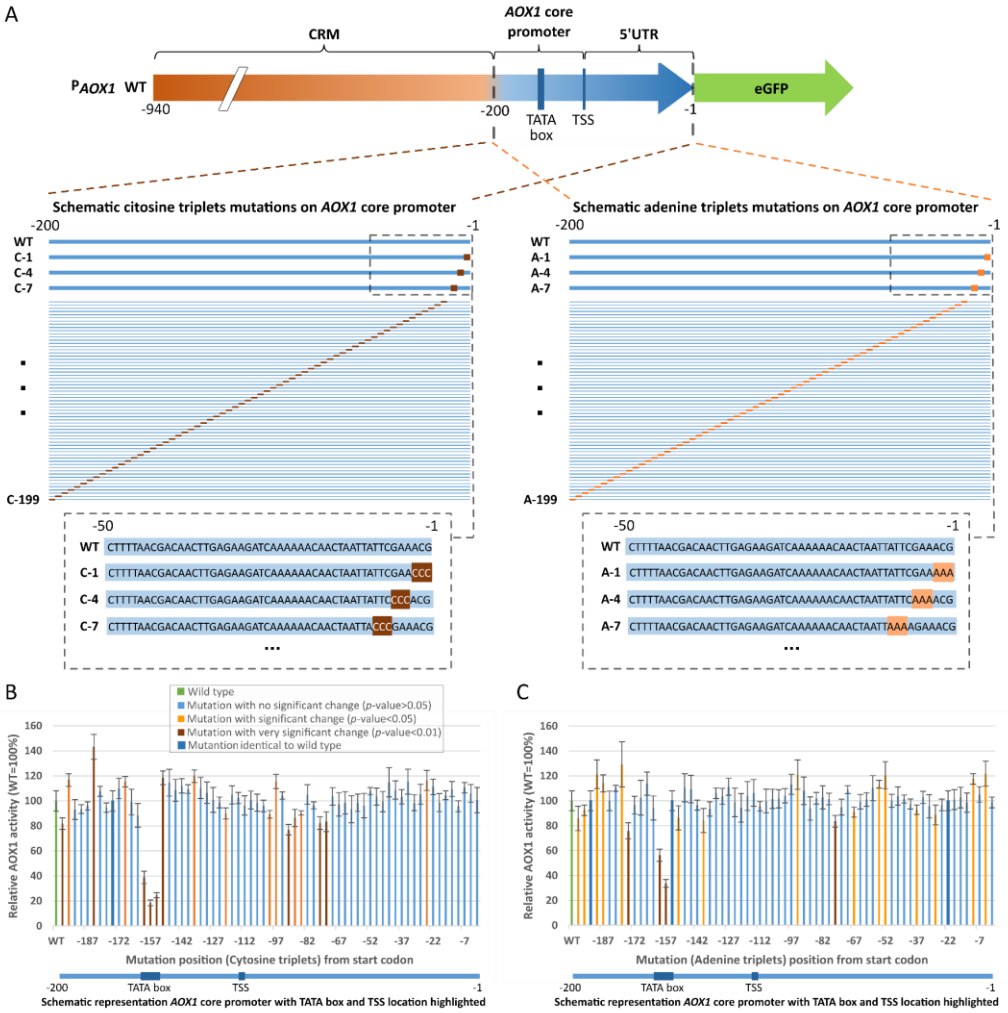

Fig. 2.2 – Design of the 130 *AOX1* core promoter variants (A) and respective reporter protein fluoresce measurements (B and C). A – Starting with the *AOX1* wild type sequence (consisting of *cis* Regulatory Module – CRM – core promoter and 5' Untranslated Region – 5'UTR), three adjacent wild type nucleotides were replaced by cytosines or adenines, at a time for each construct. 130 different constructs were created to cover the whole core promoter region (excluding the cases where natural occurring adenine and cytosine triples were found in the wild type *AOX1* promoter sequence). Overall, 64 and 66 constructs were created by replacing the respective wild type sequence with adenine and cytosine triplets, respectively. The difference between these constructs is the location of the mutations, they cover the whole promoter as a sliding windows, in such a way that the mutations are adjacent between consecutive constructs, so that whole core promoter is covered with the minimum number of constructs. The core promoter was delimited to 200bp upstream from the eGFP start codon (8, 32). The position of the TATA box and Transcriptional Start Site (TSS) are highlighted in the core promoter scheme; **B –** Promoter strength of 66 triple cytosine (CCC) constructs measured 48h after induction with BMM2 and BMM10 and schematic representation of the *AOX1* core promoter with TATA box and TSS location highlighted. Each one of the 66 constructs is named with a number (the position of the first of the three nucleotides mutated into cytosine, starting from the eGFP start codon). Activity of the wild type *AOX1* promoter is represented in green. Mean values and standard deviations of biological triplicates are shown. Mutant constructs with no significant change (*p-value*>0.05) in promoter strength are represented in light blue, while orange and brown represent the constructs with significant and highly significant changes, respectively (*p-value*<0.05 and *p-value*<0.01, respectively), in promoter strength.

Fig. 2.2 (cont.) – Dark blue bars represent unmeasured constructs, since the wild type promoter sequence was the same as the mutated one (naturally occurring three cytosines). **C –** Promoter strength of 64 triple adenines (AAA) constructs. Sampling approach and color legend of panel B also applies to panel C. Likewise, mean values and standard deviations of biological triplicates are shown.

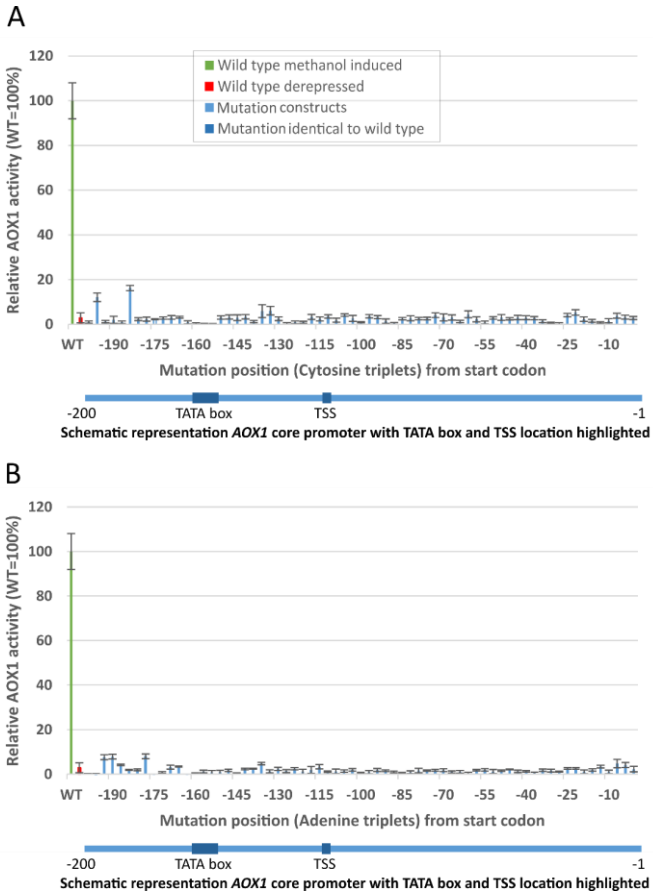

Fig. 2.3 – Reporter protein fluoresce measurements under derepressed conditions (before induction) of the 130 *AOX1* core promoter variants. A – Promoter strength of 66 triple cytosine (CCC) constructs measured before induction and schematic representation of the *AOX1* core promoter with TATA box and TSS location highlighted. Same numbering as in Fig. 2.2. Activity of the methanol induced wild type *AOX1* promoter is represented in green. The wild type promoter under derepressed conditions is shown as a red bar. Mean values and standard deviations of biological triplicates are shown. Mutant constructs in light blue. Dark blue bars represent unmeasured constructs, since the wild type promoter sequence was the same as the mutated one (naturally occurring three cytosines). **B –** Promoter strength of 64 triple adenines (AAA) constructs. Sampling approach and color legend of panel A also applies to panel B. Likewise, mean values and standard deviations of biological triplicates are shown.

A similar effect was noticed in this region when mutated to adenine, however, the derepression effect is not as intense (Fig. 2.2 C and Fig. 2.3 B). Mutations and deletions in this region have previously been reported to have an effect on regulation (13, 26, 30). Additionally, this derepression effect confirms that the whole core promoter region was covered, with the most upstream mutations already reaching CRM regions where specific regulators are binding, causing, in this case the described derepression. In accordance with previous studies reporting the TATA box as an important region on the core promoter (*e. g.* (16)), this region was also the most affected

by the mutations in our study. Even when just one nucleotide was changed (strain C-160 on Fig. 2.2 B) the promoter lost most of its functionality. Surprisingly, strain C-151 (Fig. 2.2 B) showed a highly significant increase in expression just next to TATA box when mutated to cytosine, suggesting that also sequences adjacent to the TATA box can have a profound influence on expression, as supported by (16).

Interestingly, mutations in TSS from -110bp to -116bp (represented by strains 109 and 112 in Fig. 2.2 B and C) did not significantly affect expression. However, downstream from this region seven strains from the cytosine mutations and three strains from the adenine mutations showed a significant expression change, suggesting that this area is more sensitive to mutations than the TSS itself.

The last region affecting reporter protein fluorescence was the area near the start codon. Here, the changes were more pronounced when the sequence is mutated to adenine. These results were expected, since the Kozak sequence (yeast genome wide consensus sequence directly upstream of the start codon (35)) strongly influences translation efficiency.

Lastly, it should be highlighted that the change in the expression was mostly due to the position of the mutation, rather than the mutation type (transition or transversion). Regardless of the mutation type, the variants with significant change in expression are clustered in these four regions, *i. e,* there are transitions and transversions examples in these regions with a significant effect on expression. Additionally, there were some unaffected regions (no change in expression outside of the mentioned four clustered regions) where adenine triplets were mutated to cytosine, and *vice versa*.

2.5. Conclusions

To the best of our knowledge, the present study represents the highest resolution systematic study of eukaryotic core promoters regarding the effect of up to three point mutations on gene expression. Our results show that the *AOX1* core promoter is remarkably robust, tolerating few point mutations over almost the whole core promoter sequence with non-significate expression changes. The only exception are the protein binding regions (*e. g.* TATA box), or regions with some biological importance (*e. g.* downstream of transcriptional start site and 5'UTR next to the protein start codon). The information obtained from this high resolution mutation studies, on its own, appears to be insufficient to fine-tune expression since the generated promoter library does not cover evenly a wide expression window. Yet, an approach that couples the information derived from some larger datasets on natural core promoters (15) with the testing of cumulative effects of those mutations that increased or decreased the promoter strength may constitute a promising approach.

Altogether these findings indicate that yeast core promoters show a high tolerance towards up to three point mutations, supporting regulatory models of degenerate regulatory motifs or redundant design.

2.6. Supplementary information

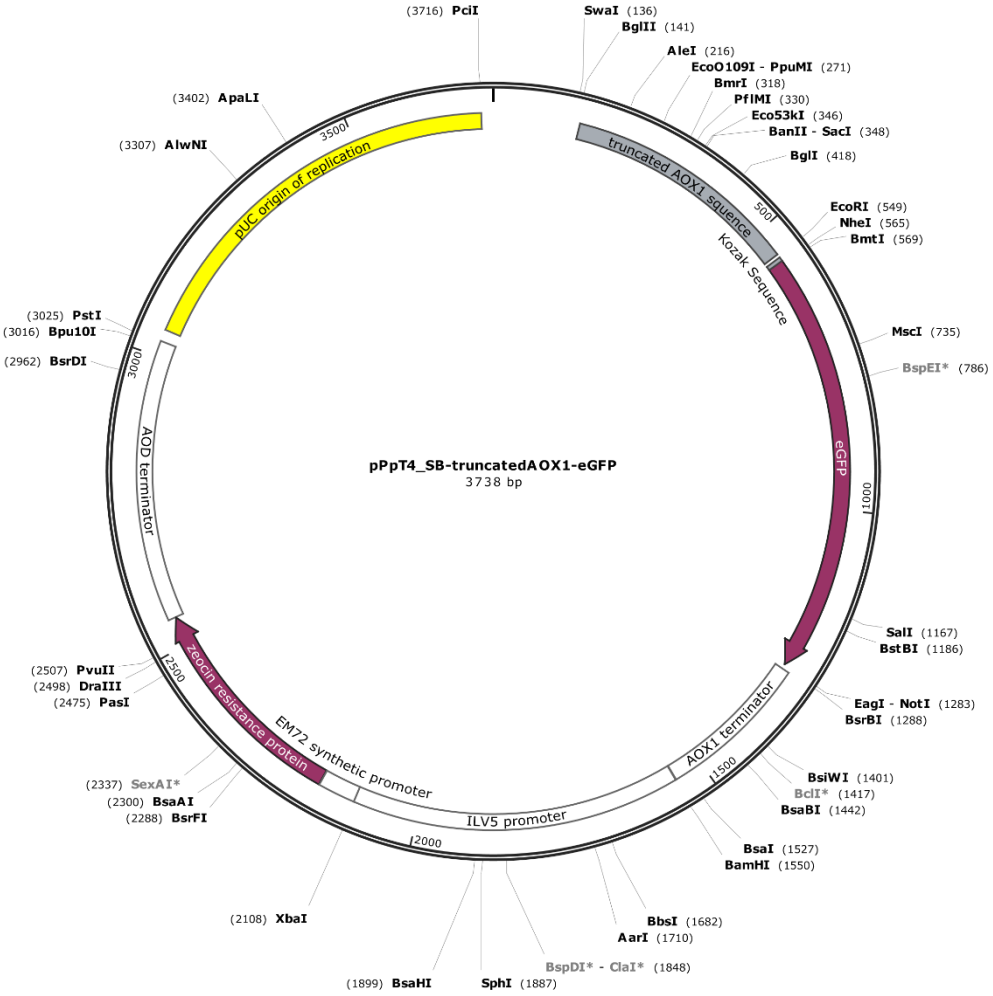

Fig. 2.4 – Map of *P. pastoris/E. coli* shuttle vector pPpT4_SB-truncated*AOX1*-eGFP with main features highlighted: Restriction enzymes, eGFP, zeocin resistance marker, promoters and terminators and origin of replication

Table 2.1 – List of primers used for Chapter 2.

Name according to mutated position	Name according to mutated position and insertion plasmid	Sequence
	p*AOX1*_Syn_dBamHI_SwaI-forward	GATCGGGAACACTGAAAAATACACAGTTATTATTCATTTAAATAGATCTAACATCCAAAGACGAAAGGTTGAATGAAAC
	pTrunk*AOX1*-50 back	CTCCTTTGCTAGCCATCGTTTCGAATAATTAGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAGAATTCGTC AGTTTTGGGCCATTTG
	pTrunk*AOX1*-100 back	TGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCAATTGGAACCAGTCGCAATTATGAAAGTAAGCT AAGAATTCGTCAGTTTTGGGCCATTTG
	pTrunk*AOX1*-150 back	GAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAAAAAAAGGTTTAAGACAGGGCAGCTTCCTTCTGTTTA GAATTCGTCAGTTTTGGGCCATTTG
C-1	**C1I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATGGGTTCGAATAATTAGTTGTTTTTTGATCTTCTCAAGTTGTC
C-4	**C4I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTGGGGAATAATTAGTTGTTTTTTGATCTTCTCAAGTTGTCG
C-7	**C7I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGGGTAATTAGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAG
C-10	**C10I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAAGGGTTAGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAG
C-13	**C13I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAAGGGGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTC
C-16	**C16I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGGGGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTC G
C-19	**C19I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGGGTTTTGATCTTCTCAAGTTGTCGTTAAAAGTC GTTAAAATC
C-22	**C22I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGTTGGGTGATCTTCTCAAGTTGTCGTTAAAAGTC GTTAAAATC
C-25	**C25I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGTTTTTGGGTCTTCTCAAGTTGTCGTTAAAAGTCG TTAAAATCAAAAG
C-28	**C28I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGTTTTTTGAGGGTCTCAAGTTGTCGTTAAAAGTC GTTAAAATCAAAAGC
C-31	**C31I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGTTTTTTGATCTGGGCAAGTTGTCGTTAAAAGTC GTTAAAATCAAAAGCTTG
C-34	**C34I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGTTTTTTGATCTTCTGGGGTTGTCGTTAAAAGTCG TTAAAATCAAAAGCTTGTC
C-37	**C37I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGTTTTTTGATCTTCTCAAGGGGTCGTTAAAAGTC GTTAAAATCAAAAGCTTGTCAATTG
C-40	**C40I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGTTTTTTGATCTTCTCAAGTTGGGGTTAAAAGTCG TTAAAATCAAAAGCTTGTCAATTGGAAC
C-43	**C43I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGTTTTTTGATCTTCTCAAGTTGTCGGGAAAAGTC GTTAAAATCAAAAGCTTGTCAATTGGAACC

Table 2.1 (cont.) – List of primers used for Chapter 2.

Name according to mutated position	Name according to mutated position and insertion plasmid	Sequence
C-46	C46I	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGTTTTTTGATCTTCTCAAGTTGTCGTTGGGAGTCG TTAAAATCAAAAGCTTGTCAATTGGAACC
C-49	C49I	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAGGGCG TTAAAATCAAAAGCTTGTCAATTGGAACCAG
C-52	C2II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTGGGTAAAATCAAAAGCTTGTCAATTGGAACCAGTCG
C-55	C5II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTGGGAATCAAAAGCTTGTCAATTGGAACCAGTCGC
C-58	C8II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAGGGCAAAAGCTTGTCAATTGGAACCAGTCGC
C-61	C11II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATGGGAAGCTTGTCAATTGGAACCAGTCGCAATTATG
C-64	C14II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAGGGCTTGTCAATTGGAACCAGTCGCAATTATG AAAG
C-67	C17II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGGGGGTCAATTGGAACCAGTCGCAATTATG AAAGT
C-70	C20II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGGGAATTGGAACCAGTCGCAATTATG AAAGTAAGC
C-73	C23II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCGGGTGGAACCAGTCGCAATTATG AAAGTAAGCTA
C-76	C26II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCAATGGGAACCAGTCGCAATTATG AAAGTAAGCTAATAATGATG
C-79	C29II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCAATTGGGGGCAGTCGCAATTATG AAAGTAAGCTAATAATGATGATAA
C-82	C32II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCAATTGGAACGGGTCGCAATTATGA AAGTAAGCTAATAATGATGATAAAAAAAAAGG
C-85	C35II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCAATTGGAACCAGGGGCAATTATG AAAGTAAGCTAATAATGATGATAAAAAAAAAGGTTTAAG
C-88	C38II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCAATTGGAACCAGTCGGGGTTATG AAAGTAAGCTAATAATGATGATAAAAAAAAAGGTTTAAGAC
C-91	C41II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCAATTGGAACCAGTCGCAAGGGTG AAAGTAAGCTAATAATGATGATAAAAAAAAAGGTTTAAGACAG
C-94	C44II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCAATTGGAACCAGTCGCAATTAGG GAAGTAAGCTAATAATGATGATAAAAAAAAAGGTTTAAGACAGG
C-97	C47II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCAATTGGAACCAGTCGCAATTATGA GGGTAAGCTAATAATGATGATAAAAAAAAAGGTTTAAGACAGGG
C-100	C50II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCAATTGGAACCAGTCGCAATTATGA AAGGGGGCTAATAATGATGATAAAAAAAAAGGTTTAAGACAGGG

Table 2.1 (cont.) – List of primers used for Chapter 2.

Name according to mutated position	Name according to mutated position and insertion plasmid	Sequence
C-103	**C3III**	AATTGGAACCAGTCGCAATTATGAAAGTAAGGGAATAATGATGATAAAAAAAAAGGTTTAAGACAGGGCAG
C-106	**C6III**	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTGGGAATGATGATAAAAAAAAAGGTTTAAGACAGGGCAGC
C-109	**C9III**	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATGGGGATGATAAAAAAAAAGGTTTAAGACAGGGCAGC
C-112	**C12III**	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGGGGATAAAAAAAAAGGTTTAAGACAGGGCAGCTTC
C-115	**C15III**	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGGGAAAAAAAAAGGTTTAAGACAGGGCAGCTTCC
C-118	**C18III**	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATGGGAAAAAAGGTTTAAGACAGGGCAGCTTCCTTC
C-121	**C21III**	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAGGGAAAGGTTTAAGACAGGGCAGCTTCCTTC
C-124	**C24III**	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAAAAGGGGGTTTAAGACAGGGCAGCTTCCTTC
C-127	**C27III**	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAAAAAAAGGGTTAAGACAGGGCAGCTTCCTTCT GTTTATAT
C-130	**C30III**	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAAAAAAAGGTGGGAGACAGGGCAGCTTCCTTCT GTTTATATATTG
C-133	**C33III**	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAAAAAAAGGTTTAGGGCAGGGCAGCTTCCTTCT GTTTATATATTGC
C-136	**C36III**	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAAAAAAAGGTTTAAGAGGGGGCAGCTTCCTTCT GTTTATATATTGCTGTC
C-139	**C39III**	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAAAAAAAGGTTTAAGACAGGGGAGCTTCCTTCT GTTTATATATTGCTGTCAAGTAG
C-142	**C42III**	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAAAAAAAGGTTTAAGACAGGGCGGGTTCCTTCT GTTTATATATTGCTGTCAAGTAGGG
C-145	**C45III**	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAAAAAAAGGTTTAAGACAGGGCAGCGGGCTTCT GTTTATATATTGCTGTCAAGTAGGG
C-148	**C48III**	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAAAAAAAGGTTTAAGACAGGGCAGCTTCGGGCT GTTTATATATTGCTGTCAAGTAGGGGTTAG
C-151	**C1IV**	AGGTTTAAGACAGGGCAGCTTCCTTGGGTTTATATATTGCTGTCAAGTAGGGGTTAGAACAG
C-154	**C4IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGGGGATATATTGCTGTCAAGTAGGGGTTAGAACAG
C-157	**C7IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTGGGTATTGCTGTCAAGTAGGGGTTAGAACAGTTA
C-160	**C10IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATAGGGTGCTGTCAAGTAGGGGTTAGAACAGTTAAAT
C-163	**C13IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATGGGTGTCAAGTAGGGGTTAGAACAGTTAAATTTTGATC

Table 2.1 (cont.) – List of primers used for Chapter 2.

Name according to mutated position	Name according to mutated position and insertion plasmid	Sequence
C-166	**C16IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCGGGCAAGTAGGGGTTAGAACAGTTAAATTTTGATCATG
C-169	**C19IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTGGGGTAGGGGTTAGAACAGTTAAATTTTGATCATGAAC
C-172	**C22IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTCAAGGGGGGGTTAGAACAGTTAAATTTTGATCATGAAC G
C-178	**C28IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTCAAGTAGGGGGGGAGAACAGTTAAATTTTGATCATGAA CGTTAGGCTATC
C-181	**C31IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTCAAGTAGGGGTTGGGACAGTTAAATTTTGATCATGAAC GTTAGGCTATCAG
C-184	**C34IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTCAAGTAGGGGTTAGAGGGGTTAAATTTTGATCATGAAC GTTAGGCTATCAGC
C-187	**C37IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTCAAGTAGGGGTTAGAACAGGGAAATTTTGATCATGAA CGTTAGGCTATCAGCAG
C-190	**C40IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTCAAGTAGGGGTTAGAACAGTTGGGTTTTGATCATGAAC GTTAGGCTATCAGCAGT
C-193	**C43IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTCAAGTAGGGGTTAGAACAGTTAAAGGGTGATCATGAA CGTTAGGCTATCAGCAGTATTC
C-196	**C46IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTCAAGTAGGGGTTAGAACAGTTAAATTTGGGTCATGAAC GTTAGGCTATCAGCAGTATTCC
C-199	**C49IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTCAAGTAGGGGTTAGAACAGTTAAATTTTGAGGGTGAA CGTTAGGCTATCAGCAGTATTCCC
A-1	**A1I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATTTTTTCGAATAATTAGTTGTTTTTTGATCTTCTCAAGTTGTC
A-4	**A4I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTTGAATAATTAGTTGTTTTTTGATCTTCTCAAGTTGTCG
A-7	**A7I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCTTTTAATTAGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAG
A-10	**A10I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATTTTTAGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAG
A-13	**A13I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTTGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTC
A-16	**A16I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTATTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCG
A-19	**A19I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTTTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGT TAAAATC
A-25	**A25I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGTTTTTTTTTCTTCTCAAGTTGTCGTTAAAAGTCGT TAAAATCAAAAGC

Table 2.1 (cont.) – List of primers used for Chapter 2.

Name according to mutated position	Name according to mutated position and insertion plasmid	Sequence
A-28	**A28I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGTTTTTTGATTTTCTCAAGTTGTCGTTAAAAGTCGT TAAAATCAAAAGC
A-31	**A31I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGTTTTTTGATCTTTTCAAGTTGTCGTTAAAAGTCGT TAAAATCAAAAGCTTG
A-34	**A34I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGTTTTTTGATCTTCTTTTGTTGTCGTTAAAAGTCGT TAAAATCAAAAGCTTGTC
A-37	**A37I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGTTTTTTGATCTTCTCAATTTGTCGTTAAAAGTCGT TAAAATCAAAAGCTTGTCAATTG
A-40	**A40I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGTTTTTTGATCTTCTCAAGTTTTTGTTAAAAGTCGT TAAAATCAAAAGCTTGTCAATTGGAAC
A-43	**A43I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGTTTTTTGATCTTCTCAAGTTGTCTTTAAAAGTCGT TAAAATCAAAAGCTTGTCAATTGGAACC
A-46	**A46I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGTTTTTTGATCTTCTCAAGTTGTCGTTTTTAGTCGT TAAAATCAAAAGCTTGTCAATTGGAACC
A-49	**A49I**	GTGAAAAGTTCTTCTCCTTTGCTAGCCATCTTTTCGAATAATTAGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAATTTCGT TAAAATCAAAAGCTTGTCAATTGGAACCAG
A-52	**A2II**	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTTTTTAAAATCAAAAGCTTGTCAATTGGAACCAGTCG
A-55	**A5II**	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTTTAATCAAAAGCTTGTCAATTGGAACCAGTCGC
A-58	**A8II**	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAATTTCAAAAGCTTGTCAATTGGAACCAGTCGC
A-61	**A11II**	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATTTTAAGCTTGTCAATTGGAACCAGTCGCAATTATG
A-64	**A14II**	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAATTTCTTGTCAATTGGAACCAGTCGCAATTATGA AAG
A-67	**A17II**	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGTTTGTCAATTGGAACCAGTCGCAATTATGA AAGT
A-70	**A20II**	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTTTTAATTGGAACCAGTCGCAATTATGA AAGTAAGC
A-73	**A23II**	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCTTTTGGAACCAGTCGCAATTATGA AAGTAAGCTA
A-76	**A26II**	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCAATTTTAACCAGTCGCAATTATGA AAGTAAGCTAATAATGATG
A-79	**A29II**	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCAATTGGTTTCAGTCGCAATTATGA AAGTAAGCTAATAATGATGATAA
A-82	**A32II**	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCAATTGGAACTTTTCGCAATTATGA AAGTAAGCTAATAATGATGATAAAAAAAAAGG

Table 2.1 (cont.) – List of primers used for Chapter 2.

Name according to mutated position	Name according to mutated position and insertion plasmid	Sequence
A-85	A35II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCAATTGGAACCAGTTTCAATTATGAA AGTAAGCTAATAATGATGATAAAAAAAAAAGGTTTAAG
A-88	A38II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCAATTGGAACCAGTCGTTTTTATGAA AGTAAGCTAATAATGATGATAAAAAAAAAAGGTTTAAGAC
A-91	A41II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCAATTGGAACCAGTCGCAATTTTGAA AGTAAGCTAATAATGATGATAAAAAAAAAAGGTTTAAGACAG
A-94	A44II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCAATTGGAACCAGTCGCAATTATTTA AGTAAGCTAATAATGATGATAAAAAAAAAAGGTTTAAGACAGG
A-97	A47II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCAATTGGAACCAGTCGCAATTATGAT TTTAAGCTAATAATGATGATAAAAAAAAAAGGTTTAAGACAGGG
A-100	A50II	AGTTGTTTTTTGATCTTCTCAAGTTGTCGTTAAAAGTCGTTAAAATCAAAAGCTTGTCAATTGGAACCAGTCGCAATTATGAA AGTTTGCTAATAATGATGATAAAAAAAAAAGGTTTAAGACAGGG
A-103	A3III	AATTGGAACCAGTCGCAATTATGAAAGTAATTTAATAATGATGATAAAAAAAAAAGGTTTAAGACAGGGCAG
A-106	A6III	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTTTTAATGATGATAAAAAAAAAAGGTTTAAGACAGGGCAGC
A-109	A9III	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATTTTGATGATAAAAAAAAAAGGTTTAAGACAGGGCAGC
A-112	A12III	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATTTTGATAAAAAAAAAAGGTTTAAGACAGGGCAGCTTC
A-115	A15III	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATTTTAAAAAAAAAAGGTTTAAGACAGGGCAGCTTCC
A-118	A18III	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATTTTAAAAAAGGTTTAAGACAGGGCAGCTTCCTTC
A-121	A21III	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAATTTAAAGGTTTAAGACAGGGCAGCTTCCTTC
A-124	A24III	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAAAATTTGGTTTAAGACAGGGCAGCTTCCTTC
A-127	A27III	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAAAAAAATTTTTAAGACAGGGCAGCTTCCTTCTGT TTATAT
A-130	A30III	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAAAAAAAGGTTTTAGACAGGGCAGCTTCCTTCTGT TTATATATTG
A-133	A33III	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAAAAAAAGGTTTATTTCAGGGCAGCTTCCTTCTGT TTATATATTGC
A-136	A36III	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAAAAAAAGGTTTAAGATTTGGCAGCTTCCTTCTGT TTATATATTGCTGTC
A-139	A39III	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAAAAAAAGGTTTAAGACAGTTTAGCTTCCTTCTGT TTATATATTGCTGTCAAGTAG
A-142	A42III	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAAAAAAAGGTTTAAGACAGGGCTTTTTCCTTCTGT TTATATATTGCTGTCAAGTAGGG

Table 2.1 (cont.) – List of primers used for Chapter 2.

Name according to mutated position	Name according to mutated position and insertion plasmid	Sequence
A-145	**A45III**	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAAAAAAAGGTTTAAGACAGGGCAGCTTTC TTCTGTTTATATATTGCTGTCAAGTAGGG
A-148	**A48III**	AATTGGAACCAGTCGCAATTATGAAAGTAAGCTAATAATGATGATAAAAAAAAAGGTTTAAGACAGGGCAGCTTCT TTCTGTTTATATATTGCTGTCAAGTAGGGGTTAG
A-151	**A1IV**	AGGTTTAAGACAGGGCAGCTTCCTTTTTTTTATATATTGCTGTCAAGTAGGGGTTAGAACAG
A-157	**A7IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTTTTTATTGCTGTCAAGTAGGGGTTAGAACAGTTA
A-160	**A10IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATTTTGCTGTCAAGTAGGGGTTAGAACAGTTAAAT
A-163	**A13IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTTTTGTCAAGTAGGGGTTAGAACAGTTAAATTTTGATC
A-166	**A16IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTTTCAAGTAGGGGTTAGAACAGTTAAATTTTGATCAT
A-169	**A19IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTTTTGTAGGGGTTAGAACAGTTAAATTTTGATCAT GAAC
A-172	**A22IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTCAATTTGGGGTTAGAACAGTTAAATTTTGATCAT GAACG
A-175	**A25IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTCAAGTATTTGTTAGAACAGTTAAATTTTGATCAT GAACGTTAGGC
A-178	**A28IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTCAAGTAGGGTTTAGAACAGTTAAATTTTGATCAT GAACGTTAGGCTATC
A-181	**A31IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTCAAGTAGGGGTTTTTACAGTTAAATTTTGATCAT GAACGTTAGGCTATCAG
A-184	**A34IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTCAAGTAGGGGTTAGATTTGTTAAATTTTGATCAT GAACGTTAGGCTATCAGC
A-187	**A37IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTCAAGTAGGGGTTAGAACATTTAAATTTTGATCAT GAACGTTAGGCTATCAGCAG
A-190	**A40IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTCAAGTAGGGGTTAGAACAGTTTTTTTTTGATCAT GAACGTTAGGCTATCAGCAGT
A-196	**A46IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTCAAGTAGGGGTTAGAACAGTTAAATTTTTTTCAT GAACGTTAGGCTATCAGCAGTATTCC
A-199	**A49IV**	AGGTTTAAGACAGGGCAGCTTCCTTCTGTTTATATATTGCTGTCAAGTAGGGGTTAGAACAGTTAAATTTTGATTTT GAACGTTAGGCTATCAGCAGTATTCCC

2.7. References

1. Allison,L.A. (2007) Transcription in eukaryotes. In *Fundamental Molecular Biology*.pp. 312–391.

2. Lelli,K.M., Slattery,M. and Mann,R.S. (2012) Disentangling the many layers of eukaryotic transcriptional regulation. *Annu. Rev. Genet.*, **46**, 43–68.

3. Istrail,S. and Davidson,E.H. (2005) Logic functions of the genomic *cis*-regulatory code. *Proc. Natl. Acad. Sci. U. S. A*, **102**, 4954–4959.

4. Juven-Gershon,T. and Kadonaga,J.T. (2010) Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev. Biol.*, **339**, 225–229.

5. Smale,S.T. and Kadonaga,J.T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem.*, **72**, 449–479.

6. Blazeck,J. and Alper,H.S. (2013) Promoter engineering: Recent advances in controlling transcription at the most fundamental level. *Biotechnol. J.*, **8**, 46–58.

7. 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**, 1–8.

8. Vogl,T., Ruth,C., Pitzer,J., Kickenweiz,T. and Glieder,A. (2014) Synthetic Core Promoters for *Pichia pastoris*. *ACS Synth. Biol.*, **3**, 188–191.

9. Nevoigt,E., Kohnke,J., Fischer,C.R., Alper,H., Stahl,U. and Stephanopoulos,G. (2006) Engineering of promoter replacement cassettes for fine-tuning of gene expression in *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol.*, **72**, 5266–5273.

10. Juven-Gershon,T., Cheng,S. and Kadonaga,J.T. (2006) Rational design of a super core promoter that enhances gene expression. *Nat. Methods*, **3**, 917–922.

11. Basehoar,A.D., Zanton,S.J. and Pugh,B.F. (2004) Identification and distinct regulation of yeast TATA box-containing genes. *Cell*, **116**, 699–709.

12. Redden,H. and Alper,H.S. (2015) The development and characterization of synthetic minimal yeast promoters. *Nat. Commun.*, **7810**, 1–9.

13. Berg,L., Strand,T.A., Valla,S. and Brautaset,T. (2013) Combinatorial mutagenesis and selection to understand and improve yeast promoters. *Biomed Res. Int.*, **2013**, 1–9.

14. Staley,C.A., Huang,A., Nattestad,M., Oshiro,K.T., Ray,L.E., Mulye,T., Li,Z.H., Le,T., Stephens,J.J., Gomez,S.R., *et al.* (2012) Analysis of the 5’ untranslated region (5'UTR) of the alcohol oxidase 1 (*AOX1*) gene in recombinant protein expression in *Pichia pastoris*. *Gene*, **496**, 118–127.

15. Lubliner,S., Keren,L. and Segal,E. (2013) Sequence features of yeast and human core promoters that are predictive of maximal promoter activity. *Nucleic Acids Res.*, **41**, 5569–5581.

16. Lubliner,S., Regev,I., Lotan-Pompan,M., Edelheit,S., Weinberger,A. and Segal,E. (2015) Core promoter sequence in yeast is a major determinant of expression level. *Genome Res.*, **25**, 1008–1017.

17. 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.

18. Hahn,S. and Young,E.T. (2011) Transcriptional regulation in *Saccharomyces cerevisiae*: Transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics*, **189**, 705–736.

19. Fickers,P. (2014) *Pichia pastoris*: a workhorse for recombinant protein production. *Current Research in Microbiology and Biotechnology* **2**, 354–363.

20. Ahmad,M., Hirz,M., Pichler,H. and Schwab,H. (2014) Protein expression in *Pichia pastoris*: Recent achievements and perspectives for heterologous protein production. *Appl. Microbiol. Biotechnol.*, **98**, 5301–5317.

21. Vogl,T. and Glieder,A. (2013) Regulation of *Pichia pastoris* promoters and its consequences for protein production. *New Biotechnol.*, **30**, 385–404.

22. Vogl,T., Sturmberger,L., Kickenweiz,T., Wasmayer,R., Schmid,C., Hatzl,A.-M., Gerstmann,M.A., Pitzer,J., Wagner,M., Thallinger,G.G., *et al.* (2015) A toolbox of diverse promoters related to methanol utilization – functionally verified parts for heterologous pathway expression in *Pichia pastoris*. *ACS Synth. Biol.*, 10.1021/acssynbio.5b00199.

23. Sahu,U., Krishna Rao,K. and Rangarajan,P.N. (2014) Trm1p, a Zn(II)2Cys6-type transcription factor, is essential for the transcriptional activation of genes of methanol utilization pathway, in *Pichia pastoris*. *Biochem. Biophys. Res. Commun.*, **451**, 158–164.

24. Ruth,C., Zuellig,T., Mellitzer, a., Weis,R., Looser,V., Kovar,K. and Glieder,A. (2010) Variable production windows for porcine trypsinogen employing synthetic inducible promoter variants in *Pichia pastoris. Syst. Synth. Biol.*, **4**, 181–191.

25. Wang,X., Wang,Q., Wang,J., Bai,P., Shi,L., Shen,W., Zhou,M., Zhou,X., Zhang,Y. and Cai,M. (2016) Mit1 Transcription Factor Mediates Methanol Signaling and Regulates *Alcohol Oxidase 1* Promoter in *Pichia pastoris*. *J. Biol. Chem.*, **291**, 6245–6261.

26. Lin-Cereghino,G.P., Godfrey,L., de la Cruz,B.J., Johnson,S., Khuongsathiene,S., Tolstorukov,I., Yan,M., Lin-Cereghino,J., Veenhuis,M., Subramani,S., *et al.* (2006) Mxr1p, a key regulator of the methanol utilization pathway and peroxisomal genes in *Pichia pastoris*. *Mol. Cell. Biol.*, **26**, 883–897.

27. Kranthi,B.V., Kumar,R., Kumar,N.V., Rao,D.N. and Rangarajan,P.N. (2009) Identification of key DNA elements involved in promoter recognition by Mxr1p, a master regulator of methanol utilization pathway in *Pichia pastoris*. *Biochim. Biophys. Acta - Gene Regul. Mech.*, **1789**, 460–468.

28. Kumar,N.V. and Rangarajan,P.N. (2012) The zinc finger proteins Mxr1p and repressor of phosphoenolpyruvate carboxykinase (ROP) have the same DNA binding specificity but regulate methanol metabolism antagonistically in *Pichia pastoris*. *J. Biol. Chem.*, **287**, 34465–34473.

29. Ohi,H., Miura,M., Hiramatsu,R. and Ohmura,T. (1994) The positive and negative *cis*-acting elements for methanol regulation in the *Pichia pastoris AOX2* gene. *Mol. Gen. Genet. MGG*, **243**, 489–499.

30. Hartner,F.S., Ruth,C., Langenegger,D., Johnson,S.N., Hyka,P., Lin-Cereghino,G.P., Lin-Cereghino,J., Kovar,K., Cregg,J.M. and Glieder,A. (2008) Promoter library designed for fine-tuned gene expression in *Pichia pastoris*. *Nucleic Acids Res.*, **36**, e76.

31. Xuan,Y., Zhou,X., Zhang,W., Zhang,X., Song,Z. and Zhang,Y. (2009) An upstream activation sequence controls the expression of *AOX1* gene in *Pichia pastoris*. *FEMS Yeast Res.*, **9**, 1271–1282.

32. Hartner,F., Ruth,C., Langenegger,D., Johnson,S.N., Hyka,P., Lin-Cereghino,G.P., Lin-Cereghino,J., Kovar,K., Cregg,J.M. and Glieder,A. (2008) Promoter library designed for fine-tuned gene expression in *Pichia pastoris*. *Nucleic Acids Res.*, **36**, 1–15.

33. Lin-Cereghino,J., Wong,W.W., Xiong,S., Giang,W., Luong,L.T., Vu,J., Johnson,S.D. and Lin-Cereghino,G.P. (2005) Condensed protocol for competent cell preparation and transformation of the methylotrophic yeast *Pichia pastoris*. *Biotechniques*, **38**, 44–48.

34. Weis,R., Luiten,R., Skranc,W., Schwab,H., Wubbolts,M. and Glieder,A. (2004) Reliable high-throughput screening with *Pichia pastoris* by limiting yeast cell death phenomena. *FEMS Yeast Res.*, **5**, 179–189.

35. Kozak,M. (1987) An analysis of 5’-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.*, **15**, 8125–48.

Chapter 3

Synthetic core promoters and 5' untranslated regions as universal parts for fine-tuning expression in different yeast

This chapter was based on the paper: Synthetic core promoters as universal parts for fine-tuning expression in different yeast species, Rui M. C. Portela, Thomas Vogl, Claudia Kniely, Jasmin Elgin Fischer, Rui Oliveira, and Anton Glieder, *ACS Synthetic Biology*, DOI: 10.1021/acssynbio.6b00178

3.1. Abstract

Synthetic Biology and metabolic engineering experiments frequently require the fine-tuning of gene expression to balance and optimize protein levels of regulators or metabolic enzymes. A key concept of Synthetic Biology is the development of modular parts that can be used in different contexts. Here, we have applied a computational multi-factor design approach to generate *de novo* synthetic core promoters and 5' Untranslated Regions (5'UTRs) for yeast cells. In contrast to upstream *cis* Regulatory Modules (CRMs), core promoters are typically not subject to specific regulation making them ideal engineering targets for gene expression fine-tuning. Here, 112 synthetic core promoter sequences were designed based on the sequence/function relationship of natural core promoters, nucleosome occupancy and the presence of short motifs. The synthetic core promoters were fused to the *Pichia pastoris AOX1* CRM and the resulting activity spanned a more than 200-fold range (0.3% to 70.6% of the wild type *AOX1* level). The top-ten synthetic core promoters with highest activity were fused to six additional CRMs (three in *P. pastoris* and three in *Saccharomyces cerevisiae*). Inducible CRM constructs showed significantly higher activity than constitutive CRMs, reaching up to 176% of natural core promoters. Comparing the activity of the same synthetic core promoters fused to different CRMs revealed high correlations only for CRMs within the same organism. These data suggest that modularity is to some extend maintained but, only within the same organism. Due to the conserved role of eukaryotic core promoters, this rational design concept may be transferred to other organisms as generic engineering tool.

Keywords

Yeast · Synthetic Biology · Core promoter · Promoter library · Transcriptional fine-tuning · Computational rational design

3.2. Introduction

Metabolic pathways and genetic circuits are commonly introduced into microbes such as *Saccharomyces cerevisiae* or *Escherichia coli* to produce chemicals or to implement novel functions (1, 2). Such experiments typically require the fine-tuning of gene expression to balance and optimize protein levels of metabolic enzymes or regulators. In prokaryotes, protein production can be controlled in a relatively easy manner using synthetic Ribosomal Binding Sites (RBSs) (3). However, to fine-tune gene expression and protein levels in unicellular eukaryotes, transcription is the most targeted step (4–7) and, to this end, various engineering tools have been developed (4, 8–10). Most promoter engineering efforts in eukaryotes were focused on yeasts, since they are the most commonly used eukaryotic expression systems for complex multi-gene pathways (11–13). *S. cerevisiae* has most commonly been used for metabolic engineering endeavors. However, recently other alternative yeasts, such as *Pichia pastoris*, have been increasingly used (14, 15). Yeast promoter libraries were designed either by random sequence modifications (9, 16) or by rational approaches (8, 17–19) with a focus on *cis* Regulatory Modules (CRMs) (20). CRM is a general

term referring to regulatory DNA sequences, also named enhancers in higher eukaryotes, while in yeasts rather the terms Upstream Activating/Repressing Sequences (UAS/URS) are used (20, 21). CRMs interact with particular transcription factors conferring specific activation/repression regulatory profiles.

CRMs alone are however nonfunctional, requiring a core (minimal) promoter sequence to recruit general transcription factors and RNA polymerase II for transcription initiation (4, 22, 23). Similarly, the core promoter alone results in basal to no expression at all, and requires a CRM for strong expression and specific regulation. Engineering the core promoter and 5' untranslated region (5'UTR) has mainly an impact on transcription strength, translation initiation and most probably mRNA stability. In contrast, engineering CRMs affects transcription strength but also impacts regulation (*i. e.,* constitutive or inducible). For instance, studies on the methanol inducible *Alcohol Oxidase 1* (*AOX1)* promoter (P_{AOX1}) in *P. pastoris* (8, 24, 25) showed that deletions or insertions of CRMs (more specifically in predicted Transcription Factor Binding Sites, TFBSs) resulted in promoter activity variations and also in regulatory differences. One example for altered regulation were derepressed P_{AOX1} variants (8). The wild type P_{AOX1} is tightly repressed on glucose, remaining repressed even when glucose is depleted and strictly requires methanol for induction. Depressed variants however start expression once glucose is depleted not requiring methanol induction.

In contrast to such mutations in CRMs, modifications of the core promoter sequence impacted only promoter strength, leaving induction/repression profiles unchanged (25). Additionally, studies on CRMs are typically limited to one promoter, *i. e.*, their conclusions cannot be easily transferred to other promoters, even in the same organism. For instance, information gained from deletion studies of P_{AOX1} (8, 24, 26) cannot be transferred to other methanol inducible promoters in *P. pastoris* due to the low sequence similarly between these co-regulated promoters (5). In contrast, core promoter function is conserved even between related species (8, 27).

Hence, we hypothesized that *de novo* designed synthetic core promoters could be used as interchangeable parts between related organisms. Such universal "tuning knobs" could be used for regulating the strength of gene expression without interfering with specific regulation in a given organism for different promoters, or in different species. Since the designed promoters are artificial, they have lower probability of recombining with natural sequences in the genome. This favors strain stability and also facilitates the expression cassette assembly. To design such promoters, we used a genome scale data set available for *S. cerevisiae* (28, 29).

S. cerevisiae is the most commonly used yeast for basic research on transcription regulation and synthetic promoter design (4, 30, 31). Recently, comprehensive studies have also addressed the sequence/function relationship of natural core promoters (27, 28, 32–34) and 5'UTRs (34). Two genome-scale studies were performed in this yeast by measuring the expression of 859 natural promoters under different conditions (29) and using this data set to deduce core promoter properties affecting expression (28). Also, nucleosome affinity in the core promoter was shown to

be an effective modification target for designing core promoters (18). For the interspecies comparisons we selected *S. cerevisiae* and *P. pastoris*.

P. pastoris is, after *E. coli*, the most commonly used expression system for single proteins (35). The exceptionally strong and tightly methanol regulated P_{AOX1}, has motivated several research studies on transcriptional regulation mechanisms (reviewed by (36) and summarized in Supplementary Text S1 on page 58 alongside *S. cerevisiae* studies (7–9, 16, 18, 19, 24–27, 32–34, 37–44)). Recently, it has been reported that at least 15 promoters of genes involved in methanol utilization are co-regulated with P_{AOX1}, some of which show higher expression (5). Hence, *P. pastoris* offers one of the largest sets of promoters that are co-regulated and easily applicable strategies for regulating their strengths would be desirable.

In the present study, we designed generic synthetic promoters for protein production fine--tuning in yeasts. Acknowledging the fact that manifold structural features contribute to the promoter strength, we have incorporated in our design several factors, which were derived from a *S. cerevisiae* core promoter data-set (*e. g.,* TATA box position and nucleosome affinity) (28, 29). Using this design approach, we have created a library of 112 synthetic core promoters and 5'UTRs that were validated with the *P. pastoris* P_{AOX1} CRM ($P_{AOX1\text{-}R}$). Additionally, we tested the best performing synthetic core promoters with alternative CRMs of *P. pastoris* and *S. cerevisiae* promoters, demonstrating their applicability in different contexts.

3.3. Materials and methods

3.3.1. Strains

The *P. pastoris* CBS7435 (*Komagataella phaffii*, NRLLY-11430 (60)) wild type strain and the *S. cerevisiae* FY 1679-01B strain (isogenic to *S. cerevisiae* S288c with an uracil auxotrophy (61)) were used as host organisms to screen the synthetic promoter activity, while *E. coli* TOP10 F' was used to perform the cloning work.

3.3.2. Vectors and cloning - Controls and synthetic core promoters fused to the $P_{AOX1\text{-}R}$

Ten different controls were created using the genomic wild type P_{AOX1} sequence as template: deletion of the entire upstream regulatory region (CRM) upstream of the core promoter, deletion of the core promoter, replacement of the natural *AOX1* core promoter with the core promoter of the *HHF2* gene (46, 47) and seven completely random sequences. For the first control (deletion of CRM) primers C-WO-CRM1 and eGFP-p*AOX1*-3prime were used. For the remaining controls, p*AOX1*_Syn_dBamHI_SwaI-forward was used as forward primer, while as reverse primers were C-WO-Core1, C-W-HHF2+10 and R1 to R7, respectively. The primers sequences are provided in Table 3.1.

The synthetic core promoters were ordered as long primers (Ultramer® DNA Plate Oligo by Integrated DNA Technologies (Leuven, Belgium) in 96 wells microtiter plates), attached by PCR to

the $P_{AOX1\text{-}R}$ and cloned into the *P. pastoris/E. coli* shuttle vector pPpT4_SB-truncated*AOX1*-eGFP, reported by Vogl *et al.* (25). The plasmid genbank file and respective map are available in the Supplementary File 3.1 (digital version) and Fig. 3.6. The synthetic promoters were amplified using forward primer p*AOX1*_Syn_dBamHI_SwaI-forward and the reverse primers listed in Table 3.2 to Table 3.5.

The final PCR product was gel purified and cloned by assembly cloning into the SwaI and NheI digested vector backbone. All constructs were verified by Sanger sequencing.

3.3.3. Controls and entry vectors to assess synthetic core promoters with different CRMs in *P. pastoris* and *S. cerevisiae*

The best synthetic core promoters were tested when fused to the CRMs of six additional promoters (*CAT1, DAS1, GAP, ADH1, GAL1* and *GPD1*, named $P_{CAT1\text{-}R}$, $P_{DAS1\text{-}R}$, $P_{GAP\text{-}R}$, $P_{ScADH\text{-}R}$, $P_{ScGAL1\text{-}R}$ and $P_{ScGPD1\text{-}R}$, respectively). Three CRMs were tested in *P. pastoris* ($P_{CAT1\text{-}R}$, $P_{DAS1\text{-}R}$ and $P_{GAP\text{-}R}$), while the remaining three were tested in *S. cerevisiae* ($P_{ScADH\text{-}R}$, $P_{ScGAL1\text{-}R}$ and $P_{ScGPD1\text{-}R}$). At first, the positive controls were created. To do so, the genomic wild type sequences of the *P. pastoris* promoters were amplified using the following three primers groups: CAT-core and CAT-CRM-forw, DAS-core and DAS-CRM-forw and GAP-core and GAP-CRM-forw (Table 3.6), resulting in promoter fragments of 500, 552, and 486bp, respectively. In each of the three PCR reactions, the respective wild type whole promoter sequence was used as template. It was then cloned into the *P. pastoris/E. coli* shuttle vector used in the previous screening, where the *AOX1* truncated sequence had been removed (digestion with SwaI and NheI). For the *S. cerevisiae* whole promoter plasmids (used as positive controls), the promoter sequences were amplified from *S. cerevisiae* genomic DNA and cloned into a reporter vector (named Sc_eGFP_RFP_ARS) comprised by pUC origin of replication for *E. coli*, the ARS/CEN sequence for low-copy replication in *S. cerevisiae*, URA3-3' and URA3-5' integration sequences, a stuffer sequence flanked by enhanced Green Fluorescence Protein (eGFP) and Red Fluorescence Protein (RFP) and the two transcriptional terminators *PRM9* and *SPG5.* It was further composed by a Kanamycin resistance cassette which was comprised by: *TEF1* and *EM72* promoters for expression in yeast and *E. coli*, respectively, the *KanMX6* resistance gene and terminator TIF51A (plasmids kindly provided by Pitzer, J., unpublished results). The plasmid genbank file is available in the Supplementary File 3.2 (digital version) and the respective map is shown in Fig. 3.8.

For each CRM, an entry vector was created to facilitate cloning of the synthetic core promoter fusions. Such entry vectors had a CRM sequence (without core promoter), a placeholder fragment and the eGFP coding sequence. The primers used to amplify the CRMs sequences for *P. pastoris* were the following three groups: CAT-CRM-rev and CAT-CRM-forw, DAS-CRM-rev and DAS-CRM-forw and GAP-CRM-rev and GAP-CRM-forw (Table 3.6). While for *S. cerevisiae* CRMs sequences amplification the reverse primer used were: ADH-CRM-rev, GAL-CRM-rev and GPD-CRM-rev (Table 3.5). The forward primer was, in these three cases, seqTomato19-41rev. The backbones used were Sc_eGFP_RFP_ARS for *S. cerevisiae* and pPpT4-bidi-sTomato-eGFP

(Vogl, T., unpublished results) for *P. pastoris* (both genbank files are available in Supplementary File 3.2 and 3.3 (digital version) and respective maps in Fig. 3.7 and Fig. 3.8). The *S. cerevisiae* vector was digested with AscI while the *P. pastoris* vector was linearized with AscI and SwaI. The digestion was gel purified and an assembly cloning was performed for each of the PCR results, yielding six entry vectors (one for each CRM) and three plasmids containing a wild type promoter of interest each (P_{CAT1}, P_{DAS1}, P_{GAP} – *P. pastoris*), which were verified by Sanger sequencing.

3.3.4. Cloning a subset of synthetic core promoters with different CRMs in *P. pastoris* and *S. cerevisiae*

Each of the ten best synthetic core promoters identified with the $P_{AOX1\text{-}R}$ (T22, T23, T24, T25, T26, T27, T28, M28, A27 and A28) was amplified by PCR six times to include the different CRMs overhangs to be used for assembly cloning. The reverse primers used for each of the 10 best core promoters were T22-GFP-rev, T23-GFP-rev, T24-GFP-rev, T25-GFP-rev, T26-GFP-rev, T27-GFP-rev, T28-GFP-rev, M28-GFP-rev, A27-GFP-rev and A28-GFP-rev. Different forward primers were used depending on the CRM to be fused. For instance, to amplify the 10 synthetic promoters to be cloned in the $P_{CAT1\text{-}R}$ plasmid, the following forward primers were used: T22-CAT-rev, T23-CAT-rev, T24-CAT-rev, T25-CAT-rev, T26-CAT-rev, T27-CAT-rev, T28-CAT-rev, M28-CAT-rev, A27-CAT-rev and A28-CAT-rev. The three different entry vectors for *P. pastoris* containing the $P_{GAP\text{-}R}$, $P_{DAS1\text{-}R}$ and $P_{CAT1\text{-}R}$ were digested by AscI and NheI to remove the placeholder fragment. The digestion products were gel purified. The linearized plasmids were used for assembly cloning with each of the respective 10 PCR core promoter fragments.

A similar approach was performed to screen the top 10 synthetic promoters in *S. cerevisiae*. The synthetic core promoters used the same reverse primers, while the forward primers vary according to the CRM sequence as explained above. The entry vectors containing the $P_{ScADH\text{-}R}$, $P_{ScGAL1\text{-}R}$ and $P_{ScGPD1\text{-}R}$ were digested by AscI and NheI to remove the placeholder fragment. They were gel purified. The linearized plasmids were used for assembly cloning with each of the respective 10 PCR core promoter fragments.

All the primers used to clone the ten synthetic promoters with highest activity with different CRMs in *P. pastoris* and *S. cerevisiae* and the respective entry vectors are listed in Table 3.6.

3.3.5. Transformation of *P. pastoris* and cultivations conditions

The aforementioned plasmids were digested with SwaI for linearization. *P. pastoris* was transformed with low amounts of linearized plasmid (approximately 1µg of DNA) using the condensed protocol reported by Lin-Cereghino *et al.* (62). This low amount of expression cassette was used to reduce multi copy integration and variability between transformants (25). Then, from the resulting transformants, 28 were screened using a previously reported high throughput method (25, 63). Briefly, cells were grown for 60h on 250µl BMD1 and subsequently induced with methanol (250µl BMM2 [1% methanol] at 60h and 50µl BMM10 [5% methanol] at 72h). The transformants were screened for uniformity and three representative transformants from the linear range of the landscape were selected for rescreening using the same protocol. Lastly, one transformant per

construct was used for comparison of the variants under the same growth conditions. Biological replicates from at least three-fold cultivations of the same transformant were used to calculate the mean and standard deviations values, which are shown in Fig. 3.2 to Fig. 3.5. These values represent the eGFP fluorescence values normalized per Optical Density at 600nm (OD600), where the background measurements of diluted medium were subtracted. eGFP fluorescence (excitation at 488nm and emission at 507nm) and OD600 were measured in micro titer plates, 48h after the first induction for the methanol inducible promoters (derived of $P_{AOX1\text{-}R}$, $P_{CAT1\text{-}R}$ *and* $P_{DAS1\text{-}R}$), while the fluorescence values of the constitutive $P_{GAP1\text{-}R}$ variants were taken 48h after the inoculation.

3.3.6. Transformation of *S. cerevisiae* and cultivations conditions

S. cerevisiae was transformed with circular plasmids (0.5µg of DNA) using chemically competent cells (64). Then, from the resulting transformants, 28 were screened using a similar protocol to the one used for *P. pastoris*. Briefly, cells were grown for 24h on 250µl YPD. The $P_{ScGAL1\text{-}R}$ variants were additionally screened using YPGal medium instead of YPD. The transformants were screened for uniformity and three representative transformants from the linear range of the landscape were selected for rescreening using the previous protocol. Lastly, one transformant per construct was used for comparison of the variants under the same growth conditions. Measurements were made in an identical way as to the *P. pastoris* protocol.

3.4. Results

3.4.1. Computational design of synthetic core promoters

Several factors were simultaneously incorporated in the synthetic core promoter design: *i*) nucleotide occurrence along the sequence of 140 strong natural *S. cerevisiae* core promoters (as reported by (28)), *ii*) the presence and position of the TATA box, *iii*) the position and number of other motifs (other than TATA box, as defined by (28)) and *iv*) nucleosome occupancy profiles (28, 45). Using this approach, we have created a library of synthetic core promoters and 5'UTRs for generic yeast cells. The method adopted in this study is represented schematically in Fig. 3.1.

The input sequences were taken from a library of *S. cerevisiae* natural core promoter sequences (28). We used the genome wide *S. cerevisiae* core promoter sequences data published by Lubliner *et al.* (28), in which 729 native *S. cerevisiae* promoters were segmented into four groups (low, medium, high and very high maximal expression). Subsequently, different structural features were examined. Namely, nucleotide frequency, nucleosome occupancy and presence/number of short motifs (up to four nucleotides). Lubliner *et al.* (28) showed that some of these features are highly predictive of maximal promoter activity, namely the high A and T content and TATA box like elements around the transcriptional start site (TSS). Also, it had previously been demonstrated that there is a correlation between promoter strength and low nucleosome affinity (18).

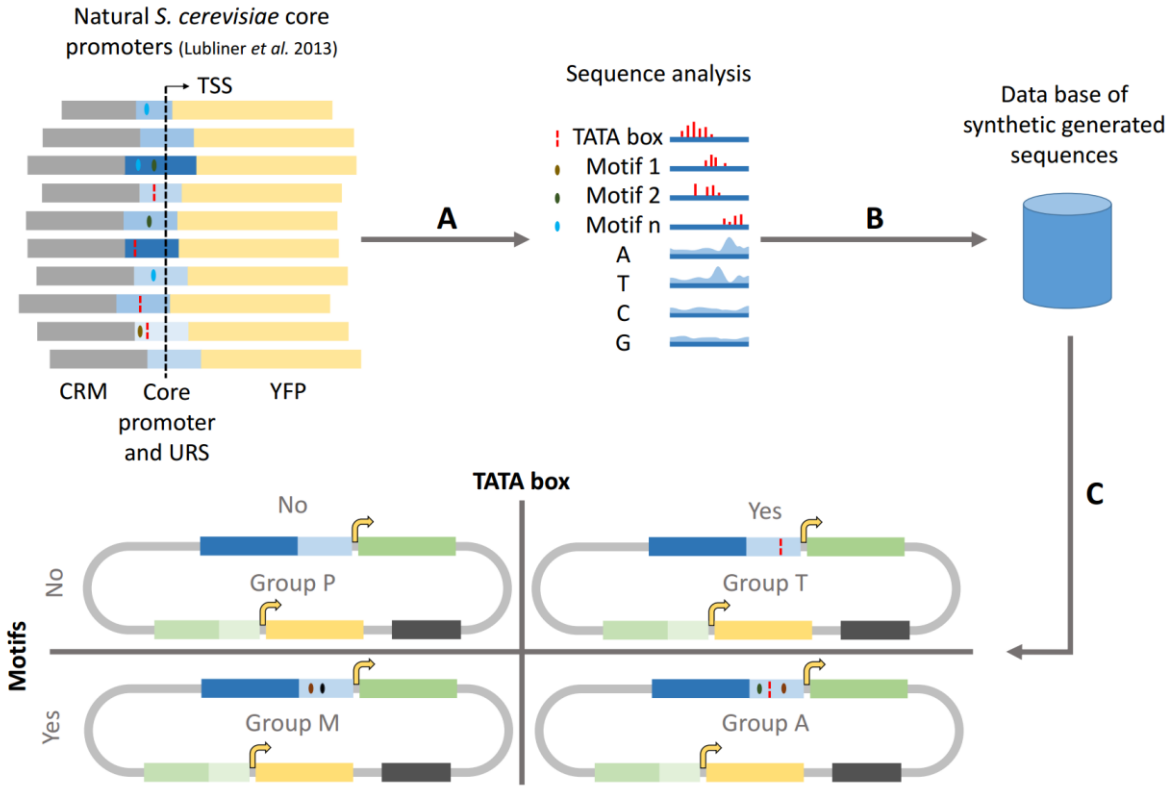

Fig. 3.1 – Design strategy for synthetic core promoters. Three steps were followed: **A –** Computation of: *i*) nucleotide probability distribution, *ii*) TATA box position, *iii*) position and frequency distribution of motifs with high impact on expression and *iv*) average nucleosome occupancy along the sequence of 140 *S. cerevisiae* natural strong core promoters (28) aligned by the transcription start site. **B –** Generation of 400 random sequences using information of nucleotide probability distribution. **C –** Partitioning of sequences in four groups (dubbed P, T, M and A), to which TATA boxes and motifs were added, according to the group they belonged to (group P: without TATA box nor motifs; group T: with TATA box and without motifs; group M: with motifs and without TATA box; group A: with TATA box and motifs). Subsequently, the nucleosome occupancy profile of each of the generated sequences was compared to the average profile for the natural strong promoters. The generated sequences with higher similarity to the average natural nucleosome occupancy were selected to be tested *in vivo*.

We reasoned that this data set (input sequences) could also be used in a reverse way to generate a model and create synthetic core promoters *de novo*. We started from the subset of 140 strong core promoters and the respective 5'UTRs. Firstly, we have selected sequences of 150bp (50bp downstream and 100bp upstream of the TSS) for analysis. Then, to extract important sequence features, we have applied the following computational procedure (Fig. 3.1 A):

i) Computation of the nucleotide probability distribution along the sequence, calculated with a 20bp windows size and 10bp windows step;

ii) Computation of the TATA box position distribution along the sequence;

iii) Computation of the position and frequency distribution of motifs along the sequence. Only the subset of motifs with highest effect (positive or negative) on the promoter strength were considered (defined by Lubliner *et al.* (28));

iv) Computation of the average nucleosome occupancy along the promoter sequence (using the software package by Kaplan *et al.* (45)).

Using this information, we have designed 4 groups (named P, T, M, A) of 28 sequences each for experimental screening (Fig. 3.1, Fig. 3.2 B and Table 3.2 to Table 3.5). They differ in the presence or absence of a TATA box and/or selected motifs (group P: without TATA box nor motifs; group T: with TATA box and without motifs; group M: with motifs and without TATA box; group A: with TATA box and motifs). In this way, the synthetic core promoters were named according to their group and to the respective measured activity, *i. e.*, the 4 groups with 28 sequences each were termed “P#”, “T#”, “M#”, “A#”, where the letters stand for P – (nucleotide) probability, T – TATA box, M – motifs and A – all, respectively. They were ordered in increasing expression strength. The general properties of the designed sequences are available in Table 3.7 to Table 3.10.

The sequences were computed in a 4-steps procedure as follows:

Step 1: Generation of 400 random sequences. To generate the initial synthetic sequences the information of nucleotide probability distribution only was used (Fig. 3.1 B). TATA boxes or any of the selected motifs were searched and replaced by a newly generated synthetic sequence. This procedure was repeated until no motif or TATA box were found in the generated sequences. Start codons upstream of the protein codon region were also removed to avoid frame shift mutations and different N-*termini* of the reporter protein. Lastly, due to the known relevance of the nucleotides adjacent to the start codon (34), this region was replaced by the P_{AOX1} Kozak sequence (CGAAACG) in the generated sequences. These 400 sequences were partitioned in four groups of 100 sequences each.

Step 2: Addition of a TATA box to groups T and A. The TATA box positioning followed a Gaussian distribution model with mean and standard deviation computed on the natural strong core promoter sequences. One TATA box was inserted per core promoter sequence (Fig. 3.1 C).

Step 3: Addition of motifs to groups M and A. The frequency of each motif in each sequence also followed a Gaussian distribution model inferred from the natural sequences, meaning that some motifs might be present more than once while others might be absent in a given sequence (Fig. 3.1 C).

Step 4: Design space reduction. Selection of 28 synthetic sequences, out of the 100 sequences of each group, were selected for experimental screening based on the nucleosome occupancy (45). The 28 synthetic sequences with higher similarity to the natural promoters concerning the predicted nucleosome average occupancy were selected for screening.

Before fusing these final 112 synthetic core promoters to the $P_{AOX1\text{-}R}$, we aimed to validate the core promoter structure of this promoter.

3.4.2. Assessing core promoter-CRM structure in the *P. pastoris* P_{AOX1} system

The natural (wild type) *P. pastoris* P_{AOX1} fused to an eGFP reporter gene, was used as positive control in this study. eGFP has widely been used as reporter for promoter characterization studies in *P. pastoris* (5, 8, 25, 26). All reporter protein fluorescence measurements of promoter variants are given relative to the wild type level normalized to 100% (shown in green – bar plots of Fig. 3.2 A and C-F).

A negative control variant was generated by deleting the *P. pastoris* P_{AOX1}-R (-769 to -172bp from start codon) to probe for its function. In a second negative control, the core promoter was deleted (-171 to -1bp from start codon). In both control variants, there was no detectable fluorescence thus the expression was completely disrupted (Fig. 3.2 A). This confirms that the core promoter sequence with high affinity to RNA polymerase II was completely removed in the variant without core promoter. Likewise, the variant in which the CRM was removed showed no fluorescence, confirming that all the relevant regulatory protein binding sites were removed resulting in complete functionality loss.

To ascertain the principle of modularity in this system, we characterized a variant in which the *AOX1* core promoter was replaced by the *HHF2* core promoter (another equally strong core promoter) (46, 47). The promoter activity level was identical to the natural P_{AOX1}, showing that different core promoters can be used interchangeably in this system (Fig. 3.2 A).

Given the complete loss of functionality when the core promoter or CRM are removed, as well as the modularity verified in this system, the determined core promoter-CRM boundary was maintained in all subsequent core promoter replacements. Namely, the core promoter boundary was set to 10bp upstream of the TATA box.

3.4.3. Establishing a baseline expression level

Seven control variants were generated in which the *P. pastoris AOX1* core promoter was replaced by completely random sequences (Fig. 3.2 A R1-R7). The resulting expression levels measured the basal expression of the $P_{AOX1\text{-}R}$ given that there is enough spacing between the CRM and the protein coding sequence for RNA polymerase II to bind. We performed this experiment to determine basic background transcription in our system. The average relative promoter activity of the seven control variants was 5.9% of the wild type promoter fluorescence (Fig. 3.2 A). We have used this value as threshold to evaluate whether the synthetic core promoters are significantly different from random sequences. In this way, synthetic core promoters with an expression value significantly lower than 5.9% were considered non-functional. For this purpose, we have adopted the One-way Analysis of Variance (ANOVA) statistical test.

3.4.4. Synthetic core promoters under the control of the *P. pastoris* $P_{AOX1\text{-}R}$

The aforementioned 112 synthetic constructs were assessed by replacing the native *P. pastoris AOX1* core promoter by each of the 112 synthetic sequences and measuring eGFP

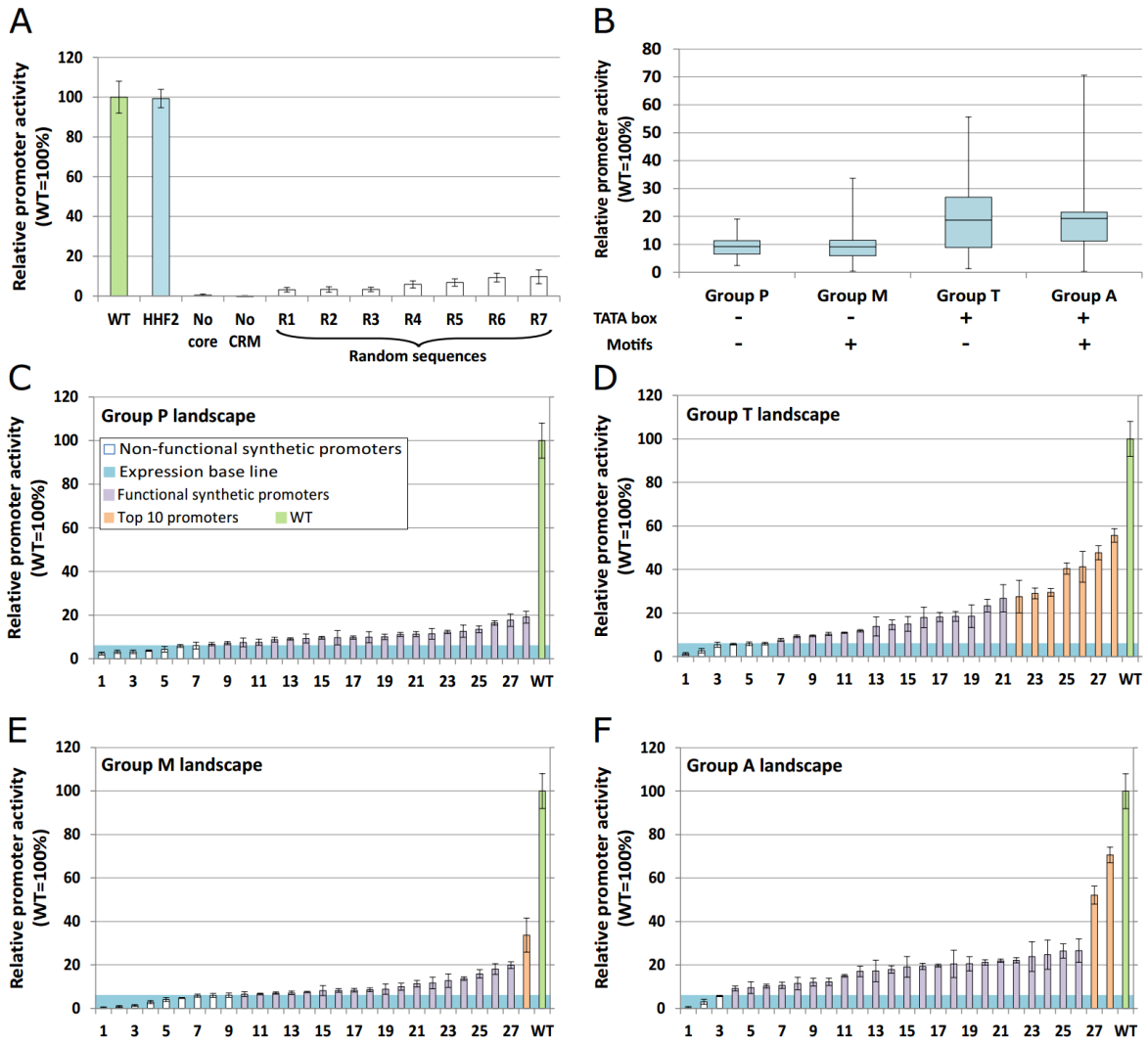

Fig. 3.2 – Establishing the $P_{AOX1\text{-}R}$ screening system (A) and testing the 112 synthetic core promoters (B-F). Promoter activity was measured by fluorescence intensity of the reporter protein after cultivation in 96-well deep-well plates and under methanol induction for 48h. **A –** Promoter activity mean and respective standard deviation of control constructs: *i*) wild type P_{AOX1} (green), *ii*) $P_{AOX1\text{-}R}$ fused to HHF2 core promoter, *iii*) $P_{AOX1\text{-}R}$ without core promoter, *iv*) *AOX1* core promoter without CRM and *v*) seven completely random sequences fused to $P_{AOX1\text{-}R}$. **B –** Overview of the groups of synthetic core promoters tested. Box plot of the minimum, first quartile, average, third quartile and maximum promoter activities for each of the four groups of synthetic core promoters (Groups P, M, T and A). **C-F –** Landscape of mean promoter activity, and respective standard deviation, for each of the four groups of synthetic core promoters (Group P, T, M and A, respectively). The individual synthetic core promoter activity is presented in increasing activity order. The legend of panel C applies as well to the other panels. Mean values and standard deviations shown in this figure were calculated from at least three independent cultivations in separate deep-well plates.

fluorescence. The overall promoter activity landscape is shown for each group (P, T, M and A) in Fig. 3.2 C-F, respectively. Seventy-eight percent of sequences showed a statistically significant (*p-value*<0.05) higher activity than baseline expression, thus they are considered as functional. Within the functional subset, reporter protein fluorescence levels range between 6.5% to 70.6%

with mean 17.0% and standard deviation 11.5%. Additionally, it was observed that the mean activity levels in groups T and A – 18.7% and 19.3%, respectively – are roughly two-fold higher than groups P and M – 9.2% and 9.1%, respectively. Furthermore, 16 out of the 25 non-functional core promoters do not have a TATA box. This is a strong indication that the TATA box is a key structural component in the P_{AOX1-R} system (Fig. 3.2 B).

Regarding the presence of motifs (group M and A), our data suggest that their presence does not significantly affect the expression level, given that the mean activity level is similar in groups with or without motifs (group P and T, respectively). However, we might speculate that the presence of motifs in association with other factors may explain the higher expression levels observed for promoters A28 and A27, given that both have motifs (Fig. 3.2 F).

Focusing the analysis on the ten promoters with highest activity (orange in Fig. 3.2 C-F), it strikes that the presence of a TATA box is a common feature, whereas the presence of motifs is not. The only exception might be the M28 promoter, which belongs to a TATA-less group. However, M28 has a TATA box like sequence in position -115 from the start codon.

3.4.5. Analysis of the top-ten synthetic core promoter sequences

The top ten synthetic core promoter sequences obtained in the screening with the P_{AOX1-R} (T22, T23, T24, T25, T26, T27, T28, M28, A27 and A28) were scrutinized in detail. They were examined by: *i*) BLAST analysis against the *P. pastoris* genome to search for similarities to naturally occurring sequences, *ii*) multiple sequence alignment to assess the presence of common motifs and *iii*) nucleosome occupancy analysis to evaluate its importance and common patterns.

To search for fragments of natural sequences, a standard nucleotide BLAST searching procedure against the whole *P. pastoris* CBS 7435 genome was adopted. With this method no significant matches were found. The detailed results are provided in Table 3.11. The highest *e-value* (0.083) was obtained for A28, T27, T26 and M28 sequences BLAST. The A28 and M28 matches were in protein coding regions and in an inter gene sequence in the case of T26. It is thus unlikely that this matches are characteristic regulatory sequences. In the case of T27, the match was in a possible promoter region in the *P. pastoris* genome (10bp upstream of nucleolar protein coding sequence). The match position in the synthetic core promoter sequence was however further upstream, close to the P_{AOX1-R} (-147 to -130bp).

To perform the multiple sequence alignments we used the EMBL-EBI Clustal Omega tool (48). The resulting alignment (Fig. 3.3 A) shows the conserved positions in seven or more sequences (shaded in blue in Fig. 3.3 A). Some of the marked positions are isolated, possibly caused by the higher adenine and thymine content, characteristic of strong core promoter sequences (28). In addition, three different common motifs (with more than one consecutive position conserved) were identified. The first one is located close to the TATA box region (position 40 in Fig. 3.3 A). However, two sequences had the respective TATA boxes positioned downstream from this region (T28 and T23), around position 70. This may influence the subsequent AT rich

motif (position 74). The last conserved region is a thymine rich sequence (position 146), followed by an adenine rich sequence (not marked), which may be related to the TSS, as suggested by Lubliner *et al.* (28).

Lastly, we calculated the nucleosome occupancy for the 10 best synthetic core promoters (Fig. 3.3 B) using the model developed by Kaplan *et al.* (45). Fig. 3.3 C shows the sum of nucleosome affinity for all the synthetic promoters. The data in Fig. 3.3 B unveil relatively low nucleosome occupancy in several synthetic core promoters (*e. g.,* T28, A27 and T26) but without a clear pattern. There are however a few exceptions (T27 and T25) with relatively high nucleosome occupancy. To ascertain a possible correlation between promoter expression and nucleosome affinity, we calculated cumulative nucleosome affinity for all the synthetic promoters and compared it with the respective expression levels. It revealed no statistically significant correlation, with a correlation coefficient of 0.07 (Fig. 3.3 C). This somewhat unexpected result might be explained by the diversity of synthetic sequences (discussed later).

The average position of the TATA box in the ten best promoters is position -120 (Fig. 3.3 B) with variations of 20 base pair around the mean. There are some promoters with lower activity with TATA boxes considerably downstream of this interval. Yet it is not possible to draw a direct causal relationship between TATA box position and promoter strength since many other features differ between them.

3.4.6. Second round screening: top-ten synthetic core promoters in different yeasts and CRMs

In the previous section, we validated the designed method and its capacity to create complete novel core promoters, demonstrating its functionality with the $P_{AOX1\text{-}R}$. Yet, we aimed to use synthetic core promoters as general tools for fine-tuning expression. Thus, they should be functional when fused to CRMs of any promoter. Hence, the top ten synthetic core promoters obtained from fusion with the $P_{AOX1\text{-}R}$ (Fig. 3.2 and summarized in Fig. 3.4 B) were fused to six different CRMs (Fig. 3.4 A), three from *P. pastoris* ($P_{DAS1\text{-}R}$, $P_{CAT1\text{-}R}$ *and* $P_{GAP\text{-}R}$ – Fig. 3.4 C to E, respectively), and the other three from *S. cerevisiae* ($P_{ScGAL1\text{-}R}$, $P_{ScGPD1\text{-}R}$ and $P_{ScADH1\text{-}R}$ – Fig. 3.4 F to H, respectively). These additional CRMs were chosen so that we could benchmark the synthetic core promoters in different conditions, *i. e.*, under the control of inducible ($P_{DAS1\text{-}R}$, $P_{CAT1\text{-}R}$ and $P_{ScGAL1\text{-}R}$) and constitutive CRMs ($P_{GAP\text{-}R}$, $P_{ScGPD1\text{-}R}$ *and* $P_{ScADH1\text{-}R}$). In all constructs, the synthetic core promoter was delimited to 10bp upstream of the TATA box. Therefore, the core promoters have a different length depending on the location of the TATA box and on the CRM (Fig. 3.4 A).

A key result is that the top-ten synthetic promoters show significantly higher expression when fused to CRMs of inducible promoters, irrespectively of the yeast and inducible mechanism. That is, the tested CRMs are inducible by methanol ($P_{CAT1\text{-}R}$, $P_{DAS1\text{-}R}$, and $P_{AOX1\text{-}R}$ CRMs in *P. pastoris*) and galactose ($P_{ScGAL1\text{-}R}$ CRM in *S. cerevisiae*). The minimum relative promoter activity was 38% for $P_{CAT1\text{-}R}$ and $P_{DAS1\text{-}R}$, 27% and 53% for $P_{AOX1\text{-}R}$ and $P_{ScGAL1\text{-}R}$, respectively. With all these CRMs, the strongest synthetic core promoter gave a higher relative expression than the $P_{AOX1\text{-}R}$, namely 82%, 122% and 176% for $P_{CAT1\text{-}R}$, $P_{DAS1\text{-}R}$ and $P_{ScGAL1\text{-}R}$, respectively, compared to

70% for the $P_{AOX1\text{-}R}$. Notably, $P_{DAS1\text{-}R}$ and $P_{ScGAL1\text{-}R}$ gave a higher expression value than the respective natural wild type core promoter. It should be stressed that these synthetic core promoters seem to be independent of the regulatory mechanism, since they are functional under the control of CRMs that respond to different *stimuli* (methanol and galactose) and in different yeasts.

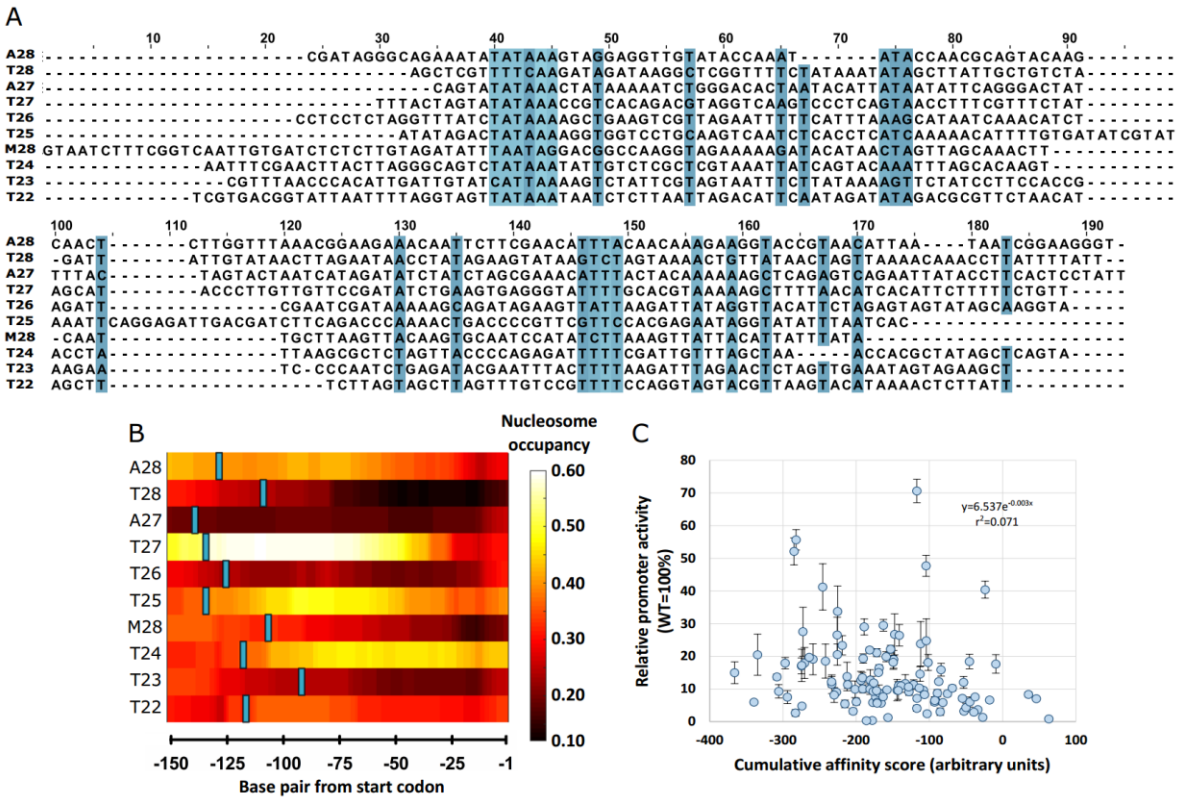

Fig. 3.3 – Analysis of the top ten synthetic core promoter sequences obtained from screenings with the $P_{AOX1\text{-}R}$. A – Multiple sequence alignment, using Clustal Omega (48), (ranked in increasing order of promoter activity). The positions conserved in seven or more sequences are marked with a blue shade. **B –** Nucleosome occupancy profile heatmap of the top ten synthetic core promoters when fused to $P_{AOX1\text{-}R}$. The nucleosome occupancy was calculated with Kaplan *et al.* prediction package (45). The core promoter is limited to 150bp form the protein start codon. The TATA box location is marked in blue. **C –** 112 synthetic core promoter activity mean and standard deviation as function of the respective cumulative nucleosome affinity scores calculated using the Xi *et al.* software package (65).

Fusions of the core promoters to CRMs of constitutive promoters show a limited functionality with the maximum relative promoter activity around 20% in $P_{GAP\text{-}R}$, $P_{ScADH1\text{-}R}$ and $P_{ScGPD1\text{-}R}$. All these CRMs have a TATA box in their natural sequence. In yeast there are mainly two types of promoters, TATA-positive and TATA-less promoters (31). Most of the available promoter studies were developed for the former group of promoters (31), thus we lack detailed understanding of critical sequence elements for transcription initiation in the TATA-less promoters. Hence, we have hypothesized that, although these promoters have a TATA box in their sequence, the transcription initiation might be TATA box independent. This would explain the apparent synthetic core promoter

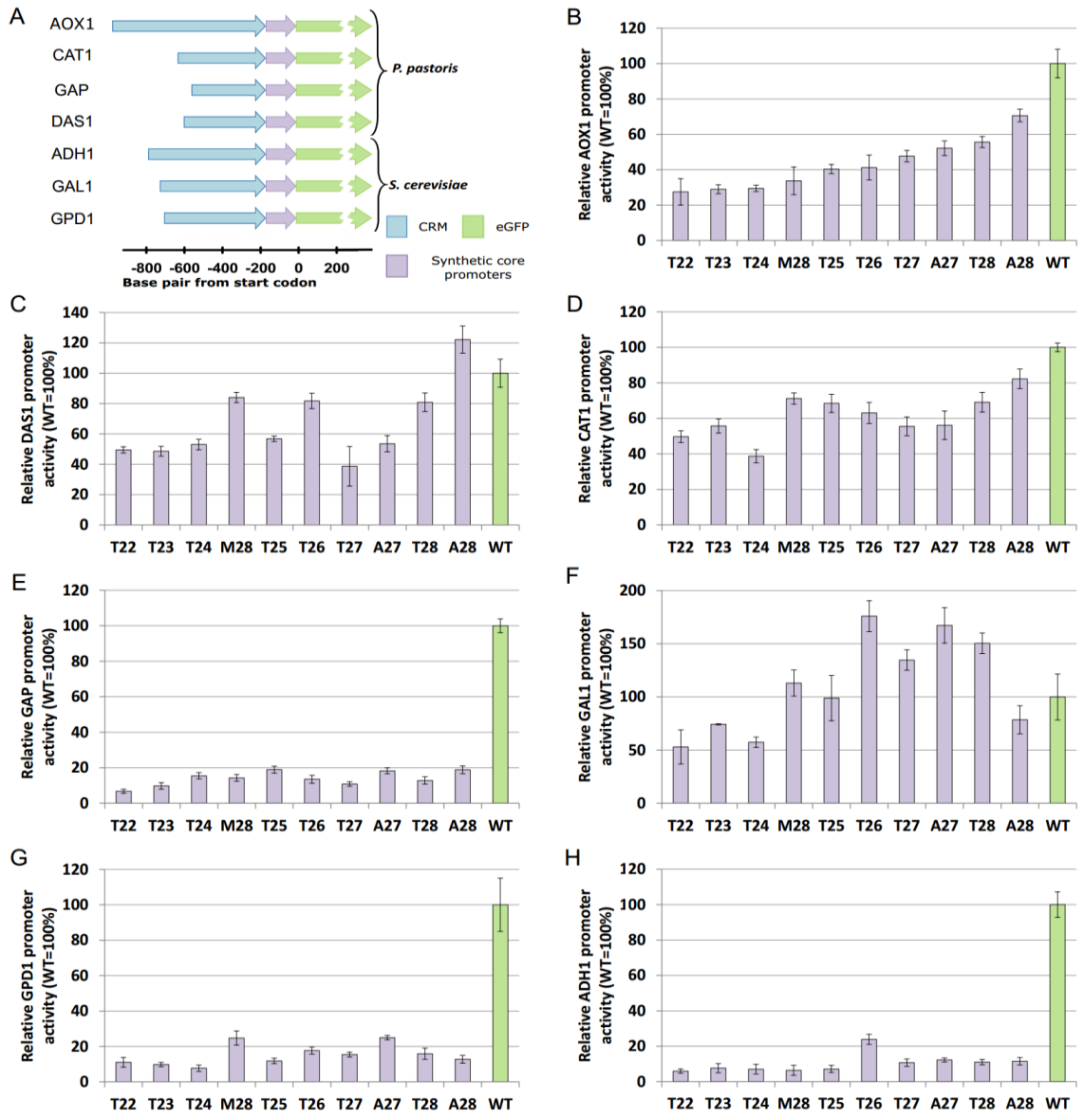

Fig. 3.4 – Testing modularity of the synthetic core promoters by fusing them to CRMs of different promoters in *P. pastoris* and *S. cerevisiae*. **A –** Relative size and CRM-core promoter fusion location of the different CRMs (P_{AOX1}-R, P_{CAT1}-R, P_{GAP}-R, P_{DAS1}-R, P_{ScADH1}-R, P_{ScGAL1}-R and P_{ScGPD1}-R, respectively) tested. **B-H –** Mean promoter activity (normalized reporter protein fluorescence), and respective standard deviation, of fusions of the ten synthetic core promoters with highest activity (when screened under the control of P_{AOX1}-R, Fig. 3.2) to different CRMs (blue bars). Synthetic core promoters were fused to different CRMs and tested in different yeasts: the P_{AOX1}-R, P_{DAS1}-R, P_{CAT1}-R and P_{GAP}-R were tested in *P. pastoris* (**B-E**), while P_{ScGAL1}-R, P_{ScGPD1}-R and P_{ScADH1}-R were tested in *S. cerevisiae* (**F-H**). The fluorescence measurements were performed after 48h induction with methanol for P_{AOX1}-R, P_{CAT1}-R and P_{DAS1}-R and 48h after inoculation for P_{GAP}-R, P_{ScGPD1}-R, P_{ScADH1}-R and P_{ScGAL1}-R promoters. All values represent single measurements of at least three independent cultivations in separate 96-well deep-well plates. In each case, the corresponding wild type promoter activity is represented in green. The order of the synthetic core promoters is kept the same (increasing promoter activity when fused to P_{AOX1}-R) to facilitate interpretation. The data of the core promoter fusions to P_{AOX1}-R is also shown in Fig. 3.2 (dispersed) and summarized here in panel B.

failure when fused to constitutive CRMs since the presence of a TATA box and adjacent nucleotides in the synthetic core promoters might favor a TATA box dependent transcription initiation mechanism. To test this hypothesis, we have disrupted the TATA box in the natural promoter sequence by mutating it. We have replaced three nucleotides of this motif by cytosine in the P_{AOX1} (control), P_{GAP}, P_{ScADH1} and P_{ScGPD1}. The resulting activity data showed that the expression is disrupted after the TATA box mutation in all promoters (18%, 20%, 8% and 2% of the wild type promoters for P_{AOX1}, P_{GAP}, P_{ScGPD1} and P_{ScADH1}, Fig. 3.9). Expression is therefore depending on the TATA box element in all cases. This finding does not confirm our hypothesis and suggests that others so far unknown elements might be essential for strong transcription for constitutive TATA box dependent yeast promoters.

3.4.7. Correlation between the activities of synthetic core promoters fused to different CRMs

We have evaluated context dependency and modularity of the top ten synthetic promoters by correlating the activity data of each synthetic core promoter under the control of different CRMs in different yeasts. This resulted in the correlation matrix depicted in Fig. 3.5 A (heatmap showing all possible combinations of CRMs and yeasts, Fig. 3.10). It was observed that the highest correlation coefficients are obtained within the subset of inducible CRMs in *P. pastoris* – $P_{CAT1\text{-}R}$, $P_{DAS1\text{-}R}$ and $P_{AOX1\text{-}R}$ (*e. g.* Fig. 3.5 B), with correlation coefficients ranging between 0.40 and 0.63. Also, relatively high correlation coefficients (around 0.5) were found when comparing the $P_{ScGAL1\text{-}R}$ and the constitutive CRMs in *S. cerevisiae* (*e. g.,* Fig. 3.5 C). However, a low correlation was observed between the synthetic promoters controlled by the $P_{ScADH1\text{-}R}$ and $P_{ScGPD1\text{-}R.}$ This might be explained by the much lower expression levels in these particular experiments. It should also be pointed that low correlations are observed when comparing CRMs of *P. pastoris* against CRMs of *S. cerevisiae* (*e. g.,* Fig. 3.5 D). Finally, it should be noted that, even when correlation is high, the relative expression levels of the same synthetic core promoter under the control of two different CRMs varies significantly. This means that although functional and correlated, the synthetic core promoters are not completely independent of the CRM to which they are fused.

3.5. Discussion

3.5.1. Functionality of synthetic core promoters

In this study, we have followed a *de novo* design approach to generate synthetic core promoter sequences for yeast cells. The design was based on natural *S. cerevisiae* core promoters resulting in synthetic core promoters that were at first experimentally tested in *P. pastoris*. We have chosen this approach, because we were primarily interested in developing regulatory elements for *P. pastoris*, where generally applicable promoter engineering strategies are scarce (36). In contrast to *S. cerevisiae*, where large sets of experimental data on core promoters from large scale high throughput studies are available, no such studies have been performed in the widely used protein production host *P. pastoris*. Hence, due to the reported conservation of core promoters (27) and

previous studies which demonstrated functionality of *S. cerevisiae* core promoters in *P. pastoris* (8), we used the data set from *S. cerevisiae* as starting point.

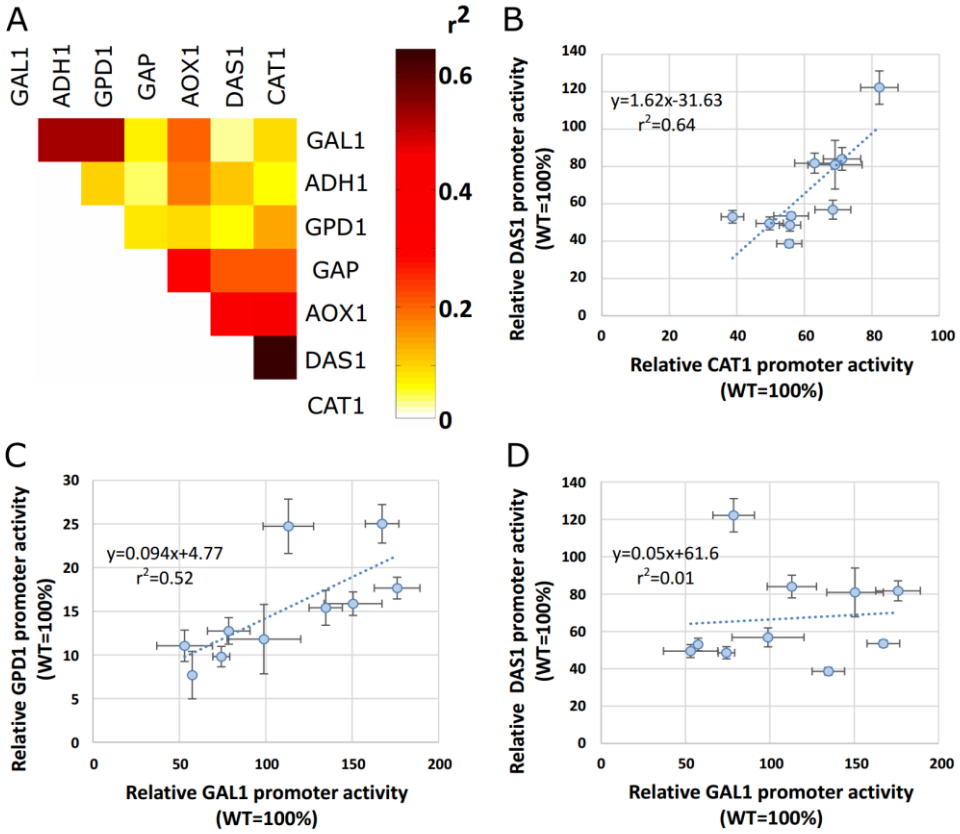

Fig. 3.5 – Correlation analysis of the top ten synthetic core promoter activities fused to seven different CRMs. A – Heatmap of the correlation coefficients of the top-ten synthetic core promoter activities fused to different CRMs. All the possible combinations of CRMs are shown. The correlation coefficients were calculated based on the average of single measurements of at least three independent cultivations in separate 96-well deep-well plates. The samples were taken 48h after induction ($P_{AOX1\text{-}R}$, $P_{CAT1\text{-}R}$ and $P_{DAS1\text{-}R}$, cases) or inoculation ($P_{GAP\text{-}R}$, $P_{ScGPD1\text{-}R}$, $P_{ScADH1\text{-}R}$ and $P_{ScGAL1\text{-}R}$ CRMs cases). $P_{AOX1\text{-}R}$, $P_{CAT1\text{-}R}$, $P_{DAS1\text{-}R}$ and $P_{GAP\text{-}R}$ were tested in *P. pastoris*, while the remaining CRMs were tested in *S. cerevisiae*. Panels B-D show representative data on the correlation coefficients: **B –** Synthetic core promoter activities when fused to $P_{DAS1\text{-}R}$ as function of its respective activities when fused to $P_{CAT1\text{-}R}$. **C –** Synthetic core promoter activities when fused to $P_{ScGPD1\text{-}R}$ as function of its respective activities when fused to $P_{ScGAL1\text{-}R}$. **D –** Synthetic core promoter activities when fused to $P_{DAS1\text{-}R}$ as function of its respective activities when fused to $P_{ScGAL1\text{-}R}$. The correlation diagrams between other promoters' combinations are shown in Fig. 3.10**.**

This design method delivered 77,6% of functional core promoter sequences with the *P. pastoris* $P_{AOX1\text{-}R}$ (Fig. 3.2). These sequences are markedly different from naturally occurring sequences (no clear matches to natural promoters were found by BLAST search, Table 3.11), between each other (Fig. 3.3 A) and substantially more diverse than variants typically obtained by local random mutations of a natural core promoter (7, 16, 43). This lack of resemblance to natural sequences is an important feature of this set of promoters. It may increase the genetic stability in the genomic context, as these sequences have low probability of recombining with any natural

sequence in the genome. This feature will be valuable for future *in vivo* and *in vitro* pathway assembly (49), when combining a multi-gene pathway using a different core promoter for each enzyme, with the objective of fine-tuning the production of each one while using a single inductor.

In a recent study, eleven artificial core promoter sequences were assessed in *P. pastoris* (25). Of these, only two were generated *de novo* by consensus sequence analysis of the natural core promoters of the *AOX1*, *GAP*, *HIS4* and *ADH2* genes. The other nine sequences were generated by replacements of short stretches in the natural *AOX1* core promoter. For the two consensus derived sequences, the activity levels were within the range of the basal activity level obtained from randomized sequences in this study (Fig. 3.2 A). This suggests that the previous design considerations had a non-significant effect over using random sequences. The replacement method was more successful, with activity levels as high as 117% of the natural P_{AOX1}. However, with the replacement method the resulting sequences share a high degree of similarity with the natural sequence, thus questioning the ability of the method to generate truly synthetic sequences. As discussed by Dehli *et al.* a diversity inherent component design approach (as the one adopted here), is advantageous for Synthetic Biology problems, as it facilitates orthogonality, modularity and standardization of new components (50).

We obtained an average activity level of 17% with a dispersion of 11.5% and a maximum activity of 70% of the wild type P_{AOX1}. Overall, this reflects the ability of the design method to span a wide spectrum of highly diverse synthetic sequences. However, the relatively low average activity might be in part explained by the way the experimental input data from *S. cerevisiae* was obtained (28). Lubliner *et al*. deduced core promoter functionality from reporter protein fluorescence measurements of the entire promoter (including the respective CRMs), whereas we fused all core promoters to the same CRM. Hence, expression strength of the *S. cerevisiae* measurements may also be influenced by the CRM and not solely the core promoter. Additionally, the phylogenetic distance between *S. cerevisiae* and *P. pastoris* may have complicated our efforts. It has been shown that core promoters in distant related yeasts maintain their functionality but with lower expression (27). To further support this statement, it should be underlined that the highest relative expression levels (176%) were obtained for $P_{ScGAL1\text{-}R}$ in *S. cerevisiae*.

Another characteristic that could compromise the synthetic core promoters' strength is the boundary between the core promoter and the CRM. Here we maintained the same boundary condition in all experiments (-10bp from the TATA box), however, it might have some influence in promoters' strength and might be the target for future expression fine-tuning and promoter optimization studies.

3.5.2. No motifs except the TATA box clearly affect expression

Fig. 3.2 B shows an expression box plot for the four groups of sequences. The comparison between groups P and M and groups T and A show that the introduction of motifs does not affect the mean expression level, but might have an effect in specific cases (*e. g.,* A28 and A27). Indeed, the effect of motifs in core promoter strength is not consensual in the literature. Recently, Seizl *et*

al. (51) suggested that the GAAAA 5-mer is a conserved yeast promoter element, functioning as a TATA binding protein binding site in promoters lacking a consensus TATA box element. However, Lubliner *et al.* (52) studied knockout mutations of 122 GAAAA 5-mers. These modifications showed little to no effect on protein expression. Other studies have concluded that, with the possible exception of the TATA box (when present), motifs are not determinant for *S. cerevisiae* core promoter functionality (32, 53).

The comparison of groups P and M with groups T and A reveals that the presence of a TATA box motif is a key effector of high expression levels, which corroborates the data presented in previous studies (32, 53). Indeed, within the top ten promoters only one sequence (M28) does not have a TATA box. However, this apparent exception is discarded after a careful sequence analysis. Such analysis reveals that the M28 has a TATA box like sequence, namely TATTTAATA at position -115. Several previous studies have shown that mutations in the TATA box region greatly affect promoter strength (54, 55). In another study, Mogno *et al.* (56) analyzed libraries of TATA-positive and TATA-less promoters in *S. cerevisiae* showing that the TATA box mainly affects the transcription rate by enhancing it. It was also shown that the location, orientation and flanking bases critically affect TATA box function and core promoter activity (52). However, given the size of our data set (56 synthetic core promoters with TATA box), we cannot draw solid conclusion regarding these aspects.

3.5.3. The role of nucleosome occupancy

Nucleosome occupancy has been reported as having a fundamental role in transcription initiation (18, 57). Variations in nucleosome occupancy alone may cause large differences in promoter strength. Raveh-Sadka *et al.* (57) showed that AT rich sequences are associated with low nucleosome affinity and high promoter activity. Curran *et al.* (18) have redesigned nucleosome architecture in natural *S. cerevisiae* promoters with a 1.5- to 6-fold expression increase of a reporter protein (β-GAL). They have hypothesized that nucleosome occupancy is an important causative factor limiting the strength of native promoters and is likely an evolutionary mechanism for controlling transcriptional strength (18). In our study, we observe no statistically meaningful correlation between promoter strength and nucleosome occupancy (Fig. 3.3 C). This suggests that other factors might have an even higher effect than nucleosome occupancy, which was the main design factor studied by Curran *et al.* (18). Similar results were obtained by Lam *et al.* (58), who have shown that the interplay of nucleosomes and motifs is important to explain promoter activity variations in *S. cerevisiae*. Experimental data for *P. pastoris* nucleosome occupancy are still not available and might help to explain our observations in future.

3.5.4. Modularity of synthetic core promoters

To assess modularity, we have inserted the top ten synthetic core promoters in *P. pastoris* and in *S. cerevisiae* under the control of seven different CRMs, four of which are inducible ($P_{CAT1\text{-}R}$, $P_{DAS1\text{-}R}$, $P_{ScGAL1\text{-}R}$ and $P_{AOX1\text{-}R}$) and three constitutive ($P_{GAP\text{-}R}$, $P_{ScADH1\text{-}R}$ and $P_{ScGPD1\text{-}R}$) (Fig. 3.4, Fig. 3.5 and Fig. 3.10).

The expression levels of the constitutive promoters are consistently lower than the inducible promoters (Fig. 3.4). Although the compatibility between CRMs and core promoters has previously been proven, even between different organisms (8, 27), it appears that, according to our data, it is not universal. For instance, in *S. cerevisiae* the *RPS5* CRM is compatible with *ADH1* and *CUP1* core promoters, thus being able to initiate transcription. This is however not reciprocal, *i. e.*, the *ADH1* and *CUP1* CRMs cannot initiate transcription when coupled with the *RPS5* core promoter (59). We have hypothesized that the tested constitutive promoters have a TATA box independent transcription initiation, hence being incompatible with this set of TATA box containing synthetic core promoters. We tested this hypothesis by mutating the TATA box in the respective natural promoter sequences. The results show that the expression is disrupted (Fig. 3.9), indicating that the transcription initiation of all constitutive promoters in this study is TATA box dependent. Hence, the lower expression of synthetic promoters fused to constitutive CRMs must rather be attributed to unknown regulatory mechanisms specific for constitutive promoters.

Within the group of inducible promoters, expression levels are high, irrespective of the yeast and CRM specific regulatory mechanism. Although the different CRMs in different yeasts respond to different *stimuli* (namely, methanol and galactose) it had no effect on its functionality. Some CRMs even outperformed the activity levels of the wild type promoter, namely $P_{DAS1\text{-}R}$ and $P_{ScGAL1\text{-}R}$ in *P. pastoris* and *S. cerevisiae*, respectively. *S. cerevisiae* $P_{ScGAL1\text{-}R}$-T26 showed the highest relative activity level (176% of the wild type P_{ScGAL1}). This may reflect the fact that our design was based on *S. cerevisiae* core promoters.

The correlation analysis of synthetic core promoters' expression levels under the control of different CRMs (Fig. 3.5) shows that the correlations are higher when comparing CRMs in the same organism. This is the case of $P_{CAT1\text{-}R}$ against $P_{DAS1\text{-}R}$ in *P. pastoris* and $P_{ScGAL1\text{-}R}$ against $P_{GPD1\text{-}R}$ in *S. cerevisiae*. Correlations are in general very low (lower than 0.2) when comparing CRMs of different organisms. For instance, in the case of $P_{DAS1\text{-}R}$ and $P_{ScGAL1\text{-}R}$ in *P. pastoris* and *S. cerevisiae*, respectively. These data suggest that comparable expression strength irrespective of the context (*i. e.*, modularity) is maintained only within the same organism, although the core promoters' functionality is maintained in the different organisms tested. Zeevi *et al.* described the conservation of orthologous ribosomal promoter activity within closely related genus of yeasts (27). For instance, *Sacaromices paradoxus*, showed high correlation with *S. cerevisiae* while *Kluyveromyces lactis* diverged considerably. Likewise, we can anticipate that the low correlation observed in our study is due to the phylogenetic distance between *P. pastoris* and *S. cerevisiae* genus (27).

All in all, our work demonstrated the feasibility of a multi factor rational synthetic core promoter design and its applicability as general engineering tool for gene expression fine-tuning. Due to their sequence diversity and independence of natural sequences, similarly designed synthetic core promoters may become valuable tools for Synthetic Biology and metabolic engineering applications in other eukaryotic organisms.

3.6. Supplementary information

3.6.1. S1: Summary of literature references on *S. cerevisiae* and *P. pastoris* core promoters.

In this section we will succinctly describe the main studies developed to clarify the mechanisms of yeast promoters, focusing on *S. cerevisiae* and *P. pastoris*. The references follow the numbering of the main text.

S. cerevisiae: Sugihara *et al*. identified CRMs in TATA-less and TFIID-dependent core promoters, namely *RPS5* (32). Park *et al.* developed a method for improved TSS mapping, inferring relationships between core promoter, CRMs, chromatin features and TSS location (33). A study on the effect of 5'UTR features on gene expression was presented by Dvir *et al.* (34). It has been shown that yeast core promoters show a high level of conservation, maintaining their functionality even in distantly related species (27). As illustrative example, the *S. cerevisiae LEU2* core promoter has been shown to remain functional when tested in *P. pastoris* (8).

Different approaches were followed to design synthetic promoters for protein expression fine-tuning. Some are focused on CRMs (37, 38), while others target both CRMs and core promoters (7, 9, 18, 19, 39–42). For CRMs design, the interaction between Transcription Factors (TFs) and respective TFBSs have been used to model transcription, either by creating a large library of promoters based on combinatorial arrangement of different TFBS upstream of the natural core promoter (38) or by generating orthogonal synthetic zinc fingers used to wire new synthetic transcriptional cascades (37). Also, the inclusion of regulatory sequences next to the core promoter can be used to fine-tune transcription. Other design approaches consisted in adding random mutations to a natural promoter (*TEF*) (7, 9), randomizing two specific areas of the *PFY1* promoter and changing its expression profile afterwards by adding Tn10 Tet operator sites (41). Additionally, the generation of a minimal promoter based on a large scale screening of random sequences for minimal length, robustness and modulatory (19) has been used for the same purpose.

P. pastoris: Promoter libraries have been generated mostly based on deletions of CRMs sections to research the P_{AOX1} regulatory mechanisms and to determine TFBSs (*e. g.* (24, 26)). Promoter libraries have also been created to control gene expression by modifying the core promoter (16, 25, 43) and 5'UTR (44) sequeces. Berg *et al.* studied random mutagenesis of P_{AOX1} (16), located in both core promoter and CRM regions. They observed expression profile modifications (derepression) when mutating some specific nucleotides in the P_{AOX1} CRM, and modifications in the expression rate when mutating the core promoter sequence. Following a different approach, *P. pastoris* synthetic core promoters have also been designed based on four natural *P. pastoris* core promoters consensus sequence through the addition of some natural TFBSs (25).

Table 3.1 – List of primers used to clone the positive and negative controls.

Name	Sequence
C-WO-CRM1	TATTGTGAAATAGACGCAGATCGGGAACACTGAAAAATACACAGTTATTATTCATTTAAATGACAGCAATATATAAACAGA AGGAAGCTGCCCTGTCTTA
eGFP-p*AOX1*-3prime	AAAAGTTCTTCTCCTTTGCTAGCCATCGTTTCGAATAATTAGTTGTTTTTTGATCTTCTC
p*AOX1*_Syn_dBamHI_SwaI-forward	GATCGGGAACACTGAAAAATACACAGTTATTATTCATTTAAATAGATCTAACATCCAAAGACGAAAGGTTGAATGAAAC
C-WO-Core1	GTTCTTCTCCTTTGCTAGCCATAAGTAGGGGTTAGAACAGTTAAATTTTGATCATG
C-W-HHF2+10	GTTCTTCTCCTTTGCTAGCCATATTTATTGATTATTTGTTTATGGGTGAGTCTAGAAAAGGACGCACTCGTCTTGTATTTA TAGATGAAAAGAAAGTAGGGGTTAGAACAGTTAAATTTTGATC
R1	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGAGTCTGGGGCGGTCTTATGCTAATCGTTGGTGTCCTGTCCTAGAGAGTT GCGAACAGGCAAGGCGCCGGAGATAGAGTATCGTGAAGGAATTTATAGAGGGTCTATCGCCGCAGTCGTCGACGGGC AAACAAGCACCGGGCAAGTAGGGGTTAGAACAGTTAAATTTTG
R2	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAACACGTATGGGCCAAGTGTTTGAAATTCAACCTAGTTTTGAGGGTTACGA TACGGACCTCCCCTACGCTGCAGTCCTATCAGCGAGGCGCAGCAAGCGTTAGGCGTGCGGCTTCCAGACTCAAGGAG ACCTTGGCGATATGAAGTAGGGGTTAGAACAGTTAAATTTTG
R3	GTTCTTCTCCTTTGCTAGCCATCGTTTCGACGCTTTCGAAGAGGATGGCTACTCCCGGCTAGTAAGTTGGCTTGTTTGCT TGACGGATGACTTAGGGTTTTTAATGAAGGCCGTTGTGTCTACGGACACGACACGGGTGTTCTACTCGCGCCTCTTGGG AGCACTTTAATAAAGTAGGGGTTAGAACAGTTAAATTTTG
R4	GTTCTTCTCCTTTGCTAGCCATCGTTTCGACGTTAGAGATTGGCTTCCTAAGATCGAGTAGTGATTCCGTATATAGGCTC GCTTACCCAGATCCCCTAGACTGTTCGCTTTGCTATGAGTCTCAACTAACCCTAAATGTCCGTGCCCCGTTTCTATCGTA TATTAGGGGACAAGTAGGGGTTAGAACAGTTAAATTTTG
R5	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAAATGTAGCTTTGTCCGTAGAATACCCGGTATAAGACAAGAGCAGTCTAGT AGGAGAGCTCCTTTGACCTGCCGGTTTCTGGGAAGGGCCCAGGAAACGGGTTACGGTTCTTACGACACGAATGCGTGT TTCGTGACTGTTTAAGTAGGGGTTAGAACAGTTAAATTTTG
R6	GTTCTTCTCCTTTGCTAGCCATCGTTTCGATCAACCGGCAATTCAGAAATGCGGGATTTAGTACACCTATTTCTTTACCAA CTCCCGCCGCGTTTTAGCTTAATGATGAGCGGTGGGCGTGTTTTGAAAAAAGATGTTAGAGATATATTCTAGTCAGAGG GTTCGACTAGAAAGTAGGGGTTAGAACAGTTAAATTTTG
R7	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTGCTTACCGGTGGATTACGTACGGGGTTGGATGAGAAGTGGGATGGCCCA ACACCAATTGATTGTCAACTCCGACCTGAAGGCTTCACAAGAGTGGAGGGTACAGCTAGTACTGAGTGTGTTAATTGGA GTACGGTCGGCTTAAGTAGGGGTTAGAACAGTTAAATTTTG

Table 3.2 – List of primers used to clone the synthetic promoters of group P.

Name	Sequence
P1	GTTCTTCTCCTTTGCTAGCCATCGTTTCGCGAGCACTTCTAGTTTTCGAAGATACAGTTCCACAATAGGTTCTTTCTATCGTCAGTATTCTCGTTC GCAAGCAAATAATTTCGGGATTATAGCGGAGTTTACAAACTAAACAAGATGTGCCTACTAGGCTATTCTAGACTAAAAGTAGGGGTTAGAACAG TTAAATTTTG
P2	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTCGTGAAATTACTAAGGGTTTAACTGCACTCGCGTATTTCAGAAGACTACTCTCGTAGGTTAATGC AACTCTACAAGTGATGACTTTGCTATGAACTACTTGTACTACTTACAATCTGTCTAGCAAACTGCGTCCGAAGTCCTAAGTAGGGGTTAGAACAG TTAAATTTTG
P3	GTTCTTCTCCTTTGCTAGCCATCGTTTCGATAATTTCAACCCTTCGTTCCGAAAACTAGAACGTGAGTTTCTGTTTGGCTATTCCGCGTATATCG TTTCTGAATTATTGAATAGAGCACAATATCAAAATACTAAAAATCGAGTTATTGGGATCGTACCAATACGTGGTTTAAAGTAGGGGTTAGAACAG TTAAATTTTG
P4	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGTTTTCGCTTTAATCTGTTAGGCAATATCCCTATTACGCAAACAGAGACTAACACCAACGACCAAG ACTTATATTTTGTTTGCGCACTACTAGCTAGGAACGTAGATATCAGTTACAAATATAATTCACTACGAGTTATCCGGAAGTAGGGGTTAGAACAG TTAAATTTTG
P5	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAAGACTCAGTTACGACGTAGAGAGGATGGATAGGCCGTTCGTCAACAGAGCGATCTAATTGTTTC GTTACTGATGGAATTGTTGGATAGTGTAAATCTAATAATGGAATTATATTTCAATTACTTGTTTGGGTAAAACGCCTGAAGTAGGGGTTAGAACA GTTAAATTTTG
P6	GTTCTTCTCCTTTGCTAGCCATCGTTTCGCTAGAACTAGGTTCTTCTCCTACTGTCTAAATATCTCTATATTTTAAAGTGAAATTTGGAGTGGTCG TTATAATTACTCGTTTTAGTGCACCCTAGTCGGGGGTCTAAAATTACAGTATACAAGTAAAGTTGTGATGACTCCAAAGTAGGGGTTAGAACAGT TAAATTTTG
P7	GTTCTTCTCCTTTGCTAGCCATCGTTTCGCTAAACTGTTCCTATCTTCTTATGTTTAGTAATGTAGCGTGAAATAATGTCAGACGATTATCTACTA CGAGACTACACTACGATACGTACTAACGAGGAGTGACTTGGGGGTACCGTATAGTTGTAATCTACCTACTTCCGCCAAGTAGGGGTTAGAACA GTTAAATTTTG
P8	GTTCTTCTCCTTTGCTAGCCATCGTTTCGATATGATCGTATGGGCAAACTTCACTCTGTTTCTAATTATAATTTAGCTTCGGATCGTATGAGGGT GGACACCTCGGTTGACTTGACTACGGTTCTAATGAACTTTTAATAAATCGTACCCACCTAATTAGAGAAGTATATAGAAGTAGGGGTTAGAACA GTTAAATTTTG

Table 3.2 (cont.) – List of primers used to clone the synthetic promoters of group P.

Name	Sequence
P9	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGATTCTAATGATATCTTCCACGACTGTAGAGCAACGGTTAGCAAACTACTATGTAGATGTTTTAGA TTGTGATTTAGATGCAAACTATGTTCCTTATTTTAACAACAATAGTGCAACTATATTGGAACTACCTGCAGAAAGCAAAGTAGGGGTTAGAACAG TTAAATTTTG
P10	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAAAGAACACGACGTCTATGAACTAAGGTTCAGTCTAAATACTAAATAAATGAACTTGTATCTATTTT TTTGCCTGATATAAGTTGCGTTGGGAAGACTAATTATGAAGATGTTCAAGATAAGATGAATTGAATAACTAAAAAAAAGTAGGGGTTAGAACAGT TAAATTTTG
P11	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTTCTGTACTTAACAGTAGCACTACAATTTCTTAAGTACTAATTTACTTTTATTCCTACTACTAAGTGG TTTTCAGCTATGGTGGTTCAAATATTAGGTAGCCTAGTATCCCACGTACGTAATGAGACAAAAACTAATAATGCAAAGTAGGGGTTAGAACAGTT AAATTTTG
P12	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAGTTTTATTAAAGAATTTAGCTCTAGCTACAATCGATTTATGCAATCGTGCTAGACTGGAATAACTT GTAGCTACGCGTATGGCTTCGTATTGGGGAGAAGTAGTTAACACACGACTATGGATATTATGGTAAATAGTCAATAAAGTAGGGGTTAGAACAG TTAAATTTTG
P13	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTTAATCACGGGTGAAAATTAAGAGAACTTTTAACTTAATGAGACTAGGGGAATAAACTTTGAATTG GTTCTCGTACGTATGCGGTAACTCGTGTATTTGCCCTATGAGTAATAGGTAGAATCAAGAATGTACTACTAATATGGAAGTAGGGGTTAGAACA GTTAAATTTTG
P14	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTATAACTTTAATTTCTTTCAATTTTAAAAAATATAAAACGGTAACTAAAGGTATTTTTCGCGTTAACC AATAACTGATTTTAAGTATATCTGCAGAGTAAGGGTTGATGAAGCAGGTAGCTATTTGAGTAGAATCGTACAATGAAGTAGGGGTTAGAACAGT TAAATTTTG
P15	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGTATTATCTATAACCACTATACTTTTCTATTACTAATTTTCAAGCAATACGTTCGTGTTTGCGTAAGT ATACCCTGGTATCTGTGACTCAAGCTAAGTAACGAGATTATTGACCTACCTTTGGGTGTATCAAGTCTAACAAAGAAGTAGGGGTTAGAACAGT TAAATTTTG
P16	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGTATTAAGAATAAAGTTCTCAGAAGGGTTTAAATGCGAGCTTAATTTGGGATGCTTAGATGTTATC CTTATCTAAATTCACTAATACACAGATAGTTCAGTTAATGAGCAGAATTTTAGTGTGACAGAATCTGTGATGTGTCAAAGTAGGGGTTAGAACAG TTAAATTTTG

Table 3.2 (cont.) – List of primers used to clone the synthetic promoters of group P.

Name	Sequence
P17	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAACACTCTCTAATGATACGCGAGCTAATTAGAAGAGAAGCTAGCTTATGAAGTTTATCGGTTGCTC CCACTATCAAACTATAAATAAGTGGAAAAAACTTCTCGTGTCGTTGTTGAACAACTAAAGTCCTATTATCGTGTCCAAAGTAGGGGTTAGAACAG TTAAATTTTG
P18	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAATAGATACAAAATCGTTAACTATAGGTAGTGCTAGTACTTTCAGATACCTTTTTGGTAGGCTAATT TATCTATCAATATATTAAAGTACGGTCTCCCTCGTTGAATGATAACTCAGTACCTAACTAACACTTTAATTAAGGAAAGTAGGGGTTAGAACAGTT AAATTTTG
P19	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAAGCTCGGTAATTCTAGAGTTTTAATTCGTTCTGTTAATTACAATGAAATCTACGTTCTAAATTTTTT CAATAAGTTTCTAATCAATCACGGGCAATTACAAGGATTGAAGTAGTCTACCTTTGTGTTCTAGTGTGGACAGTCAAGTAGGGGTTAGAACAGTT AAATTTTG
P20	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAAATCGACTTCTATTACTAATAAAGTCTAAGACTACGTTATTCAGATGCTACCTTTGAGAGTTTATA TCTACTACTATAACTGTAGTCACACGAACCTAGAATTCAGTTTCCACGGTTTATAATTACTCTATCTAAAGGGATTAAGTAGGGGTTAGAACAGT TAAATTTTG
P21	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTTCAAGGGTCACCTACAATAAGTTCTTTAGAGTATGTATATTAAACTCTATGTTGCAAGATATGGGT AACTATCAGATCGGTAAATCGTCGGTTCTAATACTATGAACTAAAAAGGTCTAACCTATGATCCTATATTCTCCTAAAGTAGGGGTTAGAACAGT TAAATTTTG
P22	GTTCTTCTCCTTTGCTAGCCATCGTTTCGCTTAATCTCGATAATATACTTTAATTACTAAAGGAGAACGATTCACAAACTAAGCAACTACGATATT GTCAACTATGCAAATTTTGATGGGTAGAGTAGTCTAGATTGTTATATTCAACTGCGAAAAAGATAGCGGAAACTTGAAGTAGGGGTTAGAACAG TTAAATTTTG
P23	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTCAGTTCTTCTAGAGGTATCTGGTTTTCTAGATTATTCGTTTTTTATTCTACGGAATTTGGAATTTG GTGTTTGAAGGTGTTACCTCTGCGACAGATTTTGTTATTCTAAAATAGTACCTCCCGAGTGAGCAAACAATTGACGAAGTAGGGGTTAGAACAG TTAAATTTTG
P24	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTCAAAATTCAACTAATTACAATTTAAGATATTAAAAAATCTTTAGCGCTATGGGGTGAGCTAGTTAA TGGCAAGACAAGATAGCTAAATTTAAGTTCCGTATATGTCTACTAGTGTTTCGTGGACTCAAATTAAGAAAGATGAAAGTAGGGGTTAGAACAGT TAAATTTTG

Table 3.2 (cont.) – List of primers used to clone the synthetic promoters of group P.

Name	Sequence
P25	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAATTAATTATTTACTTCTAGTCAGCACTAATGAGGTTTCAGCTTTTATTGAACTTCTGTCAGATTAGT ATACTAGCTACAACTATAAATCTGCGAAGCTAAGTTTAGGAAATAATATCTGTATTTATGCTCCCGATTCGAGTCAAGTAGGGGTTAGAACAGTT AAATTTTG
P26	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTGATTTATGCGTTATGGATATGTTAAGGCAACGATAATTATTAAAGGAATAGTCGTTAGAGCCGTG TGAATTCTACTACGTATTATTCTAAAGAGTCACTACTGATGTCCTTATCTATTAGTGATATATTTCGACCAGACGAAAAGTAGGGGTTAGAACAGT TAAATTTTG
P27	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGTCTAGCTAAACCTTCTACAATGTGAATTATTCAAACGTAGATCGTAGGATTCTAAGGTTCGTGGA CAGTAGTTGTTTATAGGGGGGCTCTAGAGAGTTTGATTAGCGATATTTAGAGACCAAATTTCACCTGATAGCCTAGCAAGTAGGGGTTAGAACA GTTAAATTTTG
P28	GTTCTTCTCCTTTGCTAGCCATCGTTTCGACTAAGTTCACACGGGCTAATATTTAAGATAAACTATAGGATGCTAATATTTACGTATAATGGAAG GTCCTGTAATATCCTCCAGATGGATTTGTTAACTATTTTATACTGATCGTAAAGTGATTAATGTTGAATCCCCTAGCAAGTAGGGGTTAGAACAG TTAAATTTTG

Table 3.3 – List of primers used to clone the synthetic promoters of group M.

Name	Sequence
M1	GTTCTTCTCCTTTGCTAGCCATCGTTTCGATCTAATATTCAGAACTATATCGTTTAGCGGGATGGGCAAGTGCCGCCCCTATTTTTAAAATGAAT AACTACTAGATTTCACACGGGTTTGTTTTGTTGATATGTTATTACAATCTAGCTCAATGATTATACTTGATCTCTTTAAGTAGGGGTTAGAACAGT TAAATTTTG
M2	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGTAATATTCAAGGTGGGTTGATGCACCCTAAGGTGGCTATTCTTTTTGCCTGGTGTTTACTGCTAT TTCTGTGGCACTTTAGAACAGTTCCTTCGACAGTCTTGTTTTCCGCTAAGGAGAGTAATGGACGGATTACCGAAGCGAAGTAGGGGTTAGAACA GTTAAATTTTG
M3	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGACTCACTTGGAGCTTGTTAAGTAATCAGGGCGTCGGTTGTTGAAAGTTTGATTGGTTACGGTTCT TTCGATTTTCGGCCTTTAGTCTACTACTAGTCTCCTTTAACCGCTAGTCTGCACTAGTACACCCAATCTCTAATTCCAAGTAGGGGTTAGAACAG TTAAATTTTG

Table 3.3 (cont.) – List of primers used to clone the synthetic promoters of group M.

Name	Sequence
M4	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTTTGTCTATAAAGAGATTTCGAGGACACTACGCTAGCAGATTGTGAGATTAATCGTTTTGAGCAAG TTATCAAAGAAATTTCACTGCTGGGCTTTTCTTGGCCACTCTCACTATCTTTACTGATCTCGTACTACTAGGTAAACAAGTAGGGGTTAGAACAG TTAAATTTTG
M5	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTGTATTGTGGTCCTTATACAGGACGTTGTATCACCTGAGGTTCTTTTTCTAGCTACACCAAAAGAT TATTAATACCTAAACTTAGTGAGATAAGTTATGATGTTATAACTAGTTATGTCAAGACGGGCTAACTCCAATAGACAAAGTAGGGGTTAGAACAG TTAAATTTTG
M6	GTTCTTCTCCTTTGCTAGCCATCGTTTCGCTATTAACAATGCGCGTAGATGATTAGATGCTTAACCTTATACTAAGAGTTGATGACGCGGCCGTT GCTTTTTCAAGATCTTAAGTTTTTCAGATCTTTGCTTCAAATCGCTAACTATTAATAATACGCCCTAGAAAAATCGTAAGTAGGGGTTAGAACAGT TAAATTTTG
M7	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTAGTTTCGATATGGGTAAAAGGGCTATATGAGAGGGTACTCAGTGTCCTGGAAAAATTTTTGTTGG TAAGTTCGAATCTATAGTGTTAAGCTAGGCTGTTCTATTGCTAATAGTCCGTCTTTGCGTCTTCAACGTATTTGGTCAAGTAGGGGTTAGAACAG TTAAATTTTG
M8	GTTCTTCTCCTTTGCTAGCCATCGTTTCGCGTTAGGGTCTATTTTTATGAGGACAACAGCGCTCTAGTATACCTTTCGTAGGGCCGGAAACTAT CTAAGTGCCGCTATCGACTAGAAGCTTATTATCCCCAAGATCAAATATATTGTTGAAAAGGATTATCTCAACGGCTGCAAGTAGGGGTTAGAAC AGTTAAATTTTG
M9	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTTAACTTAACAAAATACGTAGATGATTCCTCTTTTAACTAAAGGAATGACTATTAACTCGATAGCTC CTAGAGAACGTACCAGATTTTGGTGGTTTTCTTTTTTTGTCCTACTTTCTGATTACTACTATACAAGAAGTTTAAGAAGTAGGGGTTAGAACAGTT AAATTTTG
M10	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAGTAAAATTATCTTCCTTCAGTTGGCAAGGTCCTCCACACGGATACTTTATCCTAATAGAGTTGCG ACAACTACTATGAACTATCCTTAGGTAAAAGCGGCCCCAAAAATAAACTGTACCTTGTACAACTTACTAAAGTACGTAAGTAGGGGTTAGAACAG TTAAATTTTG
M11	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGACTGAGGTCCTTCGTATTTAGGTGTATGCTAATGTAACCTCTATCCTATTCGAGTCACAGTGGCT CCCAAGTAAATGTCCACACTTAATGAACGCTACTCAATTATAACCACTATGTTAGCCTTAAAATGGCTACTCAAATAAAGTAGGGGTTAGAACAG TTAAATTTTG

Table 3.3 (cont.) – List of primers used to clone the synthetic promoters of group M.

Name	Sequence
M12	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAACTGGGGTCACTCAAAATTGAGACCCTAGACTTTAAGTATCCTACAGCTAGTACCTGCACCGCT ACCTCAAAACTTTGAACGTTGAAATCGATTGCAACGAACTTGTAACGATCCTGTTAGGAAGCTAAGTGTATAGTGAATAAGTAGGGGTTAGAAC AGTTAAATTTTG
M13	GTTCTTCTCCTTTGCTAGCCATCGTTTCGCCGATCAAAGATTTTTCTAGATAGTTATTCACTACGGTTTTATTCTATACGAACTAAGCTTCACGGT AGTGTACTATTCGCTATGTGGCTACTTTGACTCCGCTGGATTCTATCGCTTTTAGCAATATATAATGAAGTTATTTAAGTAGGGGTTAGAACAGT TAAATTTTG
M14	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGGAACCGAGCTCTATTGTGAATACTGCTGCCGGTACGCTCGGTCTGATTAGAATCTCTATAGATA TGGAGCCGACTCGCGTGTTTGGACGATGTATATCTATTTGAACCCCAAAATTTAATCGCTACAATCCTCGAAAAAAATAAGTAGGGGTTAGAAC AGTTAAATTTTG
M15	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAATTTCAATACTTGAAATAGTGAGACTTATCTAACTATTAGTAAGTATGGAAGAGAACTGAAGAAAG AGACTATTATAAGAATTCAATACCTCTTATTTTTGAAAGCTAGATCTGAGTTACTGGACTTTCCTCGACACTACAGAAGTAGGGGTTAGAACAGT TAAATTTTG
M16	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAATAACCTGTGATATTACACTCTCTTATGATCTTTCTATCAGGTTCCTTATGCTTCTGGCTATTCGA TAATTCCTAGTCTGCTCACAAAACCAGATTCTATCTTTCCGTACTATCTTCTTGTTAAGAAATCCCTGTCTAGAGAAAGTAGGGGTTAGAACAGT TAAATTTTG
M17	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTCCTGTTTTGGTTGCGAGATTACCCCTTTCTAAGTTTTCTCTACGTTATGGGGCTCCAACTAGCTA ATCGGTTGTTCGTAACTGGTCTGTTCGGCGGTTAGTGAGATACCAGGTGTATTTGGGTTATTGTACGAACAAATATTAAGTAGGGGTTAGAACA GTTAAATTTTG
M18	GTTCTTCTCCTTTGCTAGCCATCGTTTCGATAAGGTTGACTTTGGACTCGCGTATTCTTGCTATCGAAGTTTGATAGAGTGATCGTCTCTATCTA TTTTTGGGAGTACTGTGCAACTGAGTATTGCAAAATAACTTGATTGTTTATGAGATATGCTAGGTGTATGAAACCCTAAGTAGGGGTTAGAACAG TTAAATTTTG
M19	GTTCTTCTCCTTTGCTAGCCATCGTTTCGATAAAATTTTGCCGGGCTGTACGTCTTATCTAGATGTCGTTAAAACCTCAGGCCAAGCTCTATATA CTGCAACAAACCGCTAGCAAAGAAATTTAATATACTACACTCTTAAAAATGTATAGAGTTATTTTACTCAATTAACTAAGTAGGGGTTAGAACAGT TAAATTTTG

Table 3.3 (cont.) – List of primers used to clone the synthetic promoters of group M.

Name	Sequence
M20	GTTCTTCTCCTTTGCTAGCCATCGTTTCGATTGTATCTACTTTTCCCACTATTCAGCGTAAACAGACTAATGCTCCTATCTCTAGAGCTTGGTAGA TTAGTACGTGAAGGTATTAAAAATCTTTTTGTTTCCAATTGAAGGAGAGTTTACCTAACCTTACTGTAAGAGTGTCAAGTAGGGGTTAGAACAGT TAAATTTTG
M21	GTTCTTCTCCTTTGCTAGCCATCGTTTCGCAAGGCAACTCGAGTATAAGTCTATTTGTATGCCGTTGATTCAAGAGAGTTCTGTGCCGTTAAAAA TTAGAATGTAATTAGAAGTAGCAATTCAGATACGATTGAATGGCCAATATCTGAAATTTAAGGTAGGGACTAACAACAAGTAGGGGTTAGAACA GTTAAATTTTG
M22	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTTACTTTCTTAATCTAAACAAGGATTCTATTCTTTGCGCCTAAAGGTTTAGGAAGTTTTCTGTTTGC CGCGTAGTTGATTTCACAAGAACAGTGAAGTATGGCTCGATCTACTAAAATTGAAAGCTAAACGTGGGATAGGTAAAAGTAGGGGTTAGAACAG TTAAATTTTG
M23	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAAAAGTCAATATATTGGAAAATTGCTCAATTAAATTGATGTACCCTGTAAGTGGTAAACTTTGTAAA CGTAGTTCACTCAGAATATATCGAGCCACGACTTTAGAAAATCCTTATCTACTTACGAAACTTAGAGTTCTCGAATAAGTAGGGGTTAGAACAGT TAAATTTTG
M24	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTGCCTTTTTTTTGAGACTGTTATTTAGTGATACACTAGTGTTAGGTATAATCGTGATTCTTAGATTG TTATTTAGATATTTCTATAGACGGCTAACTTTTTACCAACTTAAACTTACAGTATATACGCTCTCTATAATTGGCTAAGTAGGGGTTAGAACAGTT AAATTTTG
M25	GTTCTTCTCCTTTGCTAGCCATCGTTTCGATCTAGTAATTTATAAAACTACCTTGAGTGAGTGATCGATATTGTGATTCTGATCACGATGATCTCA CCTCGATATTGGATAGATGACTCCTTAAGTGCCAAGAACCTAACTAAACTTGATTGGTTAATTATTACGGGTGATAAAGTAGGGGTTAGAACAGT TAAATTTTG
M26	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAATATGTGTCCCTATAAGTGGTCCTACTCACGAAGGATGAATCAAATGAGTATGGATTCAAGGAAG GTAACTATGGATCTACTTGTGTGAAGGAACTATTTATCTTTTATGTGAGATAGGTGCCTCTAGTCAAATATTAACCTAAGTAGGGGTTAGAACAG TTAAATTTTG
M27	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGACTAAAGACCTTTATAAGGGACTGCGTGGTAAATAATGCTATTAAACTCTTTTTGATATTAGGTAT AGTACCTTATCGCTCAACAGTTTTACTACTTCTGCCGCCTCTGATTACTGATCGACAATATTATAGTATTATAAATAAGTAGGGGTTAGAACAGTT AAATTTTG
M28	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTATAAATAATGTAATAACTTTAAGATATGGATTGCACTTGTAACTTAAGCAATTGAAGTTTGCTAAC TAGTTATGTATCTTTTTCTACCTTGGCCGTCCTATTAAATATCTACAAGAGAGATCACAATTGACCGAAAGATTACAAGTAGGGGTTAGAACAGT TAAATTTTG

Table 3.4 – List of primers used to clone the synthetic promoters of group T.

Name	Sequence
T1	GTTCTTCTCCTTTGCTAGCCATCGTTTCGCCTATATACGTAGCTACGAGTTAATGACTACAAATATCGCTCTACTCGTAGATGGAGGATAAGGAA CACAAAGGAGCTCGGTTCTATACAAATTCGTTTAGTATTGATTTTATTTATATCAATGTCGTAACGTCGTGTGAAATAAGTAGGGGTTAGAACAG TTAAATTTTG
T2	GTTCTTCTCCTTTGCTAGCCATCGTTTCGATTACTAATCGAAATTACGTCACGTAGTAAGTCAATGTATTAATTACAGAGTATCCTTAATTATAAG ATTGCGTCAGCGAATCGTGTATCTATTCCTATTTACTGTGGATAACGTGAACTTATATATATCTGGTTAAAGCGCCAAGTAGGGGTTAGAACAGT TAAATTTTG
T3	GTTCTTCTCCTTTGCTAGCCATCGTTTCGACTATACGATCCACTTCTACTTTCTAGGTATGAAGATGTATGTTAGATCTCGTTTTGTTAGTCGTTA GCCGTGCAATACGTTACTTGACCCTGATACTTATATAGACTATTATACTTTTACTTGTGGTGTTTCTATCAATTTAAAGTAGGGGTTAGAACAGTT AAATTTTG
T4	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGAGGCTACGCGATCGATAGGTTATTATAACCACACACTCCAATATATTTATACTGTCCTCGTCTAA CTTTAGTCTCTAAACGTTCTCGTGTGATACTTATCCTAGTTAATATGGTTACTACGTAAAACTTGCAGTATCCCCCGAAGTAGGGGTTAGAACAG TTAAATTTTG
T5	GTTCTTCTCCTTTGCTAGCCATCGTTTCGACTAGCTATTATGGGTAAATCTGAACAACTGTAGTAACTGGAAAATCGTTGTTTTTATAAACTACAC TGCTAGAGCTATCCCTCTGAGTTAAGAATTCGTTGAAGTCTATCGTCCTACTACACGTAATTCGTTCTATTCACTAAAGTAGGGGTTAGAACAGT TAAATTTTG
T6	GTTCTTCTCCTTTGCTAGCCATCGTTTCGACTAATTAGGGGAAGCGTTTCTTACAGCTACTGAATCTAGTGCGTTACCTCTATACGTATAAGTAC TGTGAAACCAATGCTATCTACTATATTCGTTAACTTTTTATATATTTAATGTTTTTTTTAATAACACTGGCTATTTAAAGTAGGGGTTAGAACAGTT AAATTTTG
T7	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGGGTCTAGGAAGGAGCGAGTTCTACAAACACGATGGGGGGATGAAGCTCCTATTATTTAATGTAG ATCTAGAGTAATCGTATTTATATTTTAATTATGGGGGTGATCAAGGATACTAGACTTACAATCCTCTAATCTGACTGAAAGTAGGGGTTAGAACA GTTAAATTTTG
T8	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAGACTAATCAGATAATGAGTACGGGAGAAGAGAAAATCTAAGTAGACTACTCGTAGCGAACGTTA CCCTCGTTCTACTATGCTTTTATACTATCGAACAAATAGACTATTCCTATAAAAAAGTATTCCTTGGGATTCGTCCCTAAGTAGGGGTTAGAACA GTTAAATTTTG

Table 3.4 (cont.) – List of primers used to clone the synthetic promoters of group T.

Name	Sequence
T9	GTTCTTCTCCTTTGCTAGCCATCGTTTCGATAGGATTATGGTGTAGGACAGCACTCACCTATCAAATTCGTTTTTCCTACTCGTACCCTTCTAATA TTCCAATTCTAAATAGGACTATTATATCGGATGTTACACTGCGCCTTATATACTTGGCTATCAAATCTCTAAAGTTAAGTAGGGGTTAGAACAGTT AAATTTTG
T10	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGTAAAAGACTACTTCGGCTATCTAAAAGTTTAACAGTTCAGGGTTTATCGCAAGATTACTCGAGTA ACTGATTTCCTTAAGATAAGTTTTCGCGAACTCACCTAGGAGGTACCTTTTATAATGGATGTCCTTGGTAAGAGAGAAAGTAGGGGTTAGAACA GTTAAATTTTG
T11	GTTCTTCTCCTTTGCTAGCCATCGTTTCGATCCTTAGTAAGAATCTTCCTAGAAAAGGCAAAATAGATATCTCTCTCTAGTTGCTATGGTTGGTTT CCAGGCTACCTTGACCAAATTTTCTATGTCTAGCTTCAGACGATCTATTTTTATATAAAAGTAAATGTTTATAACAAAGTAGGGGTTAGAACAGTT AAATTTTG
T12	GTTCTTCTCCTTTGCTAGCCATCGTTTCGCGGCTCTCCTTCTGTTTATCGATATTCTAATTTATCGATGGTTATGTTGCGACGTGTTCTAACCTAT ATTCTGGGAGATTATTTATAACTGATAAACTGTCTATTATAGTTTTCTAACTGTGGCTATAGATTTCAGGGTGATTAAGTAGGGGTTAGAACAGTT AAATTTTG
T13	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGGCTTTTCTCAAAACGATTTCTTATCGGACCGATATTAGTGTTAACTATATAATATTCGCCGGAAC GGAGACTAGATATATTAAACCTTATATAGACGTAGTTCGTCTAGAATTCAAGTCGTATAGTGAGAAGTTTAGCTAAGAAGTAGGGGTTAGAACAG TTAAATTTTG
T14	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTATTCCTAAGACTAGAATACTATCACTAATCTAGAACGGGAGATGTAGGCTCGAGAGATCCAGTCT GCGTAAATATAGCGCCACCAAACGTACTGATTTTTATATGTTTACCAAAAGTACGACCCCTATTGTTTCCTACCAATAAGTAGGGGTTAGAACAG TTAAATTTTG
T15	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGATAACTATAATAGTATATAGGGAGCTATAATAATGCAAATTCCTCGGTGGTTGGAAATTATTACTT TAAATATTATAATTTATATACTAGTCGCCCGTTCGTATGCCCAAGATATATTCGTTTTGCCACTATCGTTTCGTTTAAGTAGGGGTTAGAACAGTT AAATTTTG
T16	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTATTTTTCTAACCAGCGTAGGATGGTAAATGAAATGTAGATCTATGGGTTGCGTTACTACTTTCCA GATTCGTTTTAATTTACAAATATAAGTATAGCTTATATACAGTAGATATATATGATTGGACGATCAACAAGACGTATAAGTAGGGGTTAGAACAGT TAAATTTTG

Table 3.4 (cont.) – List of primers used to clone the synthetic promoters of group T.

Name	Sequence
T17	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTCAACGTATTAGGCAAGAATATTTTTATGTAAAAAGCTTCTTAAATTCGGACTATAGTAGTTCCCTC CCCTTCCCACTACGCCGCAAAGTTTGTTTTGATCGTTTATATAGATTCGGTTCTGTACCTTGATCCCAAGCAAAAAAAGTAGGGGTTAGAACAGT TAAATTTTG
T18	GTTCTTCTCCTTTGCTAGCCATCGTTTCGCAACTATTATGGGCTCTGTTATCAGTAGTATTCGATGTTCAGTAGTGTGTCTATGTTTATCGAAGAT CTCTGTAGTTACCAATGGTGCGTATTTTTCTACCGTTTTTTATACTATCCAGACAGTTAATCTATCGGCCTCCTAGAAGTAGGGGTTAGAACAGT TAAATTTTG
T19	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTATGATATATAGCTATAAACTTACACGTCAGGAATATGTTCACTCTCGCAACAAATCACACACTAAC CCCTATAAATTCGTTTAGCTTTCGTTCTAACAGTGTCTGGTAGAGTACTTATATAAATTTATAACAAGATCAATTAAAGTAGGGGTTAGAACAGTT AAATTTTG
T20	GTTCTTCTCCTTTGCTAGCCATCGTTTCGCTTATAAACTTATGCAATGATCAACTAAATCTACGAATTAGATAATATTTGCACGCGAGGTTACCTT CTATACGATAACCTGATATTTATAAACTGAGAACCCCAAAGAAAGGTTTTAAGTATCTAACACTATCGTGCAGAACAAGTAGGGGTTAGAACAGT TAAATTTTG
T21	GTTCTTCTCCTTTGCTAGCCATCGTTTCGATATGTATCAGTACCCTAGGAATATCGTATAAATGTTCTATTTTCTATGGAATATGGGTGTCAGCTT TATATATTGTGCACCTGATAATTGTAGGACCAAATTTGGGTCGTAATCTCTACTAATCAAGATTGGCAAGAATTCTAAGTAGGGGTTAGAACAGT TAAATTTTG
T22	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAATAAGAGTTTTATGTACTTAACGTACTACCTGGAAAACGGACAAACTAAGCTACTAAGTAAGCTA TGTTAGTACGCGTCTATATCTATTGAATGTCTAATTAAGAGATTATTTATAACTACCTAAAATTAATACCGTCACGAAAGTAGGGGTTAGAACAGT TAAATTTTG
T23	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAGCTTCTACTATTTCAACTAGAGTTCTAAATCTTAAAAGTAAATTCGTATCTCAGATTGGGGATTCG TCGGTGGAAGGATACAACTTTTATAAGAAATTACTACGAATAGACTTTTAATGATACAATCAATGTGGGTTAAACGAAGTAGGGGTTAGAACAGT TAAATTTTG
T24	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTACTGAGCTATAGCGTGGTTTAGCTAAACAATCGAAAAATCTCTGGGGTAACTAGAGCGCTTAATA GGTACTTGTGCTAAATTTGTACTGATAATTTACGAGCGAGACAATATTTATAGACTGCCCTAAGTAAGTTCGAAATTAAGTAGGGGTTAGAACAG TTAAATTTTG

Table 3.4 (cont.) – List of primers used to clone the synthetic promoters of group T.

Name	Sequence
T25	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGTGATTAAATATACCTATTCTCGTGGAACGAACGGGGTCAGTTTTGGGTCTGAAGATCGTCAATCT CCTGAATTTATACGATATCAGATAATGTTTTTGATGAGGTGAGATTGACTTGCAGGACCACCTTTTATAGTCTATATAAGTAGGGGTTAGAACAG TTAAATTTTG
T26	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTACCTTCCTATACTACTCTAGAATGTAACCTATAATCTTAATAACTTCTATCTGCTTTTTATCGATTC GAATCTAGATGTTTGATTATGCTTTAAATGAAAAATTCTAACGACTTCAGCTTTTATAGATAAACCTAGAGGAGGAAGTAGGGGTTAGAACAGTT AAATTTTG
T27	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAACAGAAAAAGAATGTGATGTTAAAAGCTTTTTACGTGCAAAATACCCTCACTTCAGATATCGGAA CAACAAGGGTATGCTATACAAACGAAAGGTTACTGAGGGACTTGACCTACGTCTGTGACGGTTTATATACTAGTAAAAAGTAGGGGTTAGAACA GTTAAATTTTG
T28	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAATAAAATAAGGTTTCTTTTAACTAGTTATAACAGTTTTACTAGACTTATACTTCTATAGGTTATTCT AAGTTATACAATAATCTAGACAGCAATAAGCTATATTTATAGAAAACCGAGCCTTATCTATCGTGAAAACGAGCTAAGTAGGGGTTAGAACAGTT AAATTTTG

Table 3.5 – List of primers used to clone the synthetic promoters of group A.

Name	Sequence
A1	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTTTTGTTGTTGTTTGCGAATATATACTAATTCAGCTTAATGAACACCTAACTATACTCTTTTGAAATA CAAGTGGGGGTGACTATTTCACCTACAGGTGTCAAGTTCTAAAGGTATTTAAATCCCTTTTATAAGGTAATTTATAAGTAGGGGTTAGAACAGTT AAATTTTG
A2	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTCTTAAAATGTTGAGCCGCACCTGGGCACAAAGAAAGTATTGCGTTGTTATACGACCTTGCGGCC GACGAAATGCTATTATGATAATGCTTGATGTAGTGGGTACAAGATTATTGTGTAAAATATTTATACAACTAAGAATATAAGTAGGGGTTAGAACA GTTAAATTTTG
A3	GTTCTTCTCCTTTGCTAGCCATCGTTTCGATAACGCTACTGAGGATTCTTTAACTGCGTATGATCTAGCTCTAGAATTATTTTATTCTCGTTAGTT ACTCCTTATATACCTAAGATAGTTTACGCCGAACTCGCTCAATTGATCAACGTTATCTAACTTGAGCTTTAAAGTGAAGTAGGGGTTAGAACAGT TAAATTTTG

Table 3.5 (cont.) – List of primers used to clone the synthetic promoters of group A.

Name	Sequence
A4	GTTCTTCTCCTTTGCTAGCCATCGTTTCGACCACTTTACTGACAGCTACCCAGAAGGGGACTAAGATATCGGATTTGTGGGAGTTAAAGAAGTT AATAACAACGTGTATAGATAAAAATCTTTTTATAATGAACAAGCTTACTGTATTCGCACAACGAAGTCGCTTTGCTTAAAGTAGGGGTTAGAACA GTTAAATTTTG
A5	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAGCACAATTTTGAGCTAATTGATGCTATGCTAGGAAACTTTCTGTTGATCGTTACTATGTTTTTAGT TATAGAAGGCTATCAATTCTCGATTGTTGATAAACGGGGATCTAACTTTTATACTCAAAAACAACGGGAAATTAGGAAGTAGGGGTTAGAACAGT TAAATTTTG
A6	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGTGTGCTTGACAATCCTCCTAGCACTGGCTTTGTTCTTTGACCTACTGAATTCCTCTAGTAAATAC TACTGTGTAATATTTATATACGGGAGTCGAAAGCCCAGCAAACTTTCCTCTATCTACAACAAAATGTACTACTAACCAAGTAGGGGTTAGAACAG TTAAATTTTG
A7	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAGATACTAAATTTCCTGTGGGATTTGATTGATCTACGCTACCACGTAGAACAATATTAAACCAATAC TATCTGGCACTTATGCGAATTCTTGAGATATTGTTCAAGATTCCAATATAAGCTTTTATATCTGTGTTCTTCAATAAAGTAGGGGTTAGAACAGTT AAATTTTG
A8	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGATGGATTAATGGGTAGCTGTATGACGATAGTATAACTTTTGATGACTTTAATTTAACCTCGACTCA AGTAAACTTTATGTCGCTAATTCTTGATCTAATTTCTTATATATAACTAGACAATCAGGACCCTAAACTATGATTAAAGTAGGGGTTAGAACAGTT AAATTTTG
A9	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGGATCCTATGAGACGTATTTCTTCTATTGGATTGTTTCAATCTACTATGAAGGTGATCGTCTACTAC AGCCACCCCTACTCCAACCTAATACTTTTATATGGGTTGGTTTCTTTGCCCCGATTCCGCTAATAATACAAATTTTAAGTAGGGGTTAGAACAGT TAAATTTTG
A10	GTTCTTCTCCTTTGCTAGCCATCGTTTCGCTTACGGCTAGTAAGTACTTGATTCGGACCCCTATATACAGGATATTTGTATTGGGAACTATCGTA TAAGCTTACGTGCTCTATCTTAGGTAGGGAAAAACACCACTATCTACAGAACCTATCGTTTTTATTTATATATTAATAAGTAGGGGTTAGAACAGT TAAATTTTG
A11	GTTCTTCTCCTTTGCTAGCCATCGTTTCGGGTATCTCTAGACTATTGCGACTTTCTACTACAACCCACTACTAATTGATTGCGTCTATAGATCTTC GAACACTATATGGATTTAAAAGTACTTTTATAATGCTATTCTAAGAATTACTTTAAGAATTGGATGTTAATTCAATAAGTAGGGGTTAGAACAGTTA AATTTTG

Table 3.5 (cont.) – List of primers used to clone the synthetic promoters of group A.

Name	Sequence
A12	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTTGTAACTACTAAAAGTTAATATTAAGGTATTACTATGGATCGCTCTAATTTCAATCTTATACGTTTG TCTAACTCGCTCAAAATATTTATACGAAATTCTTATATTGGTTAAGGTTGAAAACTCAATTTAGGTTCACACGATAAGTAGGGGTTAGAACAGTTA AATTTTG
A13	GTTCTTCTCCTTTGCTAGCCATCGTTTCGACAAATTGCTATTTTATTGTCTTAATAGTATCTGTAGATTATTGATATATGATATAGAGGTAACTCTT TTCCAGAATAGAAAGGTCACTAGTCTTTCTGTACAACCTAAACCGTAATTTTTATATTGACCAAAATAGAAAGAGAAGTAGGGGTTAGAACAGTT AAATTTTG
A14	GTTCTTCTCCTTTGCTAGCCATCGTTTCGACTATAGGTTTCTGTTGGTAATTACGTTTACTCTTATCGTGGTGGTACAAAAAATACTAAAGTATCA AATCTTGAATTTATAGAACTAAACTAAGTGGATACCTATTACTACCTAGTTACTATTATAATTATATATATGAAAAAAGTAGGGGTTAGAACAGTTA AATTTTG
A15	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTTCTACCCTTTCCAATATATCGGGATATTGATTAATAATTACTGCTAAGCGCGAGGCTTATATAAGG GTATGTAACGAGAGGAATATCTTTCAGGGTCAACGTTGGATTACCCTATTTTACCTATTATATAAGGGTTGCGAGTAAGTAGGGGTTAGAACAG TTAAATTTTG
A16	GTTCTTCTCCTTTGCTAGCCATCGTTTCGCAGCTTGCGCTAACCTATTCTAAATGTAAATTGCACTATCTACTTACTTATATCTTTCTAGCAACGT AAGTAGTTGGAACTATTACTCAAAAATTCAATTTCTTTTATACCAGATCAATTACCAGTCTATGTGCGACTTATGTAAGTAGGGGTTAGAACAGTT AAATTTTG
A17	GTTCTTCTCCTTTGCTAGCCATCGTTTCGCCTACTAATCTATGGGATAAAGAAGTGAAATTGATGCGATTGCGTGACAAAGAGAGCTATACTTAC TCTGTCTTATTAGTTTTACTGTGGCTTTTTATATATTTCTACTTTAAGACTAGATTGACCTATTTAGAATATATAATAAGTAGGGGTTAGAACAGTT AAATTTTG
A18	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTAGTGGCTGTATGTAGTCTCAATTCTAATATTCTTATCAATTGATTATTTTGGTTGAATGTAGGTAC TAATCTTATAAATTAAGAAGGTTAATAAATCTTAACTCAATGAGAGAATAGTTTACAATTTTTTATATCTAAACAGAAGTAGGGGTTAGAACAGTTA AATTTTG
A19	GTTCTTCTCCTTTGCTAGCCATCGTTTCGCCTATAATTGGGAGAAACGAGGATAGTTGAATTCTAGCTCCTATAGTATGAACGGTTATGTAAAGG AATGAGATGCTATCTTATATACGAACTTGAGTTTAAAAGGTTATTTTTCTTAAATACTCCACCCAATGGAATTTAGTAAGTAGGGGTTAGAACAGT TAAATTTTG

Table 3.5 (cont.) – List of primers used to clone the synthetic promoters of group A.

Name	Sequence
A20	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTAATATATTTATTATTGTTCACCTCTCTGCTATGGAGTTACGACTCGTCGACGGGATATGTTTACTA GGAGGTCTTACAGGGTAGATCTTTCGAAACTAGTGATTTATATATATCTTTTTCTATGACCTCCTAGTACGTTATCAAGTAGGGGTTAGAACAGT TAAATTTTG
A21	GTTCTTCTCCTTTGCTAGCCATCGTTTCGCTTTAACTACGTCGAGGGACCTCTATTTAGTCTGGCCGAGTTGCGTAGCTAAAGTTTAAGATCTAC TAAAGTTTAGTATGTGTTCGATAATGATTTTCTCTCACGTTCGCTTTTCGAAGTTTTTTCGCCTTTTATAATCGTTAAAGTAGGGGTTAGAACAGT TAAATTTTG
A22	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTGTCTACAACTTACTAATCAACTCACTACGACTTTAATAGTCAACAAGAGAACGAAAGAGATTCTC CAAATACAATCTTTGTTTATATATGTGATGTCGATTAAGTCTTCCTATCTACTAATGTACACTGCAGTGAGGATGTCAAGTAGGGGTTAGAACAG TTAAATTTTG
A23	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAAGTGATCGTAAGAGAATTTGAATCTTGACTAAGCTAGGCTTAGGTGTTGAAGTAAGCGCAAATGA ACTTAATCTTGTTTGCTCTTTAGTAAAATTCAGTGATTGGAGATCACTTTTATAGAGAACAAGTATTTCGTCAGCAAAAGTAGGGGTTAGAACAGT TAAATTTTG
A24	GTTCTTCTCCTTTGCTAGCCATCGTTTCGTATTTGGCTGCCCGTTACGTAACCCAAAAATAGTGAATTTCCTGGCCGAATAGTTCTGTTCAGTGG GCCTTCTACTATGTATGGGAGCAAAAACCAACACCCTTATATAAGTATTCTACAAATTAATCTTGGCACCCAAAAGAAAGTAGGGGTTAGAACAG TTAAATTTTG
A25	GTTCTTCTCCTTTGCTAGCCATCGTTTCGATTAACAATACGCTTACCTCGCAATTTGGTATCTAGATTGCTTAAGCAGGCCAATCCTTAATGTGC TTTCTTCTCGTTGTCAATTAATCTGATTGACGCTTTCTATTATTTATACCCGTCTTGGTGAATTCAACTATATAGAAAAGTAGGGGTTAGAACAGT TAAATTTTG
A26	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAATGATATAGAAATGTCCTATATGTAATTCTAAGACAGTCTAATTTTCTATATAAAATTGCAACTACT ATCTAGGAAACTGCAATGTCACACAGTAAAAACCTGGGCCTTATCTTATATATGTTAGAGCCTATCCTTAGTGACAAGTAGGGGTTAGAACAGTT AAATTTTG
A27	GTTCTTCTCCTTTGCTAGCCATCGTTTCGAATAGGAGTGAAGGTATAATTCTGACTCTGAGCTTTTTTGTAGTAAATGTTTCGCTAGATAGATATC TATGATTAGTACTAGTAAAATAGTCCCTGAATATTATAATGTATTAGTGTCCCAGATTTTTTTAGTTTATATACTGAAGTAGGGGTTAGAACAGTTA AATTTTG
A28	GTTCTTCTCCTTTGCTAGCCATCGTTTCGACCCTTCCGATTATTAATGTTACGGTACCTTCTTTGTTGTAAATGTTCGAAGAATTGTTTCTTCCGT TTAAACCAAGAGTTGCTTGTACTGCGTTGGTATATTTGGTATACAACCTCCTACTTTATATATTTCTGCCCTATCGAAGTAGGGGTTAGAACAGT TAAATTTTG

Table 3.6 – List of primers used test the ten best synthetic core promoters fused to different CRMs in *P. pastoris* and *S. cerevisiae*.

Name	Sequence
CAT-core	AAAGTTCTTCTCCTTTGCTAGCCATCGTTTCGTTTAATTGTAAGTCTTGACTAGAGCAAG
CAT-CRM-forw	GCTGGCCTTTTGCTCACATGTATTTAAATTAATCGAACTCCGAATGCGGTTC
DAS-core	GAAAAGTTCTTCTCCTTTGCTAGCCATCGTTTCGTTTGTTCGATTATTCTCCAGATAAAATCAACAATAGTTG
DAS-CRM-forw	GCTGGCCTTTTGCTCACATGTATTTAAATAGCAATGATATAAACAACAATTGAGTGACAGG
GAP-core	GAAAAGTTCTTCTCCTTTGCTAGCCATCGTTTCGTGTGTTTTGATAGTTGTTCAATTGATTGAAATAGG
GAP-CRM-forw	GCTGGCCTTTTGCTCACATGTATTTAAATTTTTTGTAGAAATGTCTTGGTGTCCTCG
CAT-CRM-rev	TCCAACAAAGAGGCAACAGAGGTCGGCGCGCCACTGGGTGCTACTGATGAGCAACAGAGGCTATCAC
DAS-CRM-rev	TCCAACAAAGAGGCAACAGAGGTCGGCGCGCCACTGGGTGCTATGCTTTAGTTCTTTTTTGAACCCAAAGGCTATCTGATGAAAAG
GAP-CRM-rev	TCCAACAAAGAGGCAACAGAGGTCGGCGCGCCACTGGGTGCTAGTGGTTTCCAATAATCTCATGACATGCG
seqTomato19-41rev	CGCATGAACTCCTTGATAACTTC
ADH-CRM-rev	TCCAACAAAGAGGCAACAGAGGTCGGCGCGCCACTGGGTGCTAAGACAGCAAACTTTTTTTTATTTCAAATTCAAGTAAC
GAL-CRM-rev	TCCAACAAAGAGGCAACAGAGGTCGGCGCGCCACTGGGTGCTAGATCAAAAATCATCGCTTCGCTGA
GPD-CRM-rev	TCCAACAAAGAGGCAACAGAGGTCGGCGCGCCACTGGGTGCTAAAGTAGGGGAATAATTTCAGGGAACTG
A28-GFP-rev	AAAAGTTCTTCTCCTTTGCTAGCCATCGTTTCGACCCTTCCGATTATTAATGTTACGGTA
T28-GFP-rev	GAAAAGTTCTTCTCCTTTGCTAGCCATCGTTTCGAATAAAATAAGGTTTCTTTTAACTAGTTATAACAGTTTTACTAG
A27-GFP-rev	GAAAAGTTCTTCTCCTTTGCTAGCCATCGTTTCGAATAGGAGTGAAGGTATAATTCTGACTCTGAG
T27-GFP-rev	GAAAAGTTCTTCTCCTTTGCTAGCCATCGTTTCGAACAGAAAAAGAATGTGATGTTAAAAGCTTTTTACG
T26-GFP-rev	AAAGTTCTTCTCCTTTGCTAGCCATCGTTTCGTACCTTGGTATACTACTCTAGAATGTAA
T25-GFP-rev	GAAAAGTTCTTCTCCTTTGCTAGCCATCGTTTCGGTGATTAAATATACCTATTCTCGTGGAACGAAC
M28-GFP-rev	GAAAAGTTCTTCTCCTTTGCTAGCCATCGTTTCGTATAAATAATGTAATAACTTTAAGATATGGATTGCACTTG
T24-GFP-rev	GAAAAGTTCTTCTCCTTTGCTAGCCATCGTTTCGTACTGAGCTATAGCGTGGTTTAGC
T23-GFP-rev	GAAAAGTTCTTCTCCTTTGCTAGCCATCGTTTCGAGCTTCTACTATTTCAACTAGAGTTCTAAATCTTAAAAG
T22-GFP-rev	GAAAAGTTCTTCTCCTTTGCTAGCCATCGTTTCGAATAAGAGTTTTATGTACTTAACGTACTACCTGG

Table 3.6 (cont.) – List of primers used test the ten synthetic core promoters with different CRMs in *P. pastoris* and *S. cerevisiae*.

Name	Sequence
A28-CAT-for	GTGATAGCCTCTGTTGCTCATCAGCGATAGGGCAGAAATATATAAAGTAGGAGG
T28-CAT-for	GTGATAGCCTCTGTTGCTCATCAGAGCTCGTTTTCACGATAGATAAGGCTC
A27-CAT-for	GTGATAGCCTCTGTTGCTCATCAGCAGTATATAAACTAAAAAAATCTGGGACACTAATAC
T27-CAT-for	GTGATAGCCTCTGTTGCTCATCAGTTTACTAGTATATAAACCGTCACAGACGTAGG
T26-CAT-for	GTGATAGCCTCTGTTGCTCATCAGCCTCCTCTAGGTTTATCTATAAAAGCTGAAG
T25-CAT-for	GTGATAGCCTCTGTTGCTCATCAGATATAGACTATAAAAGGTGGTCCTGCAAG
M28-CAT-for	GTGATAGCCTCTGTTGCTCATCAGGTAATCTTTCGGTCAATTGTGATCTCTC
T24-CAT-for	GTGATAGCCTCTGTTGCTCATCAGAATTTCGAACTTACTTAGGGCAGTC
T23-CAT-for	GTGATAGCCTCTGTTGCTCATCAGCGTTTAACCCACATTGATTGTATCATTAAAAG
T22-CAT-for	GTGATAGCCTCTGTTGCTCATCAGTCGTGACGGTATTAATTTTAGGTAG
A28-DAS-for	GCCTTTGGGTTCAAAAAAGAACTAAAGCACGATAGGGCAGAAATATATAAAGTAGGAGG
T28-DAS-for	GCCTTTGGGTTCAAAAAAGAACTAAAGCAAGCTCGTTTTCACGATAGATAAGGCTC
A27-DAS-for	GCCTTTGGGTTCAAAAAAGAACTAAAGCACAGTATATAAACTAAAAAAATCTGGGACACTAATAC
T27-DAS-for	CCTTTGGGTTCAAAAAAGAACTAAAGCATTTACTAGTATATAAACCGTCACAGACGTAGG
T26-DAS-for	GCCTTTGGGTTCAAAAAAGAACTAAAGCACCTCCTCTAGGTTTATCTATAAAAGCTGAAG
T25-DAS-for	GCCTTTGGGTTCAAAAAAGAACTAAAGCAATATAGACTATAAAAGGTGGTCCTGCAAG
M28-DAS-for	GCCTTTGGGTTCAAAAAAGAACTAAAGCAGTAATCTTTCGGTCAATTGTGATCTCTC
T24-DAS-for	GCCTTTGGGTTCAAAAAAGAACTAAAGCAAATTTCGAACTTACTTAGGGCAGTC
T23-DAS-for	CCTTTGGGTTCAAAAAAGA ACTAAAGCACGTTTAACCCACATTGATTGTATCATTAAAAG
T22-DAS-for	GCCTTTGGGTTCAAAAAAGAACTAAAGCATCGTGACGGTATTAATTTTAGGTAG

Table 3.6 (cont.) – List of primers used test the ten synthetic core promoters with different CRMs in *P. pastoris* and *S. cerevisiae*.

Name	Sequence
A28-GAP-for	CGCATGTCATGAGATTATTGGAAACCACCGATAGGGCAGAAATATATAAAGTAGGAGG
T28-GAP-for	CGCATGTCATGAGATTATTGGAAACCACAGCTCGTTTTCACGATAGATAAGGCTC
A27-GAP-for	CATGTCATGAGATTATTGGAAACCACCAGTATATAAACTAAAAAAATCTGGGACACTAATAC
T27-GAP-for	CGCATGTCATGAGATTATTGGAAACCACTTTACTAGTATATAAACCGTCACAGACGTAGG
T26-GAP-for	CGCATGTCATGAGATTATTGGAAACCACCCTCCTCTAGGTTTATCTATAAAAGCTGAAG
T25-GAP-for	CGCATGTCATGAGATTATTGGAAACCACATATAGACTATAAAAGGTGGTCCTGCAAG
M28-GAP-for	CGCATGTCATGAGATTATTGGAAACCACGTAATCTTTCGGTCAATTGTGATCTCTC
T24-GAP-for	CGCATGTCATGAGATTATTGGAAACCACAATTTCGAACTTACTTAGGGCAGTC
T23-GAP-for	CGCATGTCATGAGATTATTGGAAACCACCGTTTAACCCACATTGATTGTATCATTAAAAG
T22-GAP-for	CGCATGTCATGAGATTATTGGAAACCACTCGTGACGGTATTAATTTTAGGTAG
A28-ADH-for	GTTACTTGAATTTGAAATAAAAAAAAGTTTGCTGTCTCGATAGGGCAGAAATATATAAAGTAGGAGG
T28-ADH-for	TACTTGAATTTGAAATAAAAAAAAGTTTGCTGTCTAGCTCGTTTTCACGATAGATAAGGC
A27-ADH-for	GTTACTTGAATTTGAAATAAAAAAAAGTTTGCTGTCTCAGTATATAAACTAAAAAAATCTGGGACACTAATAC
T27-ADH-for	GTTACTTGAATTTGAAATAAAAAAAAGTTTGCTGTCTTTTACTAGTATATAAACCGTCACAGACGTAGG
T26-ADH-for	GTTACTTGAATTTGAAATAAAAAAAAGTTTGCTGTCTCCTCCTCTAGGTTTATCTATAAAAGCTGAAG
T25-ADH-for	GTTACTTGAATTTGAAATAAAAAAAAGTTTGCTGTCTATATAGACTATAAAAGGTGGTCCTGCAAG
M28-ADH-for	ACTTGAATTTGAAATAAAAAAAAGTTTGCTGTCTGTAATCTTTCGGTCAATTGTGATCTC
T24-ADH-for	GTTACTTGAATTTGAAATAAAAAAAAGTTTGCTGTCTAATTTCGAACTTACTTAGGGCAG
T23-ADH-for	GTTACTTGAATTTGAAATAAAAAAAAGTTTGCTGTCTCGTTTAACCCACATTGATTGTATCATTAAAAG
T22-ADH-for	TTACTTGAATTTGAAATAAAAAAAAGTTTGCTGTCTTCGTGACGGTATTAATTTTAGGTA

Table 3.6 (cont.) – List of primers used test the ten synthetic core promoters with different CRMs in *P. pastoris* and *S. cerevisiae*.

Name	Sequence
A28-GAL-for	TCAGCGAAGCGATGATTTTTGATCCGATAGGGCAGAAATATATAAAGTAGGAGG
T28-GAL-for	TCAGCGAAGCGATGATTTTTGATCAGCTCGTTTTCACGATAGATAAGGCTC
A27-GAL-for	TCAGCGAAGCGATGATTTTTGATCCAGTATATAAACTAAAAAAATCTGGGACACTAATAC
T27-GAL-for	TCAGCGAAGCGATGATTTTTGATCTTTACTAGTATATAAACCGTCACAGACGTAGG
T26-GAL-for	TCAGCGAAGCGATGATTTTTGATCCCTCCTCTAGGTTTATCTATAAAAGCTGAAG
T25-GAL-for	TCAGCGAAGCGATGATTTTTGATCATATAGACTATAAAAGGTGGTCCTGCAAG
M28-GAL-for	TCAGCGAAGCGATGATTTTTGATCGTAATCTTTCGGTCAATTGTGATCTCTC
T24-GAL-for	TCAGCGAAGCGATGATTTTTGATCAATTTCGAACTTACTTAGGGCAGTC
T23-GAL-for	TCAGCGAAGCGATGATTTTTGATCCGTTTAACCCACATTGATTGTATCATTAAAAG
T22-GAL-for	TCAGCGAAGCGATGATTTTTGATCTCGTGACGGTATTAATTTTAGGTAG
A28-GPD-for	CAGTTCCCTGAAATTATTCCCCTACTTCGATAGGGCAGAAATATATAAAGTAGGAGG
T28-GPD-for	CAGTTCCCTGAAATTATTCCCCTACTTAGCTCGTTTTCACGATAGATAAGGCTC
A27-GPD-for	GTTCCCTGAAATTATTCCCCTACTTCAGTATATAAACTAAAAAAATCTGGGACACTAATAC
T27-GPD-for	CAGTTCCCTGAAATTATTCCCCTACTTTTTACTAGTATATAAACCGTCACAGACGTAGG
T26-GPD-for	CAGTTCCCTGAAATTATTCCCCTACTTCCTCCTCTAGGTTTATCTATAAAAGCTGAAG
T25-GPD-for	CAGTTCCCTGAAATTATTCCCCTACTTATATAGACTATAAAAGGTGGTCCTGCAAG
M28-GPD-for	CAGTTCCCTGAAATTATTCCCCTACTTGTAATCTTTCGGTCAATTGTGATCTCTC
T24-GPD-for	CAGTTCCCTGAAATTATTCCCCTACTTAATTTCGAACTTACTTAGGGCAGTC
T23-GPD-for	CAGTTCCCTGAAATTATTCCCCTACTTCGTTTAACCCACATTGATTGTATCATTAAAAG
T22-GPD-for	CAGTTCCCTGAAATTATTCCCCTACTTTCGTGACGGTATTAATTTTAGGTAG

Table 3.7 – Overview of group P synthetic core promoters' features: TATA box position and positive and negative motifs frequency and position.

Promoter name	Postive motifs frequency and position										Negative motifs frequency and position				TATA box position
	TTTT	Position	TTCT	Position	CAA	Position	ATCA	Position	CAAT	Position	AAGA	Position	AGCG	Position	
P1	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P2	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P3	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P4	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P5	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P6	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P7	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P8	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P9	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P10	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P11	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P12	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P13	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P14	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P15	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P16	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P17	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P18	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P19	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P20	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P21	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P22	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P23	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P24	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-

Table 3.7 (cont.) – Overview of group P synthetic core promoters' features: TATA box position and positive and negative motifs frequency and position.

Promoter name	Postive motifs frequency and position										Negative motifs frequency and position				TATA box position
	TTTT	Position	TTCT	Position	CAA	Position	ATCA	Position	CAAT	Position	AAGA	Position	AGCG	Position	
P25	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P26	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P27	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-
P28	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-

Table 3.8 – Overview of group M synthetic core promoters' features: TATA box position and positive and negative motifs frequency and position.

Promoter name	Postive motifs frequency and position										Negative motifs frequency and position				TATA box position
	TTTT	Position	TTCT	Position	CAA	Position	ATCA	Position	CAAT	Position	AAGA	Position	AGCG	Position	
M1	1	-67	0	-	0	-	0	-	0	-	1	-148	0	-	-
M2	0	-	1	-90	1	-26	0	-	0	-	1	-109	1	-119	-
M3	0	-	0	-	4	-49; -46; -23; -16	1	-59	0	-	1	-73	1	-117	-
M4	0	-	1	-83	2	-49; -10	0	-	1	-49	2	-123; -105	0	-	-
M5	1	-69	0	-	2	-34; -13	0	-	1	-13	0	-	0	-	-
M6	0	-	0	-	0	-	2	-60; -30	0	-	2	-104; -88	1	-118	-
M7	1	-62	0	-	0	-	0	-	0	-	2	-135; -127	0	-	-
M8	0	-	1	-94	0	-	0	-	0	-	0	-	1	-83	-
M9	0	-	1	-81	0	-	1	-32	0	-	1	-106	0	-	-
M10	0	-	0	-	1	-30	0	-	0	-	0	-	0	-	-

Table 3.8 (cont.) – Overview of group M synthetic core promoters' features: TATA box position and positive and negative motifs frequency and position.

Promoter name	Postive motifs frequency and position										Negative motifs frequency and position				TATA box position
	TTTT	Position	TTCT	Position	CAA	Position	ATCA	Position	CAAT	Position	AAGA	Position	AGCG	Position	
M11	0	-	0	-	0	-	0	-	0	-	0	-	1	-103	-
M12	1	-80	0	-	1	-27	0	-	1	-27	0	-	0	-	-
M13	0	-	0	-	0	-	0	-	0	-	0	-	3	-123; -110; -88	-
M14	0	-	1	-59	1	-23	1	-54	1	-23	0	-	0	-	-
M15	0	-	3	-88; -71; -63	1	-20	0	-	0	-	1	-101	0	-	-
M16	1	-96	0	-	0	-	1	-37	0	-	3	-126; -123; -110	0	-	-
M17	0	-	0	-	2	-20; -16	0	-	0	-	0	-	0	-	-
M18	0	-	0	-	4	-50; -37; -21; -15	2	-60; -52	0	-	0	-	0	-	-
M19	0	-	1	-96	1	-17	0	-	0	-	1	-116	1	-87	-
M20	1	-96	0	-	1	-10	0	-	1	-10	1	-100	0	-	-
M21	1	-72	2	-89; -78	2	-44; -34	1	-46	0	-	0	-	0	-	-
M22	0	-	1	-95	1	-42	0	-	0	-	0	-	0	-	-
M23	0	-	1	-89	3	-43; -30; -22	1	-45	3	-43; -30; -22	0	-	0	-	-
M24	0	-	0	-	1	-19	2	-63; -37	0	-	1	-67	0	-	-
M25	0	-	0	-	2	-50; -32	4	-69; -60; -54; -42	1	-50	0	-	0	-	-
M26	0	-	0	-	0	-	0	-	0	-	1	-112	0	-	-
M27	0	-	0	-	0	-	1	-64	0	-	1	-58	1	-87	-
M28	0	-	0	-	2	-46; -40	0	-	1	-40	1	-87	0	-	-

Table 3.9 – Overview of group T synthetic core promoters' features: TATA box position and positive and negative motifs frequency and position.

Promoter name	Postive motifs frequency and position										Negative motifs frequency and position				TATA box position
	TTTT	Position	TTCT	Position	CAA	Position	ATCA	Position	CAAT	Position	AAGA	Position	AGCG	Position	
T1	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-124
T2	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-134
T3	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-110
T4	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-58
T5	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-65
T6	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-116
T7	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-93
T8	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-95
T9	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-125
T10	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-126
T11	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-130
T12	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-93
T13	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-100
T14	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-110
T15	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-93
T16	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-111
T17	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-116
T18	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-117
T19	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-128
T20	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-97
T21	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-79
T22	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-123
T23	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-98
T24	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-124

Table 3.9 (cont.) – Overview of group T synthetic core promoters' features: TATA box position and positive and negative motifs frequency and position.

Promoter name	Postive motifs frequency and position										Negative motifs frequency and position				TATA box position
	TTTT	Position	TTCT	Position	CAA	Position	ATCA	Position	CAAT	Position	AAGA	Position	AGCG	Position	
T25	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-141
T26	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-132
T27	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-141
T28	0	-	0	-	0	-	0	-	0	-	0	-	0	-	-115

Table 3.10 – Overview of group A synthetic core promoters' features: TATA box position and positive and negative motifs frequency and position.

Promoter name	Postive motifs frequency and position										Negative motifs frequency and position				TATA box position
	TTTT	Position	TTCT	Position	CAA	Position	ATCA	Position	CAAT	Position	AAGA	Position	AGCG	Position	
A1	0	-	0	-	4	-21; -17; -14; -11	0	-	0	-	1	-66	0	-	-138
A2	0	-	0	-	2	-48; -19	0	-	1	-48	0	-	0	-	-136
A3	0	-	1	-53	0	-	1	-41	0	-	0	-	1	-112	-85
A4	2	-95; -94	1	-68	0	-	0	-	0	-	1	-99	0	-	-105
A5	0	-	1	-80	2	-27; -18	1	-55	1	-27	0	-	0	-	-126
A6	0	-	0	-	3	-46; -39; -15	0	-	0	-	0	-	0	-	-92
A7	0	-	1	-56	2	-36; -32	2	-38; -34	1	-36	2	-144; -97	0	-	-133
A8	0	-	0	-	1	-48	1	-50	0	-	2	-111; -98	1	-91	-118
A9	0	-	0	-	2	-40; -35	1	-63	2	-40; -35	1	-119	0	-	-105
A10	0	-	0	-	1	-27	0	-	0	-	1	-93	0	-	-144
A11	1	-94	0	-	1	-24	1	-56	1	-24	1	-72	0	-	-105
A12	1	-90	0	-	1	-9	0	-	0	-	2	-108; -65	1	-85	-98

Table 3.10 (cont.) – Overview of group A synthetic core promoters' features: TATA box position and positive and negative motifs frequency and position.

Promoter name	Postive motifs frequency and position										Negative motifs frequency and position				TATA box position
	TTTT	Position	TTCT	Position	CAA	Position	ATCA	Position	CAAT	Position	AAGA	Position	AGCG	Position	
A13	0	-	2	-87; -82	3	-50; -25; -14	2	-59; -52	3	-50; -25; -14	2	-101; -74	0	-	-130
A14	3	-60; -59; -58	1	-91	1	-23	0	-	0	-	1	-79	0	-	-143
A15	0	-	0	-	1	-36	1	-38	1	-36	1	-96	0	-	-70
A16	2	-99; -98	0	-	2	-39; -13	0	-	1	-39	2	-110; -62	0	-	-115
A17	0	-	0	-	2	-48; -39	1	-41	2	-48; -39	1	-80	0	-	-107
A18	0	-	1	-92	1	-49	1	-51	1	-49	2	-106; -80	0	-	-140
A19	0	-	0	-	2	-35; -16	0	-	1	-16	2	-123; -88	0	-	-93
A20	0	-	0	-	1	-23	0	-	1	-23	3	-123; -96; -82	0	-	-119
A21	0	-	0	-	1	-49	0	-	0	-	0	-	1	-117	-142
A22	0	-	1	-58	0	-	0	-	0	-	2	-115; -85	0	-	-95
A23	1	-100	0	-	2	-34; -27	0	-	0	-	2	-92; -82	0	-	-126
A24	2	-99; -98	0	-	1	-12	0	-	0	-	1	-136	0	-	-115
A25	0	-	0	-	2	-46; -34	0	-	1	-46	2	-128; -78	1	-106	-120
A26	1	-63	0	-	0	-	0	-	0	-	1	-121	0	-	-126
A27	1	-93	0	-	1	-46	0	-	0	-	0	-	0	-	-146
A28	0	-	1	-60	2	-45; -42	0	-	0	-	1	-69	0	-	-135

Table 3.11 – Blast result for the 10 synthetic promoters with highest activity against the *P. pastoris* CBS 7435 genome. The results consist of the minimum *e-value* for each case, the presence of gaps, match localization (in the *P. pastoris* genome and in the synthetic promoter sequence, as well as the chromosome location) and brief description of the sequence function.

Seq. Name	Min. *E* value	Gaps	Chromosome	Genome location (bp)	Syn. Seq. Location	Syn. Seq. location (from start codon)	Match description
A28	0.083	0	2	497292 to 497314	76 to 98	-74 to -52	Coding sequence of catalytic subunit of (1,3)-β-D-glucan synthase
T28	0.29	1	3	2049828 to 2049792	42 to 77	-108 to -73	130bp upstream of protein coding sequence with putative serine active lipase domain (possible promoter region)
	0.29	0	1	2266300 to 2266316	59 to 75	-91 to -75	Coding sequence of putative protein with unknown function
A27	1	0	3	943316 to 943288	125 to 149	-25 to -1	Coding sequence of hypothetical protein
	1	0	1	212956 to 212931	43 to 68	-107 to -82	Coding region of essential component of the Rix1 complex
T27	0.083	0	4	1369772 to 1369755	123 to 140	-27 to -10	10bp upstream of nucleolar protein coding sequence (possible promoter region)
T26	0.083	0	1	1928476 to 1928498	139 to 117	-33 to -11	Inter gene sequence (between a nucleolar protein and a transcription factor)
T25	0.29	0	4	565027 to 565053	20 to 46	-130 to -104	Coding sequence of hypothetical protein
	0.29	0	2	1165582 to 1165601	33 to 52	-117 to -98	Coding sequence of hypothetical protein
	0.29	0	1	1209655 to 1209634	61 to 82	-89 to -68	Coding sequence of hypothetical protein
M28	0.083	0	4	482161 to 482129	29 to 61	-121 to -89	Coding sequence of subunit of TFIIH and nucleotide excision repair factor 3 complexes
	0.083	1	1	450941 to 450977	149 to 112	-38 to -1	Coding sequence of Flavin adenine dinucleotide (FAD) synthetase
T24	1	0	1	1442424 to 1442406	47 to 65	-103 to -85	Coding sequence of phosphatidylserine decarboxylase of the mitochondrial inner membrane
T23	0.29	0	1	1877072 to 1877098	28 to 54	-122 to -96	Coding sequence of hypothetical protein
T22	1	0	3	619259 to 619274	80 to 95	-70 to -55	Coding sequence of Component of the ESCRT-II complex

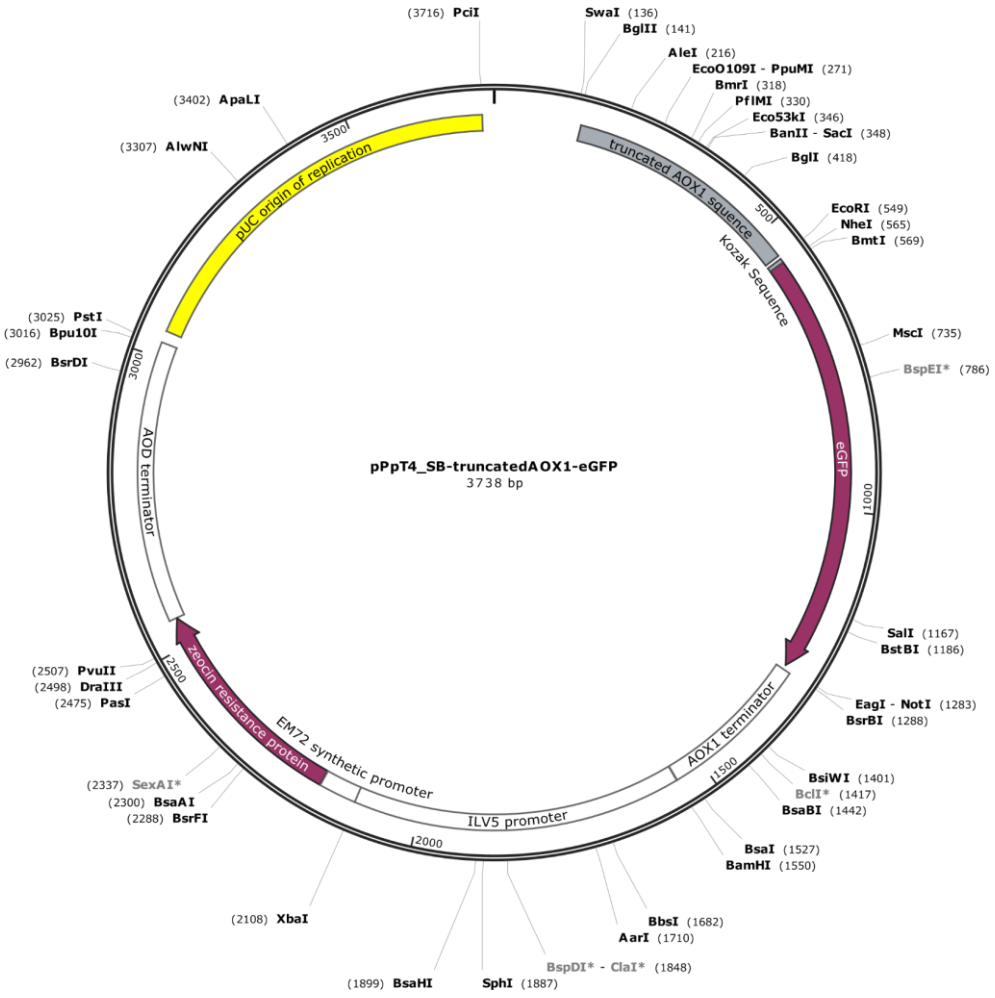

Fig. 3.6 – Map of *P. pastoris/E. coli* shuttle vector pPpT4_SB-truncated*AOX1*-eGFP with main features highlighted: Restriction enzymes, eGFP, zeocin resistance marker, promoters and terminators and origin of replication.

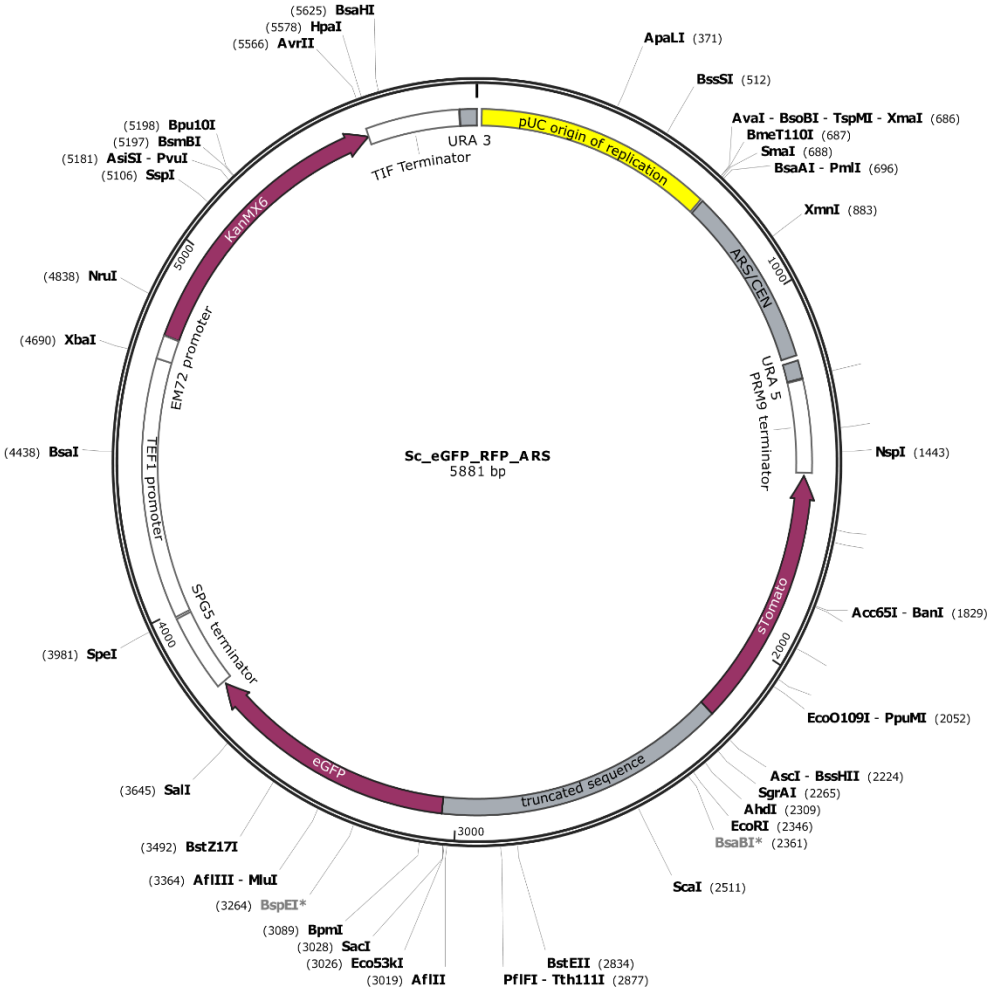

Fig. 3.7 – Map of Sc_eGFP_RFP_ARS with main features highlighted: Restriction enzymes, eGFP, sTomato (RFP), promoters and terminators, kanamycin resistance marker and autonomous replicating sequence (ARS).

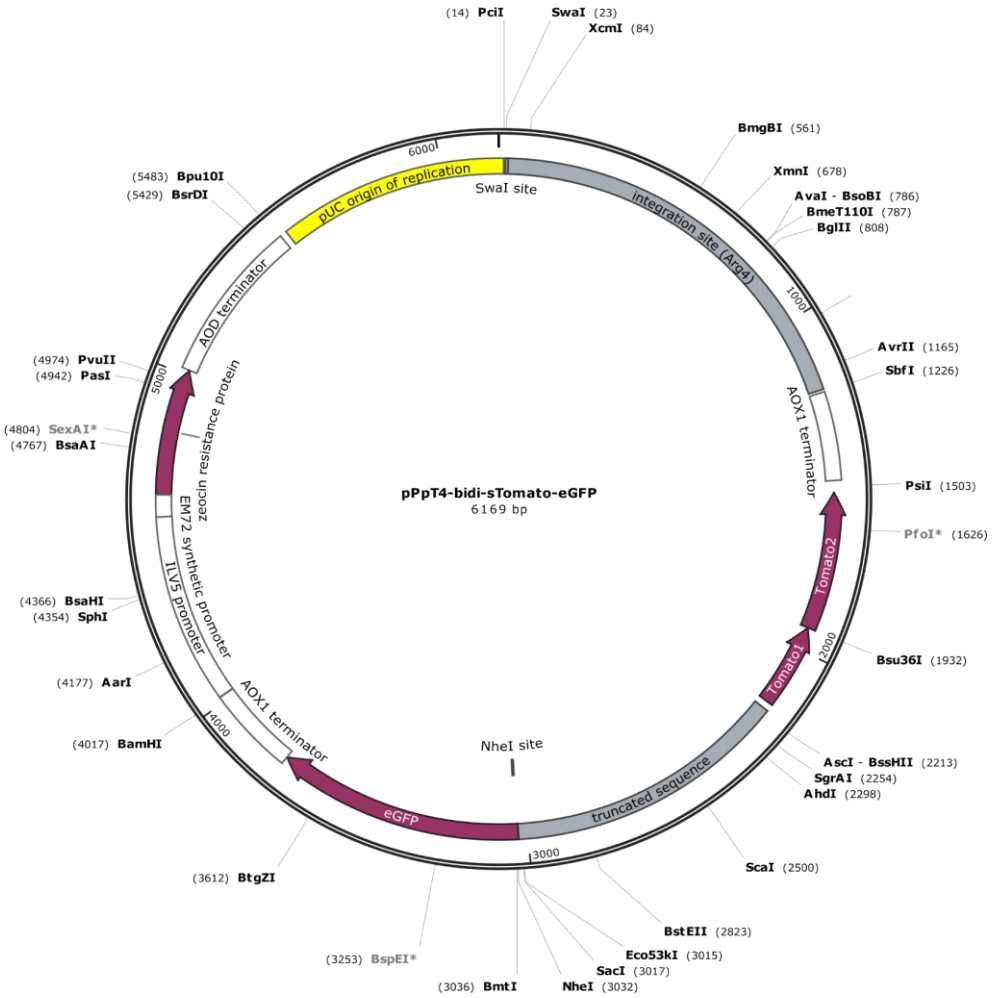

Fig. 3.8 – Map of *P. pastoris/E. coli* shuttle vector pPpT4-bidi-sTomato-eGFP with main features highlighted: Restriction enzymes, eGFP, sTomato (RFP), zeocin resistance marker, promoters and terminators and origin of replication

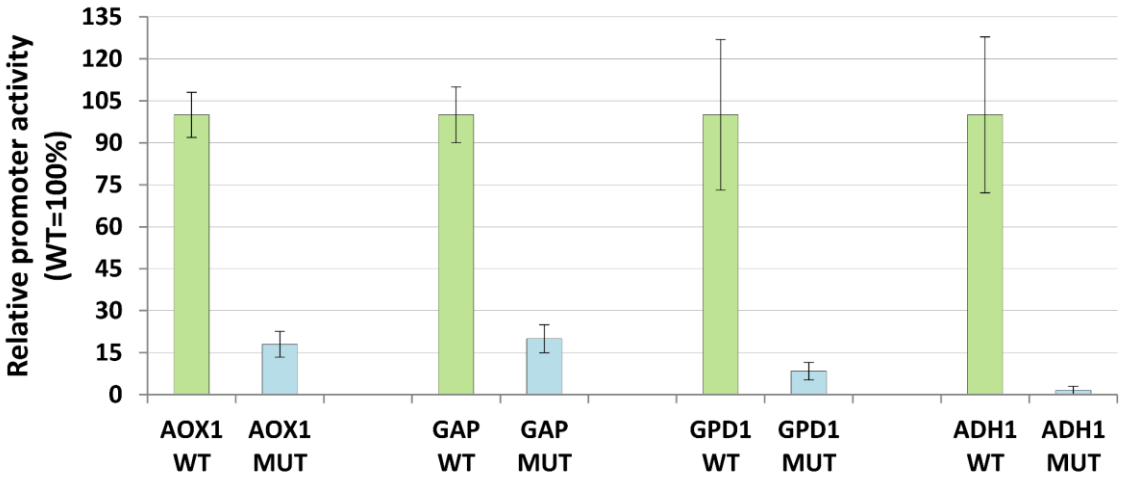

Fig. 3.9 – Expression of the P_{AOX1}, P_{GAP}, P_{ScGPD1} and P_{ScADH1} promoters with (MUT) and without (WT) the mutated TATA box. The TATA box motif in the natural promoter sequence was mutated by replaying three nucleotides of this motif by cytosine. The reporter protein fluorescence of the MUT promoters is compared to the unmodified WT promoter.

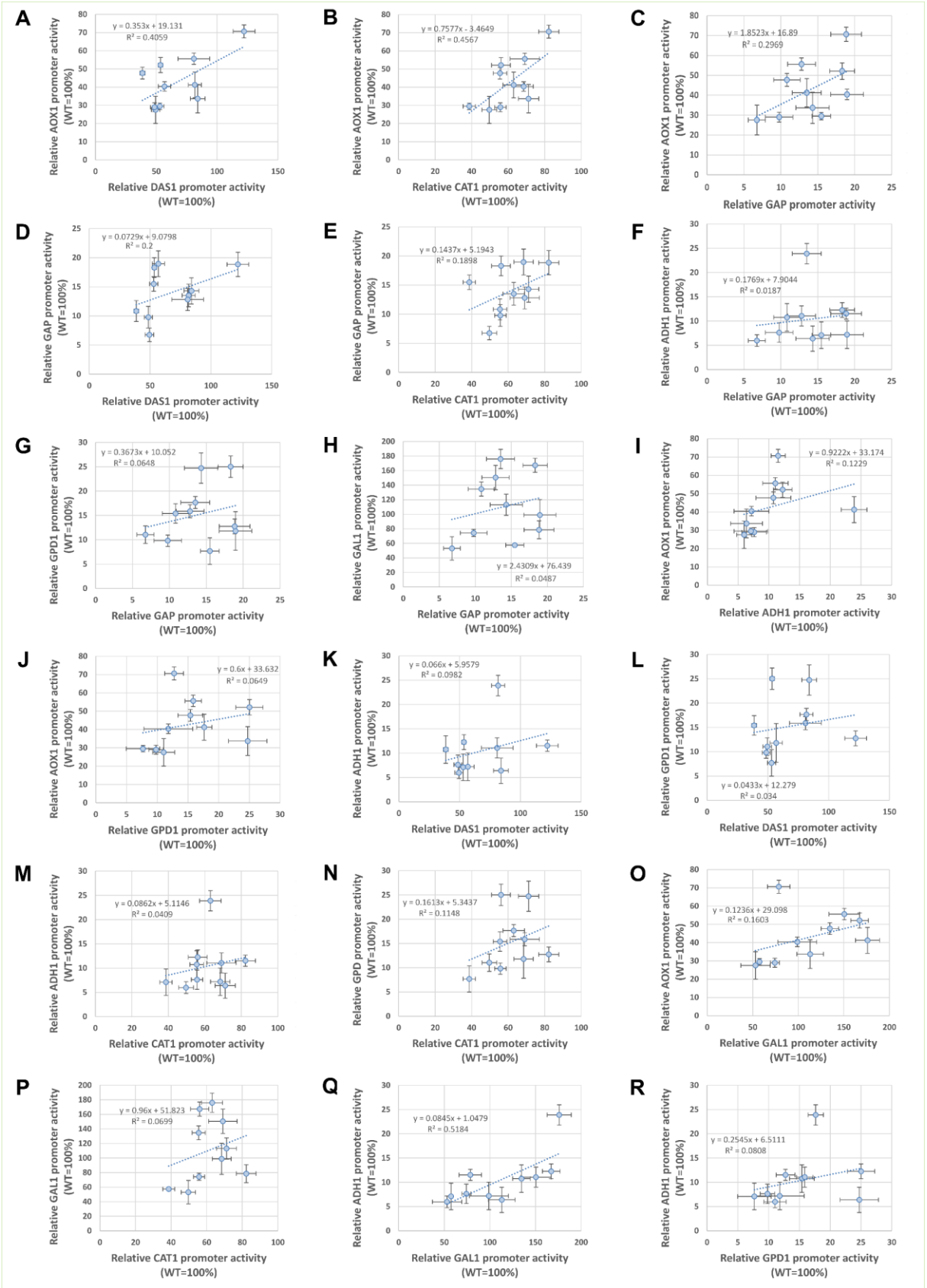

Fig. 3.10 – Additional correlation diagrams for comparisons shown in Fig. 3.5. The heatmap in Fig. 3.5 A was generated from 21 correlation diagrams. Three representative diagrams are shown as panels B-D in Fig. 3.5, the remaining 18 are shown here.

3.7. References

1. Stephanopoulos,G. (2012) Synthetic Biology and Metabolic Engineering. *ACS Synth. Biol.*, **1**, 514–525.

2. Tabor,J.J., Salis,H.M., Simpson,Z.B., Chevalier,A.A., Levskaya,A., Marcotte,E.M., Voigt,C.A. and Ellington,A.D. (2009) A Synthetic Genetic Edge Detection Program. *Cell*, **137**, 1272–1281.

3. Salis,H.M., Mirsky,E. and Voigt,C. (2010) Automated Design of Synthetic Ribosome Binding Sites to Precisely Control Protein Expression. *Nat Biotechnol*, **27**, 946–950.

4. Blazeck,J. and Alper,H.S. (2013) Promoter engineering: Recent advances in controlling transcription at the most fundamental level. *Biotechnol. J.*, **8**, 46–58.

5. Vogl,T., Sturmberger,L., Kickenweiz,T., Wasmayer,R., Schmid,C., Hatzl,A.-M., Gerstmann,M.A., Pitzer,J., Wagner,M., Thallinger,G.G., *et al.* (2016) A Toolbox of Diverse Promoters Related to Methanol Utilization: Functionally Verified Parts for Heterologous Pathway Expression in *Pichia pastoris*. *ACS Synth. Biol.*, **5**, 172–186.

6. Ellis,T., Wang,X. and Collins,J.J. (2009) Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nat. Biotechnol.*, **27**, 465–471.

7. Nevoigt,E., Kohnke,J., Fischer,C.R., Alper,H., Stahl,U. and Stephanopoulos,G. (2006) Engineering of promoter replacement cassettes for fine-tuning of gene expression in *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol.*, **72**, 5266–5273.

8. Hartner,F.S., Ruth,C., Langenegger,D., Johnson,S.N., Hyka,P., Lin-Cereghino,G.P., Lin-Cereghino,J., Kovar,K., Cregg,J.M. and Glieder,A. (2008) Promoter library designed for fine-tuned gene expression in *Pichia pastoris*. *Nucleic Acids Res.*, **36**, 1–15.

9. Alper,H., Fischer,C., Nevoigt,E. and Stephanopoulos,G. (2005) Tuning genetic control through promoter engineering. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 12678–12683.

10. Brown,A.J., Sweeney,B., Mainwaring,D.O. and James,D.C. (2014) Synthetic promoters for CHO cell engineering. *Biotechnol. Bioeng.*, **111**, 1638–1647.

11. Nielsen,J., Larsson,C., van Maris,A. and Pronk,J. (2013) Metabolic engineering of yeast for production of fuels and chemicals. *Curr. Opin. Biotechnol.*, **24**, 398–404.

12. Paddon,C.J., Westfall,P.J., Pitera,D.J., Benjamin,K., Fisher,K., McPhee,D., Leavell,M.D., *et al.* (2013) High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature*, **496**, 528–532.

13. Hong,K.K. and Nielsen,J. (2012) Metabolic engineering of Saccharomyces cerevisiae: a key cell factory platform for future biorefineries. *Cell. Mol. Life Sci.*, **69**, 2671–2690.

14. Liu,L., Redden,H. and Alper,H.S. (2013) Frontiers of yeast metabolic engineering: Diversifying beyond ethanol and *Saccharomyces*. *Curr. Opin. Biotechnol.*, **24**, 1023–1030.

15. Wagner,J.M. and Alper,H.S. (2015) Synthetic biology and molecular genetics in non-conventional yeasts: Current tools and future advances. *Fungal Genet. Biol.*, **89**, 126-136

16. Berg,L., Strand,T.A., Valla,S. and Brautaset,T. (2013) Combinatorial mutagenesis and selection to understand and improve yeast promoters. *Biomed Res. Int.*, **2013**, 1-9.

17. Blazeck,J., Garg,R., Reed,B. and Alper,H.S. (2012) Controlling promoter strength and regulation in *Saccharomyces cerevisiae* using synthetic hybrid promoters. *Biotechnol. Bioeng.*, **109**, 2884–2895.

18. Curran,K., 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.*, **4002**, 1-8.

19. Redden,H. and Alper,H.S. (2015) The development and characterization of synthetic minimal yeast promoters. *Nat. Commun.*, **6**, 7810.

20. Lelli,K.M., Slattery,M. and Mann,R.S. (2012) Disentangling the many layers of eukaryotic transcriptional regulation. *Annu. Rev. Genet.*, **46**, 43–68.

21. Allison,L.A. (2007) Transcription in eukaryotes. In *Fundamental Molecular Biology*.pp. 312–391.

22. Juven-Gershon,T. and Kadonaga,J.T. (2010) Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev. Biol.*, **339**, 225–239.

23. Smale,S.T. and Kadonaga,J.T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem.*, **72**, 449–479.

24. Ruth,C., Zuellig,T., Mellitzer, a., Weis,R., Looser,V., Kovar,K. and Glieder, a. (2010) Variable production windows for porcine trypsinogen employing synthetic inducible promoter variants in Pichia pastoris. *Syst. Synth. Biol.*, **4**, 181–191.

25. Vogl,T., Ruth,C., Pitzer,J., Kickenweiz,T. and Glieder,A. (2014) Synthetic Core Promoters for *Pichia pastoris*. *ACS Synth. Biol.*, **3**, 188–191.

26. Xuan,Y., Zhou,X., Zhang,W., Zhang,X., Song,Z. and Zhang,Y. (2009) An upstream activation sequence controls the expression of *AOX1* gene in *Pichia pastoris*. *FEMS Yeast Res.*, **9**, 1271–1282.

27. Zeevi,D., Lubliner,S., Lotan-Pompan,M., Hodis,E., Vesterman,R., Weinberger,A. and Segal,E. (2014) Molecular dissection of the genetic mechanisms that underlie expression conservation in orthologous yeast ribosomal promoters. *Genome Res.*, **24**, 1991–1999.

28. Lubliner,S., Keren,L. and Segal,E. (2013) Sequence features of yeast and human core promoters that are predictive of maximal promoter activity. *Nucleic Acids Res.*, **41**, 5569–5581.

29. Keren,L., Zackay,O., Lotan-Pompan,M., Barenholz,U., Dekel,E., Sasson,V., Aidelberg,G., Bren,A., Zeevi,D., Weinberger,A., *et al.* (2013) Promoters maintain their relative activity levels under different growth conditions. *Mol. Syst. Biol.*, **9**, 701.

30. Sharon,E., Kalma,Y., Sharp,A., Raveh-Sadka,T., Levo,M., Zeevi,D., Keren,L., Yakhini,Z., Weinberger,A. and Segal,E. (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.*, **30**, 521–530.

31. Hahn,S. and Young,E.T. (2011) Transcriptional regulation in *Saccharomyces cerevisiae*: Transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics*, **189**, 705–736.

32. Sugihara,F., Kasahara,K. and Kokubo,T. (2011) Highly redundant function of multiple AT-rich sequences as core promoter elements in the TATA-less RPS5 promoter of *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **39**, 59–75.

33. Park,D., Morris,A.R., Battenhouse,A. and Iyer,V.R. (2014) Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements. *Nucleic Acids Res.*, **42**, 3736–3749.

34. 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.

35. Bill,R.M. (2014) Playing catch-up with *Escherichia coli:* Using yeast to increase success rates in recombinant protein production experiments. *Front. Microbiol.* ***5***, 1-5.

36. Vogl,T. and Glieder,A. (2013) Regulation of *Pichia pastoris* promoters and its consequences for protein production. *New Biotechnol.*, **30**, 385–404.

37. Khalil,A.S., Lu,T.K., Bashor,C.J., Ramirez,C.L., Pyenson,N.C., Joung,J.K. and Collins,J.J. (2012) A synthetic biology framework for programming eukaryotic transcription functions. *Cell*, **150**, 647–658.

38. Gertz,J., Siggia,E.D. and Cohen,B.A. (2009) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, **457**, 215–218.

39. Mazumder,M. and McMillen,D.R. (2014) Design and characterization of a dual-mode promoter with activation and repression capability for tuning gene expression in yeast. *Nucleic Acids Res.*, **42**, 9514–9522.

40. Murphy,K.F., Balázsi,G. and Collins,J.J. (2007) Combinatorial promoter design for engineering noisy gene expression. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 12726–12731.

41. Blount,B.A., Weenink,T., Vasylechko,S. and Ellis,T. (2012) Rational diversification of a promoter providing fine-tuned expression and orthogonal regulation for synthetic biology. *PLoS One*, **7**, e33279.

42. Teo,W.S. and Chang,M.W. (2014) Development and characterization of AND-gate dynamic controllers with a modular synthetic GAL1 core promoter in *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.*, **111**, 144–151.

43. Qin,X., Qian,J., Yao,G., Zhuang,Y., Zhang,S. and Chu,J. (2011) GAP promoter library for fine-tuning of gene expression in *Pichia pastoris*. *Appl Env. Microbiol*, **77**, 3600–3608.

44. Staley,C.A., Huang,A., Nattestad,M., Oshiro,K.T., Ray,L.E., Mulye,T., Li,Z.H., Le,T., Stephens,J.J., Gomez,S.R., *et al.* (2012) Analysis of the 5' untranslated region (5'UTR) of the alcohol oxidase 1 (*AOX1*) gene in recombinant protein expression in *Pichia pastoris*. *Gene*, **496**, 118–127.

45. Kaplan,N., Moore,I.K., Fondufe-Mittendorf,Y., Gossett,A.J., Tillo,D., Field,Y., LeProust,E.M., Hughes,T.R., Lieb,J.D., Widom,J., *et al.* (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362–366.

46. Geier,M., Fauland,P., Vogl,T. and Glieder,A. (2015) Compact multi-enzyme pathways in *P. pastoris*. *Chem. Commun.*, **51**, 1643–1646.

47. Weninger,A., Hatzl,A., Schmid,C., Vogl,T. and Glieder,A. (2016) Combinatorial optimization of CRISPR/Cas9 expression enables precision genome engineering in the methylotrophic yeast *Pichia pastoris*. *J. Biotechnol.*, **235**, 139–149.

48. McWilliam,H., Li,W., Uludag,M., Squizzato,S., Park,Y.M., Buso,N., Cowley,A.P. and Lopez,R. (2013) Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res.*, **41**, 597–600.

49. Marx,H., Pflugl,S., Mattanovich,D., and Sauer,M. (2016) Synthetic Biology Assisting Metabolic Pathway Engineering. In: Glieder,A., Kubicek,C.P., Mattanovich,D. Wiltschi,B. and Sauer,M., Synthetic Biology. Springer Press, London, pp. 255-280.

50. Dehli,T., Solem,C. and Jensen,P.R. (2012) Tunable promoters in synthetic and systems biology. In: Wang,X., Chen,J., Quinn,P., Reprogramming Microbial Metabolic Pathways, Springer Press, London, pp. 181–201.

51. Seizl,M., Hartmann,H., Hoeg,F., Kurth,F., Martin,D.E., Söding,J. and Cramer,P. (2011) A conserved GA element in TATA-less RNA polymerase II promoters. *PLoS One*, **6**, e27595.

52. Lubliner,S., Regev,I., Lotan-Pompan,M., Edelheit,S., Weinberger,A. and Segal,E. (2015) Core promoter sequence in yeast is a major determinant of expression level. *Genome Res.*, **25**, 1008–1017.

53. Basehoar,A.D., Zanton,S.J. and Pugh,B.F. (2004) Identification and distinct regulation of yeast TATA box-containing genes. *Cell*, **116**, 699–709.

54. Blake,W.J., Balázsi,G., Kohanski,M.A., Isaacs,F.J., Murphy,K.F., Kuang,Y., Cantor,C.R., Walt,D.R. and Collins,J.J. (2006) Phenotypic Consequences of Promoter-Mediated Transcriptional Noise. *Mol. Cell*, **24**, 853–865.

55. Raser,J. and O'Shea,E. (2013) Control of Stochasticity in Eukaryotic Gene Expression. **18**, 1199–1216.

56. Mogno,I., Vallania,F., Mitra,R.D. and Cohen,B. (2010) TATA is a modular component of synthetic promoters. *Genome Res.*, **20**, 1391–1397.

57. Raveh-Sadka,T., Levo,M., Shabi,U., Shany,B., Keren,L., Lotan-Pompan,M., Zeevi,D., Sharon,E., Weinberger,A. and Segal,E. (2012) Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat Genet*, **44**, 743–750.

58. Lam,F.H., Steger,D.J. and O'Shea,E.K. (2008) Chromatin decouples promoter threshold from dynamic range. *Nature*, **453**, 246–250.

59. Li,X.Y., Bhaumik,S.R., Zhu,X., Li,L., Shen,W.-C., Dixit,B.L. and Green,M.R. (2002) Selective recruitment of TAFs by yeast upstream activating sequences. Implications for eukaryotic promoter structure. *Curr. Biol.*, **12**, 1240–1244.

60. Sturmberger,L., Chappell,T., Geier,M., Krainer,F., Day,K.J., Vide,U., Trstenjak,S., Schiefer, A., Richardson,T., Soriaga,L., Darnhofer,B., Birner-Gruenberger,R., Glick,B.S., Tolstorukov,I., Cregg,J., Madden,K., and Glieder,A. (2016) Refined *Pichia pastoris* reference genome sequence. *J. Biotechnol*, **235**, 121-131.

61. Winston,F., Dollard,C. and Ricupero-hovasse,S.L. (1995) Construction of a set of convenient *Saccharomyces cerevisiae* strains that are isogenic to S288C. *Yeast*, **11**, 53–55.

62. Lin-Cereghino,J., Wong,W.W., Xiong,S., Giang,W., Luong,L.T., Vu,J., Johnson,S.D. and Lin-Cereghino,G.P. (2005) Condensed protocol for competent cell preparation and transformation of the methylotrophic yeast *Pichia pastoris*. *Biotechniques*, **38**, 44-48.

63. Weis,R., Luiten,R., Skranc,W., Schwab,H., Wubbolts,M. and Glieder,A. (2004) Reliable high-throughput screening with *Pichia pastoris* by limiting yeast cell death phenomena. *FEMS Yeast Res.*, **5**, 179–89.

64. Amberg,D., Burke,D. and Strathern,J. (2005) Methods in Yeast Genetics: A Cold Spring Harbor Laboratory Course Manual, Cold Spring Harbor Laboratory Press.

65. Xi,L., Fondufe-Mittendorf,Y., Xia,L., Flatow,J., Widom,J. and Wang,J.P. (2010) Predicting nucleosome positioning using a duration Hidden Markov Model. *BMC Bioinformatics*, **11**, 1-9.

Chapter 4

Hybrid semiparametric sequence-activity modeling: the case of *E. coli* synthetic Ribosome Binding Sequences

4.1. Abstract

Quantitative Sequence-Activity Modeling (QSAM) is essential for engineering synthetic DNA sequences with desired biologic activity. However, a critical limitation is the combinatorial nature of nucleotide sequences, which tremendously increase the experimental effort for reliable model--based design. Here we explore the use of hybrid semiparametric systems for QSAM in the context of data sparsity, using the 5' Untranslated Region (5'UTR) in *Escherichia coli* as illustrative example. We compare Thermodynamic Modeling (TM), Partial Least Squares (PLS) and hybrid parallel TM+PLS modeling. Our results show that TM clearly outperforms empirical PLS in predicting high activity RNA sequences. However, hybrid TM+PLS modeling is shown to be significantly better than standalone TM or PLS. In particular, hybrid TM+PLS is shown to be 33% and 52% more accurate than standalone TM in data calibration and extrapolation respectively. All in all, this study points out that hybrid modeling is a powerful methodology for QSAM, potentially enabling more effective design of synthetic DNA sequences.

Keywords

Hybrid semiparametric systems · Thermodynamic modeling · N-way Partial Least Squares (N-PLS) · Quantitative Sequence-Activity Modeling (QSAM) · Synthetic Biology · *Escherichia coli* · Ribosome Binding Site (RBS)

4.2. Introduction

Mathematical modeling is a fundamental tool for the engineering of standard biological parts (SBPs) in Synthetic Biology (1, 2). A particular class of problems deals with the DNA sequence design of synthetic SBPs such as promoters, riboswitches, Ribosome Binding Sequences (RBSs) and other DNA/RNA parts. Given the combinatorial nature of nucleotide sequences, exploring the whole design space by random mutation is experimentally impractical. As illustrative example, the 5' Untranslated Region (5'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 only the 10bp that precede the start codon. Even using high-throughput techniques, only 0.2% out of 10^6 possible sequences were screened. Rational design, aided by mathematical models, is thus essential to save time and resources while increasing design efficiency in such large design problems.

Mechanistic modeling is the method of choice for SBPs design (4) with several successful examples published. Brewster *et al.* (5) developed a thermodynamic transcription initiation model that focused on the -10 and -35 promoter regions of *Escherichia coli* and its affinity to RNA polymerase. This model enabled the design of synthetic promoters, with three orders of magnitude difference, in their mean expression, in comparison to natural promoters. Other thermodynamic models were developed to describe translation in *E. coli*. Namely, Salis *et al.* (6) developed a predictive method for RBSs, with the objective of controlling protein translation, assuming that

translation initiation is the limiting step. This model is based on the free Gibbs energy associated with the formation of the mRNA-ribosome complex, determined from five molecular interactions. Later on, Na *et al.* (7) simplified the model by Salis *et al.* (6), focusing on three molecular interactions only, and applied it to the design of 22 *luxR* mRNA sequences with desired translation efficiencies. 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. These three models were based on previous experimental evidence of a relationship between the absence of secondary structures near the RBS and start codon and high protein production rates (*e. g.* (9)). Other studies focused on translation elongation. For instance, Racle *et al.* (10) used a sequence specific continuous model of translation and a parameter sensitivity analysis to determine the optimal codon sequence for heterologous protein expression. Furthermore, mechanistic and dynamic thermodynamic models were applied to quantitatively reprogram gene expression using RNA devices, such as ribozymes and aptazymes. Carothers *et al.* were able to design 28 RNA devices in *E. coli* with a high correlation between predicted and measured gene expression levels (11).

As alternative to mechanistic modeling, empirical data-oriented methods have been reported in the literature for Quantitative Sequence-Activity Modeling (QSAM) problems. There are two main classes of problems: classification of DNA sequences and regression to describe biologic activity as function of the respective DNA sequence. González-Díaz *et al.* (12) used Markov molecular negentropies to describe the secondary structure of putative RNA molecules and used such prediction to identify mycobacterial promoters. Tavares *et al.* (13) 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.* (14) used a mixture of Gaussian models to predict translation initiation sites when there were several start codons in particular mRNAs 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 (15). Artificial neural networks have been used for both classification and regression problems. In (16), an encoding method based on DNA helical parameters was adopted to predict DNA curvature and transcription rate in *E. coli*. Partial Least Squares (PLS) and Support Vector Machines (SVM) were compared. Jonsson *et al.* used PLS with binary encoding to design two synthetic *E. coli* promoters (17). Liang *et al.* compared the performance of SVM and PLS to predict transcription rate of *E. coli* promoters (18). Ran *et al.* (19) 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 QSAM studies, modeling either follows a parametric paradigm, where models have a fixed structure inspired by knowledge and have parameters with physical interpretation, or follows a nonparametric approach, where model structure is derived exclusively from data and parameters have no physical meaning. In this chapter, we explore the combination of both approaches in hybrid semiparametric systems. The main advantage of the semiparametric over the parametric or nonparametric frameworks lies in that it broadens the knowledge base that can be

used 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 (20). Applications to Synthetic Biology are still largely absent in the literature. However, while designing SBPs, it is unlikely that all relevant processes can be fully described by a mechanistic (parametric) approach due to lack of knowledge. Purely empirical (nonparametric) modeling approaches are limited by the availability of sufficient experimental data within a very large design space. Opting for either one of the frameworks will invariably promote reductionism. On the contrary, the "complementary" use of both types of knowledge permits to expand the model towards more comprehensive descriptions of the biological system at hand. In this chapter, we study this hybrid semiparametric modeling concept using as illustrative case study a QSAM of the 5'UTR sequence in *E. coli*.

4.3. Materials and methods

4.3.1. RNA sequences and protein expression data

In this work we have adopted the data set published by Salis *et al.* (6), where 5'UTR sequences were designed and tested in *E. coli* DH10B. For that purpose, a vector composed by a chloramphenicol resistance gene, a σ^{70} constitutive promoter, a RBS sequence followed by the *mRFP1* fluorescence reporter protein gene was used. In total, 132 mRNA sequences were designed and cloned in the referred vector, previously digested with XbaI and SacI to replace the existing RBS. *E. coli* was transformed with the resulting plasmid. Cells were cultured in 96-well plates with LB and chloramphenicol. Red fluorescence protein levels were acquired by measuring fluorescence in a flow cytometer. The final data set comprises 132 mRNA sequences and respective protein fluorescence levels. For more details, the reader is referred to (6).

The data were divided into a model identification partition and a test partition. The model identification partition served to identify model structure and to estimate the underlying parameter values. The test partition served to assess the model predictive power. Three different 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 was repeated 100 times yielding 100 different models to eliminate a possible data sampling effect.

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.

Partition R applies uniform data sampling for model identification. Partitions E33 and E67 are obviously much more demanding since the model is calibrated with low protein titer sequences and then asked to extrapolate the high protein titer sequences. However, this is a rather tough testing criterion that "measures" the ability of the model to design improved RNA sequences with higher protein expression levels.

4.3.2. Thermodynamic model

The equilibrium Thermodynamic Model (TM) proposed by Salis *et al.* (6) is one of the modules of the hybrid model (Fig. 4.1). 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 a final state, where the mRNA binds to the ribosome in the form of the 30S pre-initiation complex. The ΔG_{TOT} accounts for 5 terms:

$$\Delta G_{TOT} = \Delta G_{mRNA:rRIB} + \Delta G_{START} + \Delta G_{SPACING} - \Delta G_{STANDBY} - \Delta G_{mRNA} \quad \text{(Eq. 4.1)}$$

ΔG_{mRNA} is the Gibbs free energy value of the mRNA molecule when it is not interacting with any other molecule, thus it can 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 protein start codon. After unfolding, mRNA hybridizes with 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 (interaction too far away or too close to the start codon). ΔG_{START} accounts for the interaction between the mRNA 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 these Gibbs free energies (ΔG_{START}, $\Delta G_{mRNA:rRIB}$, $\Delta G_{STANDBY}$ and ΔG_{mRNA}) is direct (no 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 (21–23).

Finally, the predicted protein expression level (P_{TM}) is a function of the respective mRNA secondary structures, represented by the difference in Gibbs free energy (ΔG_{TOT}), as follows:

$$P_{TM} = \alpha\, t\, e^{-\beta \Delta G_{TOT}} \quad \text{(Eq. 4.2)}$$

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

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. The slope corresponds to the Boltzmann constant, β, while the y-axis intercept corresponds to ln(α t).

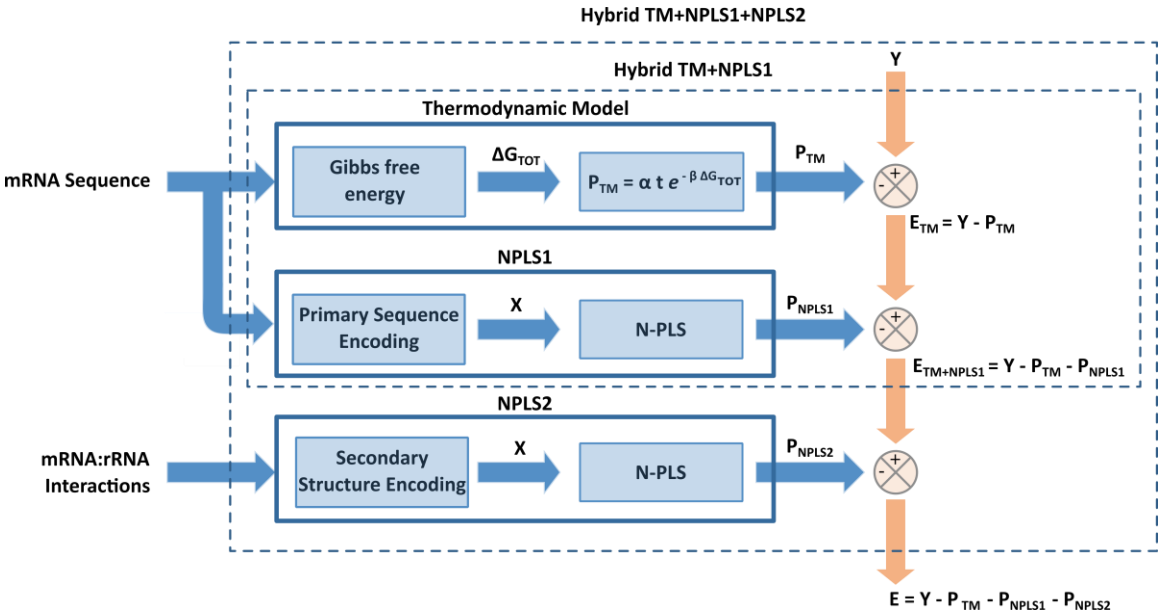

Fig. 4.1 – Parallel hybrid model structure describing protein expression (Y) as function of mRNA sequence. The first module is the thermodynamic model (TM) that predicts protein expression as function of ΔG_{TOT}. The second module (NPLS1) is an N-PLS model that runs in tandem with the TM and extracts 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.

4.3.3. Nucleotide sequences encoding

To translate the symbolic nucleotide sequences into a numerical representation, six different encoding methods 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(π/6), -sin(π/3)), Cytosine (sin(π/6), sin(π/3)), Guanine (sin(π/3), -sin(π/6)), Uracil (sin(π/3), sin(π/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* (24), *Encoding 5* (17) and *Encoding 6* (18).

4.3.4. N-PLS models

N-way Partial Least Squares (N-PLS) is a well-known multivariate regression method with data factorization. In this procedure a target matrix, **Y**, is linearly regressed against an input (regressor) matrix **X** of many possibly collinear variables (25). 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 3D **X** (*I* × *J* × *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**:

$$\mathbf{X}_{ijk} = \mathbf{t}_i \, \mathbf{w}_j^J \, \mathbf{w}_k^K + \mathbf{E} \qquad \text{Eq. 4.3)}$$

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:

$$\max_{\mathbf{w}^J \mathbf{w}^K} \left(\sum_{i=1}^{I} \mathbf{t}_i \mathbf{y}_i \middle| \mathbf{t} = \sum_{j=1}^{J} \sum_{k=1}^{K} \mathbf{x}_{ijk} \mathbf{w}_j^J \mathbf{w}_k^K \right) \qquad \text{(Eq. 4.4)}$$

Modelling multi-dimensional datasets by bilinear PLS implies that the 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 (25). In this work, we used the N-PLS MATLAB implementation described in (26). Two N-PLS models were developed, as described next.

4.3.4.1. NPLS1: primary structure N-PLS model

NPLS1 is a 3-way N-PLS model that describes protein titer as function of primary mRNA sequence. The mRNA sequences were trimmed to 70bp (35bp upstream and downstream of the start codon – same sequences as the ones used to calculate ΔG_{mRNA} for the TM). Since some of the mRNA molecules are shorter (less than 35bp upstream of the start codon), they were first aligned by their start codon and filled with blank spaces. Then one of the previously described encoding methods was applied. The encoding resulted in a 3-dimensiononal **X** matrix (*np* x *nb* x

ne), with *np*=132 the number of mRNA sequences, *nb*=70 the maximum sequence length (in base pairs) and *ne=4* or *ne=2* the number of values representing a single nucleotide (depending on the encoding used). **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 normalized by first applying the natural logarithm (note that the natural logarithm also appears in Eq. 4.2, providing a rational for this type of normalization) and then by autoscaling. The normalized **X** and **Y** were subject to the N-PLS regression using the MATLAB implementation described in (26). The optimal number of latent variables was determined by the leave-one-out method (27). The number of NPLS1 parameters, *npar*, is equal to the optimal number of latent variables, Fac_{NPLS1}.

4.3.4.2. NPLS2: mRNA:rRNA interactions N-PLS model

NPLS2 models describe protein expression level as function of the mRNA standby sequence. The mRNA standby sequences (used before to calculate the $\Delta G_{STANDBY}$), 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 (21–23), in order to determine the standby sequences for a given mRNA molecule. These sequences were organized in a 3D **X** matrix (*np* x *nb* x (*ns* x *ne)*) with *np*=132 the number of mRNA molecules, *nb*=20 the maximum standby sequence length (in base pairs), and *ns* x *ne* (3rd 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.,* a 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 and the protein expression data **Y** was normalized by first applying the natural logarithm and then by autoscaling. The normalized **X** and **Y** were subject to the N-PLS regression using MATLAB implementation described in (26). The optimal number of latent variables was determined by the leave-one-out method (27). The number of NPLS2 parameters, *npar*, is equal to the optimal number of latent variables, Fac_{NPLS2}.

4.3.5. Hybrid semiparametric model

The hybrid model structure is represented schematically in Fig. 4.1. It consists of a modular structure with input data segmentation. The first module is the TM. It predicts protein expression as function of the thermodynamic parameter ΔG_{TOT}. The second module, NPLS1, is a 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 a N-PLS model that runs in tandem with TM+NPLS1 and extracts information from the TM+NPLS1 residuals as function of mRNA-rRNA possible interactions. The output vector **Y** (natural logarithm of the measured protein expression data – $\ln(P_{MES})$) was again normalized by autoscalling. The hybrid model decomposes **Y** in 4 terms:

$$\mathbf{Y} = \ln(\mathbf{P}_{\mathrm{TM}}) + \mathbf{P}_{\mathrm{NPLS1}} + \mathbf{P}_{\mathrm{NPLS2}} + \mathbf{E} \quad \text{(Eq. 4.5)}$$

The first 3 terms represent the contribution of the 3 modules TM, NPLS1 and NPLS2, respectively, to the prediction of **Y**. The vector **E** is the final hybrid model residuals. The model identification was performed in two steps as follows:

Step 1: Identification of the hybrid TM+NPLS1 structure. Firstly, the TM model is fitted to the model identification data set (see above). Then the TM residuals, $\mathbf{E}_{\mathrm{TM}}$, are calculated:

$$\mathbf{E}_{\mathrm{TM}} = \mathbf{Y} - \ln(\mathbf{P}_{\mathrm{TM}}) \quad \text{(Eq. 4.6)}$$

Finally, NPLS1 is identified following the same method described above for the standalone NPLS1, except that the target output is $\mathbf{E}_{\mathrm{TM}}$ rather than **Y**. Also, the optimal number of latent variables (Fac_{NPLS1}) was not determined by leave-one-out. Rather, the model identification dataset 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 $\mathbf{E}_{\mathrm{TM}}$). This ensures the selection of the optimal number of latent variables that maximizes predictive power of the TM model residuals.

Step 2: Identification of the hybrid TM+NPLS1+NPLS2 structure. Firstly, the predicted protein expression level by the previously identified NPLS1, $\mathbf{P}_{\mathrm{NPLS1}}$, is calculated. Then the hybrid TM+NPLS1 residuals are calculated:

$$\mathbf{E}_{\mathrm{TM+NPLS1}} = \mathbf{Y} - \ln(\mathbf{P}_{\mathrm{TM}}) - \mathbf{P}_{\mathrm{NPLS1}} \quad \text{(Eq. 4.7)}$$

Finally, NPLS2 is identified following the same method previously described for standalone NPLS2, except that the target output is $\mathbf{E}_{\mathrm{TM+NPLS1}}$ (instead of **Y**) and that the optimal number of latent variables was determined as in step 1.

4.3.6. Model performance criteria

Three different metrics were employed for model performance assessment, namely the Mean Squared Error (*MSE*) (Eq. 4.8), Explained Variance (*Var*., %), (Eq. 4.9) and Akaike Information Criterion (*AIC*), (Eq. 4.10):

$$MSE = \frac{1}{n}\mathbf{E}^T\mathbf{E} \quad \text{(Eq. 4.8)}$$

$$Var(\%) = 100\left(1 - \frac{\mathbf{E}^T\mathbf{E}}{\mathbf{Y}^T\mathbf{Y}}\right) \quad \text{(Eq. 4.9)}$$

$$AIC = 2 \times npar + np\ln(MSE) \quad \text{(Eq. 4.10)}$$

with *np* the number of data points, **E** a vector of model residuals, *npar* the number of model parameters given by:

$$npar = 2 + Fac_{NPLS1} + Fac_{NPLS2} \quad \text{(Eq. 4.11)}$$

AIC introduces a penalty for overparameterization and is commonly used to discriminate between empirical model candidates, enabling a selection of a more parsimonious model (27).

4.4. Results and discussion

4.4.1. Thermodynamic model

4.4.1.1. Determination of Gibbs free energy and model fitting

Fig. 4.2 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 pre-initiation complex, accounting for 5 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 free Gibbs energy parameters for each of the 132 mRNA sequences are available as Table 4.4.

The TM model (Eq. 4.2) was fitted to the calculated ΔG_{TOT} and measured fluorescence data, adopting the uniform data partitioning strategy (partition R) previously described. 67% of data points are randomly selected for fitting (Eq. 4.2). The remaining 33% of data points are used to assess predictive power. The procedure is repeated 100 times to eliminate data sampling bias. The results are shown in Table 4.1 (1st row). The average Boltzmann constant among the 100 different trials was 0.37±0.034mol/kcal, which is slightly lower than the value reported by Salis *et al.* (6) (0.45±0.05mol/kcal for two different data partitions). The average explained variance of measured protein fluorescence was 70.59% for the identification data set and 68.47% for the test data set, showing that the prediction accuracy is comparable to the calibration accuracy. Even so, the model exhibits a bias in the sense that it systematically underpredicts the highest protein titer sequences, *e. g.* -36.57% for the top 5 protein expression sequences (Table 4.1, 1st row).

4.4.1.2. Effect of data sparsity on predictive power

Data sparsity is critical in QSAM because the design space is very large and typically only a small set of sequences are experimentally screened. In the problem studied here, the length of the sequence considered for design purposes is up to 35bp (mRNA sequence upstream of start codon) with 4^{35} possible combinations, of which only 132 sequences were experimentally screened. QSAM coupled with rational design are thus essential for the effective discovery of better performing sequences. To assess the ability of the TM model to extrapolate the high activity mRNA

sequences, the E33 and E67 data partitioning scenarios were studied. The overall results are shown in Table 4.1 (4th and 7th rows).

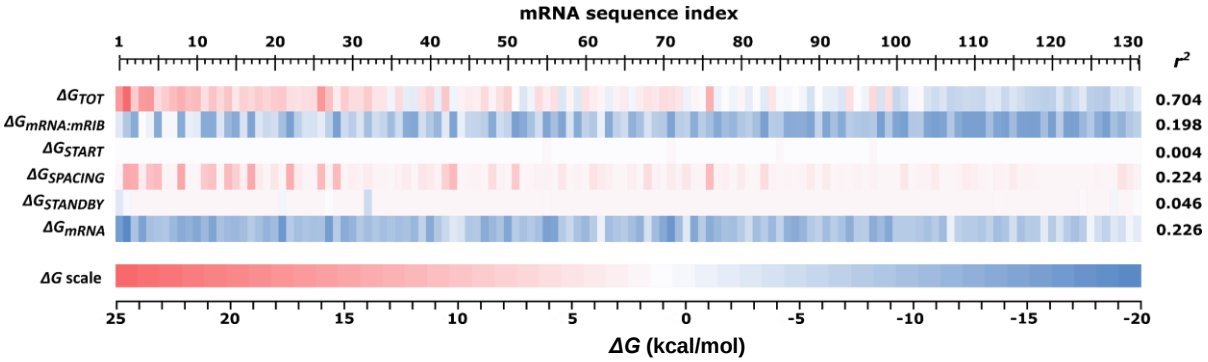

Fig. 4.2 – Heatmap representing the free Gibbs energy for each of the 132 RNA sequences sorted from high to low protein fluorescence values. The six rows refer to ΔG_{TOT} and to the individual $\Delta G_{mRNA:rRIB}$, ΔG_{START}, $\Delta G_{SPACING}$, $\Delta G_{STANDBY}$ and ΔG_{mRNA}. The correlation coefficient (r^2) on the right side denotes the correlation of measured protein fluorescence in relation to calculated free Gibbs energy. The three ΔG 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). Values represented in this figure are available in Table 4.4**.**

Table 4.1 – Explained variance, *MSE* and *AIC* values for identification and test partition in different data partitioning conditions (R, E33 and E66). The number of latent variables (*Fac*) and Boltzman factor (β) for N-PLS and thermodynamic models, respectively is also presented.

Partition	Model	Identification			Test			Relative Error	
		% *Var*	*MSE*	*AIC*	% *Var*	*MSE*	*AIC*	Top 5%	Top 1%
PARTITION R	TM[1]	70.59	0.29	-	68.47	0.31	-	-36.57	-61.17
	NPLS1[2]	71.20	0.28	-94.74	29.02	0.72	-11.57	-45.00	-21.82
	NPLS2[3]	29.43	0.71	-24.15	10.58	0.84	-1.45	-80.94	-134.60
PARTITION E33	TM[4]	68.14	0.30	-	56.59	0.49	-	-62.06	-81.47
	NPLS1[5]	63.30	0.34	-87.85	51.32	0.55	-21.31	-56.62	3.53
	NPLS2[6]	15.5	0.78	-19.29	30.43	0.78	-9.24	-85.66	-129.12
PARTITION E67	TM[7]	85.86	0.25	-	-63.51	0.98	-	-98.32	-112.83
	NPLS1[8]	90.23	0.17	-71.03	-147.00	1.48	40.65	-39.61	21.50
	NPLS2[9]	10.50	1.59	22.41	7.46	0.56	-49.74	-97.22	-101.83

[1] β=0.37, [2] With encoding 5 and *Fac*=3, [3] With encoding 4 and *Fac*=3, [4] β=0.30, [5] With encoding 3 and *Fac*=3, [6] With encoding 6 and *Fac*=1, [7] β=0.25, [8] With encoding 5 and *Fac*=3, [9] With encoding 6 and *Fac*=1

In the case of partition E33 (extrapolation of the 33% best sequences), the explained variance of the identification data set is 68.14%, but the explained variance of the test data set decreases to 56.59%. 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 explained variance of the identification data set improves to 85.86% while the explained variance of the test data decreases to -63.51% and the top 5 sequences are systematically underpredicted by -98.32%. The model is clearly overfitting the identification data set and unable to predict the test data set.

All in all, these results suggest that the ability of the TM to predict the activity of better performing sequences beyond the domain of experimental validation is significantly affected in a context of data sparsity.

Fig. 4.3 A plots predicted over measured protein expression for the data partition E33. Fig. 4.3 B and C show the residuals distribution for the identification and test data sets respectively. It may be concluded, 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 (Fig. 4.3 C) that model predictions are largely biased in the test partition (-0.62 mean and 0.32 standard deviation) in the sense of underprediction.

4.4.2. N-PLS regression as a QSAM tool

We have assessed N-PLS regression as a standalone QSAM tool and compared it to the TM model. Firstly, we have studied N-PLS regression of protein fluorescence as function of primary mRNA sequence (NPLS1). The mRNA encoding method is critical for the modeling results, thus we compared 6 different encodings. The data partitioning is also critical, especially in the context of data sparsity. We have assessed the same 3 data partitioning scenarios as before for the TM model. The results are shown in Table 4.1 for the best encodings (2nd, 5th and 8th rows). Table 4.5 to Table 4.7 provide detailed results for all encodings and different partitions. The encoding method is clearly an important factor. Encodings 1, 3, and 5 produce significantly better results than encodings 2, 4 and 6, in this conditions.

The second row of Table 4.1 summarizes the results for partition R best encoding, reporting the average performance from 100 different NPLS1 models derived from 100 randomly selected data sets (to remove data sampling bias). NPLS1 describes the identification partition with explained variance 71.20%, very similar to the thermodynamic model (70.59%). The description of the test partition is, however, much worse for the NPLS1 model (29.02% explained variance) when compared to the TM model (68.47%). In the case of partition E33 (Table 4.1 – 5th row and Table 4.6), NPLS1 shows a slightly worse but comparable performance to the TM model in terms of data fitting flexibility (68.14% explained variance for TM against 63.30% for NPLS1 in the identification partition) and of predictive power (56.59% explained variance for TM against 51.32% for NPLS1 in the test partition). In the case of partition E67 (Table 4.1 – 8th row and Table 4.7), the NPLS1 model improves the fitting power (85.86% explained variance for TM against 84.20% to 97.15% for NPLS1 in the identification partition) but degrades significantly its predictive power in relation to the TM model (-63.51% explained variance for TM against -555.74% to -147.00% for NPLS1 in the test partition).

A similar analysis was performed with all the mRNA-rRNA interactions as input to the N-PLS (NPLS2 model). Here the N-PLS framework is applied to model protein fluorescence as function of the standby sequence upstream of the mRNA-rRNA interaction locus. The results are shown in Table 4.1 (3rd, 6th and 9th rows) and Table 4.8 to Table 4.10 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 distinctive result is the much lower explained variances for the NPLS2 model in relation to the NPLS1 and to the thermodynamic model. This is not entirely surprising since the input to the NPLS2 model is restricted to mRNA-rRNA interactions, thus incomplete.

In the TM model, the standby sequence information content is given by the $\Delta G_{STANDBY}$ term, which is zero for a large number of sequences (Fig. 4.2 and Table 4.4). However, it strikes that in the NPLS2 model the explained variance of the identification and testing partitions are comparable. This suggests that N-PLS is extracting meaningful, though partial, information from the standby sequence that is relevant for describing reporter protein fluorescence. Since, this information is not effectively captured by $\Delta G_{STANDBY}$ there might be a window of opportunity to improve the TM model through a hybrid modeling framework.

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 a key factor but no clear rule could be identified regarding the best approach, where, apparently, the data partitioning influences which encoding performs best. As general trend, N-PLS tends to be more flexible in data calibration but clearly inferior to the TM model in terms of extrapolation. Thus, we conclude the TM to be a more powerful QSAM methodology than N-PLS when compared as standalone methods.

4.4.3. Hybrid semiparametric QSAM

The hybrid structure schematized in Fig. 4.1 was investigated as QSAM methodology. The key objective was to examine whether the joint use of the TM and N-PLS in a hybrid structure is advantageous in relation to the standalone TM, NPLS1 and NPLS2 models.

The design of a hybrid model structure must consider the different types of knowledge available (28, 29). For the present problem 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. Another key rule when developing a hybrid model is that reliable mechanistic knowledge has priority over heuristic or empirical knowledge (30). We have thus adopted a hybrid structure that gives full priority to the TM model to describe protein expression observations. This is in agreement with the results above, where the TM clearly outperforms NPLS1 and NPLS2 when applied as standalone models. Therefore, the hybrid models studied here may be seen as improvements of a core TM component. This was performed in two sequential steps as described below. First we have studied the improvement of TM by the inclusion of NPLS1 module (hybrid TM+NPLS1 structure) followed by the inclusion of the NPLS2 module (hybrid TM+NPLS1+NPLS2 structure).

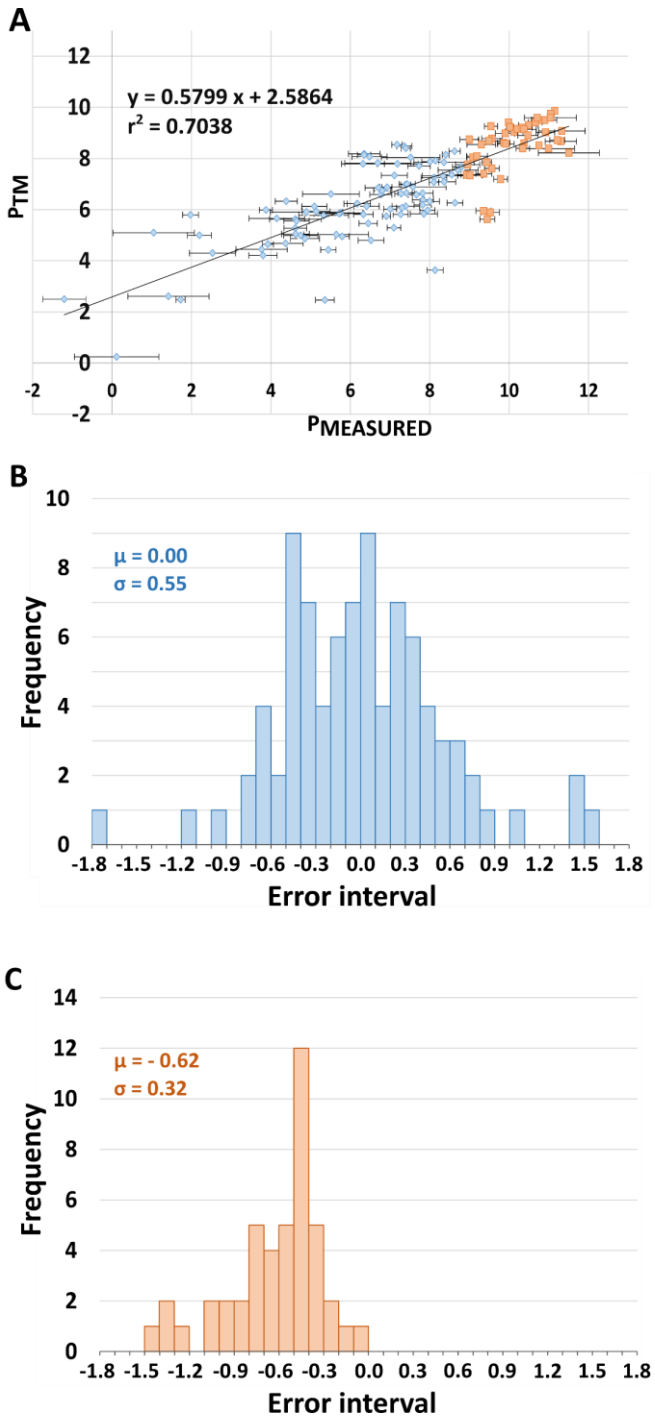

Fig. 4.3 – Thermodynamic modeling (TM) results for partition E33. 67% of mRNA sequences with lowest measured protein fluorescence are used for model calibration (identification data set) while 33% of mRNA sequences with highest measured protein fluorescence are used for model predictive power assessment (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. Data points are normalized in panels B and C.

4.4.3.1. Hybrid TM+NPLS1 model

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

The procedure used to identify the TM as module embedded in the hybrid structure is the same as for the standalone TM model (see methods section). Thus, the results of this first identification step were previously discussed and summarized in Table 4.1**,** Fig. 4.2 and Fig. 4.3.

The results of NPS1 identification in tandem with the TM model (see methods section) are summarized in Table 4.2 and Table 4.3 for partitions E33 and E67, respectively. Detailed results for all encodings are given in Table 4.11 and Table 4.12. Again, the results are highly sensitive to the encoding method, where encoding 4 stands out as the best performing method. Noteworthy, encoding 4, together with encoding 6, had the worst predictive performance with standalone NPLS1 for partition E33 (Table 4.6).

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% of in the test data set (Table 4.2). 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 this model. 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 data noise. In the case of E67, the improvement is also clear but less expressive (Table 4.3). For this reason, in what follows we focus our analysis in partition E33.

Table 4.2 – Model performance criteria for TM and hybrid models for (partition E33)

Model		Identification			Test			Relative Error	
		% Var	*MSE*	*AIC*	% Var	*MSE*	*AIC*	Top 5%	Top 1%
TM[1]		68.14	0.30	-	56.59	0.49	-	-62.06	-81.47
Hybrid TM+NPLS1	**NPLS1 alone**[2]	70.83	0.09	-201.34	52.27	0.23	-53.74	-43.50	-41.44
	TM+NPLS1[2]	90.71	0.09	-197.35	79.28	0.23	-49.75	-27.73	-33.76
Hybrid TM+NPLS1 +NPLS2	**NPLS2 alone**[3]	6.52	0.30	-104.15	4.27	0.51	-28.61	137.69	41.32
	TM+NPLS1+ NPLS2[3]	91.31	0.08	-201.20	80.16	0.22	-49.71	-26.97	-37.39

[1] β=0.30, [2] With encoding 4 and *Fac*=6, [3] With encoding 5 and *Fac*=1

The hybrid TM+NPLS1 was recalculated with the contributions of the TM and NPLS1 modules together, obtained by summing the output of both modules (Table 4.2). This procedure improved the description of the model identification data set from 68.1% explained variance (TM model) to 90.7% (hybrid model). More importantly, the test data set improved from 56.6% (TM model) to 79.3% explained variance (hybrid TM+NPLS1 model), which is very significant.

Table 4.3 – Model performance criteria for TM and hybrid models for (partition E67)

Model		Identification			Test			Relative Error	
		% Var	*MSE*	*AIC*	% Var	*MSE*	*AIC*	Top 5%	Top 1%
TM[1]		85.86	0.25	-	-63.51	0.98	-	-98.32	-112.83
HYBRID TM+NPLS1	**NPLS1 alone[2]**	28.63	0.18	-73.61	4.88	0.93	-4.05	-69.68	-62.38
	TM+NPLS1[2]	89.91	0.18	-69.61	-55.54	0.93	-0.05	-67.00	-70.38
HYBRID TM+NPLS1+ NPLS2	**NPLS2 alone[3]**	18.97	0.25	-56.78	1.34	0.99	3.89	-50.39	-39.69
	TM+NPLS1+ NPLS2[3]	91.82	0.15	-74.87	-53.44	0.92	2.76	-56.49	-69.75

[1] β=0.25, [2] With encoding 4 and *Fac* 1, [3] With encoding 5 and *Fac* 2

Fig. 4.4 plots the predicted over measured protein fluorescence data for the hybrid TM+NPLS1 and respective residuals distribution. Comparing with the standalone TM model (Fig. 4.3), it may be seen that the dispersion of model identification residuals decreases 1.7-fold for the hybrid model (Fig. 4.3 B, σ=0.32) when compared to the TM (Fig. 4.3 B, σ=0.55). In the case of the test partition, it strikes the almost 3-fold reduction in model bias (μ=-0.62 for the TM, Fig. 4.3 C, μ=0.23 for the hybrid model, Fig. 4.4 C). 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 model.

4.4.3.2. 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. Thus, the job of NPLS2 is to improve the TM+NPLS1 model by considering mRNA-rRNA interactions information that is not accounted for neither by the TM nor by the NPLS1 model.

The results of this identification step are summarized in Table 4.2 and Table 4.3 for partition E33 and E67, respectively. Additionally, the detailed results for these cases are presented in Table 4.13 and Table 4.14, respectively. The inclusion of NPLS2 module seems to not significantly improve the hybrid model performance. In the case of partition E33 (Table 4.2), the model performance improves from 90.7% to 91.3% in the identification partition and from 79.3% to 80.2% in the test partition. In the case of partition E67, the improvement is higher both in the identification partition (from 89.9% to 91.8% in explained variance) and in the test partition, 2.1% higher explained variance, although still negative (increases from -55.5% to -53.4%).

Again focusing in partition E33, we calculated the *AIC* values to discriminate the more parsimonious model. The *AIC* values are -49.75 (TM+NPLS1) and -49.71 (TM+NPLS1+NPLS2) for the test data set and -197.35 (TM+NPLS1) and -201.20 (TM+NPLS1+NPLS2) for the identification data set. According to this criterion the hybrid TM+NPLS1+NPLS2 is a more parsimonious model than TM+NPLS1 in the identification data set, but not in the test data set. Thus, such residual

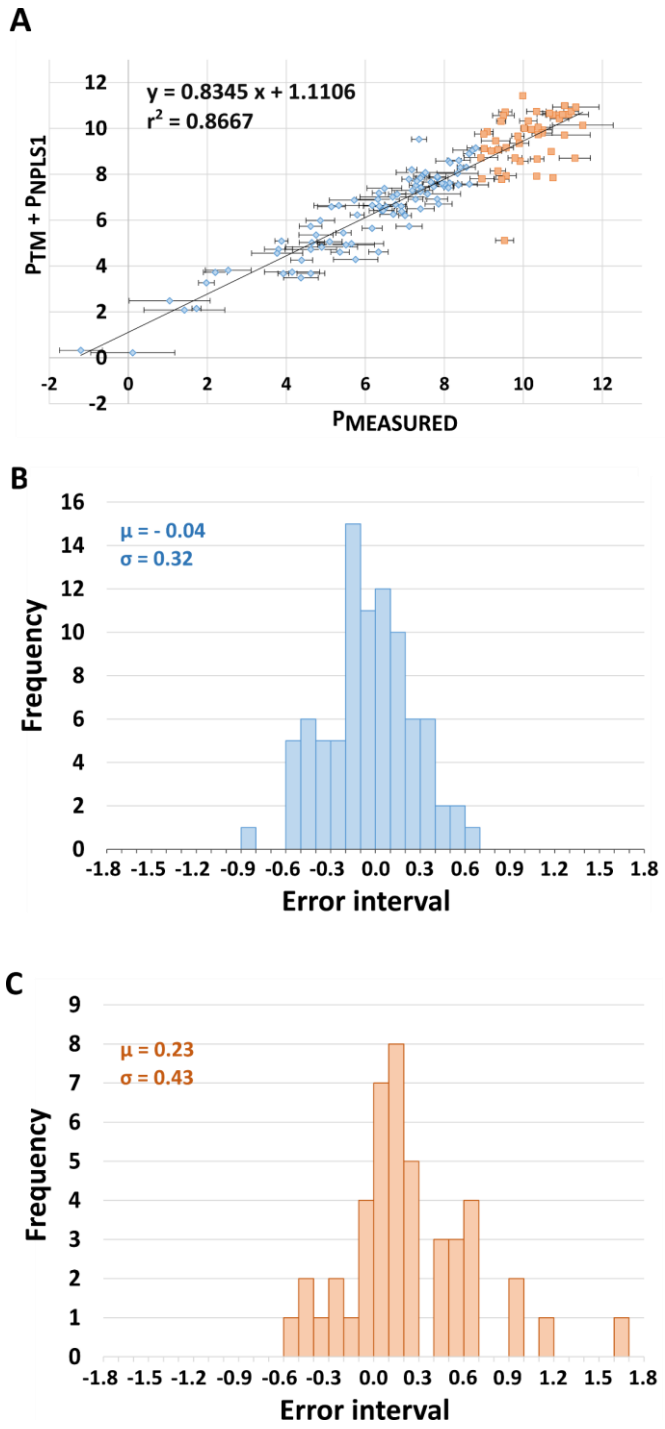

Fig. 4.4 – **Hybrid model TM+NPLS1 modeling results for partition E33.** 67% of mRNA sequences with lowest measured protein fluorescence are used for model calibration (identification data set) while 33% of mRNA sequences with highest measured protein fluorescence are used for model predictive power 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. Data points are normalized in panels B and C.

improvement is inconclusive. Even so, comparing Table 4.11 and Table 4.13, a higher consistency was obtained for TM+NPLS1+NPLS2, where the model explained variance was equivalent for all encoding in the identification and test partitions.

4.5. Conclusions

Data sparsity is a major limitation when developing QSAM with sufficient predictive power for model-based design of synthetic SBPs. In this study we have investigated the use of hybrid semiparametric systems for QSAM using a sparse 132 5'UTR sequence-activity data set in *E. coli* as application case study. The main conclusions to be withdrawn from this study are the following:

- The Thermodynamic Model by Salis *et al.* (6) when identified with the 87 lowest protein expression sequences (identification partition) and assessed with the 44 highest protein titer sequences (test partition), results in much lower accuracy in the test partition than in the identification partition. Moreover, the TM exhibits a significant bias in the sense of protein underprediction. As expected, these results degrade even further when the identification partition is restricted to the 44 lowest titer sequences.
- The standalone N-PLS model based on primary mRNA structure (NPLS1), has higher data fitting flexibility (higher accuracy in the identification data set) and considerably worse extrapolation capacity (lower accuracy in the test partition) than the TM model. Thus, it is prone to overfitting and is less powerful at predicting new sequences. The same conclusions hold for the N-PLS model based on the mRNA-rRNA possible interactions (NPLS2).
- The hybrid model structure TM+NPLS1 is clearly advantageous over standalone TM or NPLS1. Explained variance increases 23% both in the identification and testing data sets (partition E33 – Table 4.2). Hence, this hybrid structure is significantly better at both interpolating and extrapolating in relation to the referred standalone TM model. Moreover, bias is significantly reduced in the hybrid model and its residuals are normal as opposed to the TM model.
- The hybrid structure TM+NPLS1+NPLS2 produces only residual improvements in relation to TM+NPLS1 for the present data set.
- Although encoding 4 stands out as the best encoding in the TM+NPLS1 hybrid model structure, and one of the best in the TM+NPLS1+NPLS2 hybrid model structure, taking all results together, it is not straightforward to discriminate the best encoding method for this data set. That is, apparently, the data portioning greatly affects the performance of the encoding.

All in all, this study is a first step towards the demonstration of the potential of hybrid modeling in Synthetic Biology. We clearly show that, for the 5'UTR sequence-activity modeling problem in *E. coli,* hybrid models are better at interpolating and, more importantly, at extrapolating than TM or empirical N-PLS modeling when applied as standalone techniques. Many other

problems could benefit from a hybrid modeling framework. In complex eukaryotic organisms, mechanistic models embody fewer details as compared to simpler prokaryotic organisms. For instance, the nucleosome occupancy has been one of the key features to be included in a model for transcription initiation in *S. cerevisiae* (31). We envision that hybrid semiparametric modeling could be of high value for QSAM, not restricted, but particularly to complex eukaryotic organisms.

4.6. Supplementary information

Table 4.4 – The free Gibbs energy parameters for each of the 132 mRNA sequences and respective natural logarithm of the reporter protein fluorescence.

Index	*ln*(Y)	ΔG_{TOT}	$\Delta G_{mRNA:rRIB}$	ΔG_{START}	$\Delta G_{SPACING}$	$\Delta G_{STANDBY}$	ΔG_{mRNA}
1	4.62	4.81	-15.28	-1.19	12.18	0.00	-9.10
2	4.89	3.89	-15.98	-1.19	11.96	0.00	-9.10
3	5.51	1.56	-15.98	-1.19	9.63	0.00	-9.10
4	10.13	-6.55	-15.98	-1.19	1.53	0.00	-9.10
5	10.90	-8.07	-15.98	-1.19	0.01	0.00	-9.10
6	10.75	-8.07	-15.98	-1.19	0.00	0.00	-9.10
7	10.66	-7.78	-15.98	-1.19	0.29	0.00	-9.10
8	10.49	-7.40	-15.98	-1.19	0.67	0.00	-9.10
9	10.20	-6.92	-15.98	-1.19	1.15	0.00	-9.10
10	9.90	-6.34	-15.98	-1.19	1.73	0.00	-9.10
11	9.57	-5.67	-15.98	-1.19	2.40	0.00	-9.10
12	8.63	-4.04	-15.98	-1.19	4.03	0.00	-9.10
13	7.11	-0.87	-15.98	-1.19	7.20	0.00	-9.10
14	6.17	2.82	-6.18	-1.19	0.29	0.00	-9.90
15	6.77	1.62	-7.38	-1.19	0.29	0.00	-9.90
16	4.61	5.98	-5.28	-1.19	1.15	0.00	-11.30
17	6.17	2.82	-7.48	-1.19	0.29	0.00	-11.20
18	5.14	3.82	-5.38	-1.19	0.29	0.00	-10.10
19	2.53	9.22	-1.98	-1.19	0.29	0.00	-12.10
20	3.80	9.53	-4.08	-1.19	0.00	0.00	-14.80
21	8.14	11.46	-10.98	-1.19	9.63	0.00	-14.00
22	9.36	3.73	-8.88	-1.19	0.01	0.00	-13.80
23	9.57	-1.77	-9.68	-1.19	0.00	0.00	-9.10
24	8.11	-0.58	-8.78	-1.19	0.29	0.00	-9.10
25	8.37	-0.04	-9.68	-1.19	1.73	0.00	-9.10

Table 4.4 (cont.) – The free Gibbs energy parameters for each of the 132 mRNA sequences and respective natural logarithm of the reporter protein fluorescence.

Index	*ln*(Y)	ΔG_{TOT}	$\Delta G_{mRNA:rRIB}$	ΔG_{START}	$\Delta G_{SPACING}$	$\Delta G_{STANDBY}$	ΔG_{mRNA}
26	7.41	1.40	-9.68	-1.19	3.17	0.00	-9.10
27	9.86	-5.10	-15.18	-1.19	0.67	0.00	-10.60
28	9.54	-7.27	-15.18	-1.19	0.00	0.00	-9.10
29	9.09	-3.24	-15.18	-1.19	1.73	0.00	-11.40
30	7.83	2.30	-8.68	-1.19	0.67	0.00	-11.50
31	7.27	4.20	-6.08	-1.19	0.67	0.00	-10.80
32	8.64	2.70	-8.78	-1.19	0.67	0.00	-12.00
33	7.36	-4.67	-15.08	-1.19	2.40	0.00	-9.20
34	6.81	1.10	-7.78	-1.19	0.67	0.00	-9.40
35	8.00	2.50	-8.48	-1.19	0.67	0.00	-11.50
36	7.01	3.50	-6.28	-1.19	0.67	0.00	-10.30
37	7.66	1.63	-6.28	-1.19	0.00	0.00	-9.10
38	9.36	-1.10	-9.68	-1.19	0.67	0.00	-9.10
39	6.33	4.20	-6.48	-1.19	0.67	0.00	-11.20
40	5.65	6.82	-6.48	-1.19	0.29	0.00	-14.20
41	4.64	5.03	-6.78	-1.19	0.01	0.00	-13.00
42	10.33	-7.08	-16.08	-1.19	0.29	0.00	-9.90
43	10.38	-6.88	-15.98	-1.19	0.29	0.00	-10.00
44	10.91	-6.47	-15.88	-1.19	0.01	0.00	-10.60
45	10.48	-6.08	-15.88	-1.19	0.29	0.00	-10.70
46	9.50	-5.37	-15.98	-1.19	0.01	0.00	-11.80
47	9.30	-4.88	-14.38	-1.19	0.29	0.00	-10.40
48	9.92	-4.97	-15.38	-1.19	0.00	0.00	-11.60
49	6.35	-3.67	-14.88	-1.19	0.00	0.00	-12.40
50	7.19	-2.37	-10.48	-1.19	0.01	0.00	-9.30
51	6.48	-3.30	-12.48	-1.19	0.67	0.00	-9.70
52	8.36	-2.57	-12.48	-1.19	0.00	0.00	-11.10
53	6.32	-2.37	-11.38	-1.19	0.00	0.00	-10.20
54	8.96	-2.28	-13.98	-1.19	0.29	0.00	-12.60
55	7.73	-2.12	-13.18	-1.19	1.15	0.00	-11.10
56	8.93	-0.97	-13.98	-1.19	0.00	0.00	-14.20
57	8.34	-0.82	-11.48	-1.19	1.15	0.00	-10.70
58	7.58	0.63	-8.08	-1.19	0.01	0.00	-9.90
59	7.82	1.43	-10.68	-1.19	0.01	0.00	-13.30
60	5.75	4.00	-5.08	-1.19	0.67	0.00	-9.60
61	5.09	3.16	-8.38	-1.19	1.73	0.00	-11.00
62	7.40	3.13	-7.98	-1.19	0.01	0.00	-12.30

Table 4.4 (cont.) – The free Gibbs energy parameters for each of the 132 mRNA sequences and respective natural logarithm of the reporter protein fluorescence.

Index	*ln*(Y)	ΔG_{TOT}	$\Delta G_{mRNA:rRIB}$	ΔG_{START}	$\Delta G_{SPACING}$	$\Delta G_{STANDBY}$	ΔG_{mRNA}
63	5.72	4.08	-8.58	-1.19	1.15	0.00	-12.70
64	5.33	4.32	-5.48	-1.19	0.29	0.00	-10.70
65	4.15	4.73	-4.78	-1.19	0.00	0.00	-10.70
66	4.62	6.76	-4.78	-1.19	1.73	0.00	-11.00
67	4.75	6.93	-5.88	-1.19	2.40	0.00	-11.60
68	4.37	7.96	-12.18	-1.19	9.63	0.00	-11.70
69	1.42	14.85	-1.08	-1.19	1.53	0.00	-15.60
70	1.72	15.28	-2.78	-1.19	8.45	0.00	-10.80
71	-1.20	15.23	-4.28	-1.19	0.00	-4.30	-16.40
72	0.11	22.71	-9.68	-1.19	12.18	-1.80	-19.60
73	8.15	-2.80	-14.58	-1.19	0.67	0.00	-12.30
74	8.56	-0.97	-10.98	-1.19	0.01	0.00	-11.20
75	2.20	6.92	-1.98	-1.19	0.29	0.00	-9.80
76	6.41	3.13	-5.98	-1.19	2.40	0.00	-7.90
77	7.39	0.33	-5.38	-1.19	0.01	0.00	-6.90
78	8.01	-2.67	-8.08	-1.19	2.40	0.00	-4.20
79	10.38	-5.04	-11.58	-1.19	1.73	0.00	-6.00
80	10.75	-4.77	-8.78	-1.19	0.01	0.00	-5.20
81	10.71	-8.37	-11.18	-1.19	0.00	0.00	-4.00
82	11.15	-9.27	-15.98	-1.19	0.01	0.00	-7.90
83	6.52	7.58	-4.18	-1.19	6.05	0.00	-6.90
84	5.36	15.32	-3.68	-1.19	9.79	0.00	-10.40
85	10.34	-4.47	-11.18	-1.19	0.01	0.00	-7.90
86	9.44	-2.57	-10.08	-1.19	0.01	0.00	-8.70
87	5.79	7.05	-11.18	-1.19	1.53	-6.60	-11.30
88	7.83	2.96	-5.48	-1.19	4.03	0.00	-5.60
89	7.43	0.26	-3.28	-1.19	1.73	0.00	-3.00
90	6.69	-2.48	-15.38	-1.19	9.79	0.00	-4.30
91	11.00	-4.37	-12.38	-1.19	0.01	-0.30	-8.90
92	11.30	-5.30	-13.08	-1.19	3.17	0.00	-5.80
93	9.99	-7.77	-9.68	-1.19	0.00	0.00	-3.10
94	7.85	4.13	-9.38	-1.19	0.01	0.00	-14.70
95	8.72	-1.55	-9.38	-0.08	0.01	0.00	-7.90
96	9.79	-0.40	-9.58	-1.19	0.67	0.00	-9.70
97	9.45	4.82	-9.58	-0.08	0.67	0.00	-13.80
98	10.35	-4.40	-15.28	-1.19	0.67	-0.20	-11.20

Table 4.4 (cont.) – The free Gibbs energy parameters for each of the 132 mRNA sequences and respective natural logarithm of the reporter protein fluorescence.

Index	*ln*(Y)	ΔG_{TOT}	$\Delta G_{mRNA:rRIB}$	ΔG_{START}	$\Delta G_{SPACING}$	$\Delta G_{STANDBY}$	ΔG_{mRNA}
99	7.28	1.52	-15.28	-0.08	0.67	-0.20	-16.00
100	9.00	-5.54	-15.38	-1.19	1.73	0.00	-9.30
101	7.94	3.67	-15.38	-0.08	1.73	0.00	-17.40
102	5.45	8.80	-8.28	-1.19	0.67	-1.20	-16.40
103	11.50	-3.80	-8.28	-1.19	0.67	-1.20	-3.80
104	4.86	7.33	-10.98	-1.19	0.00	-2.20	-17.30
105	11.21	-5.47	-10.98	-1.19	0.00	-2.20	-4.50
106	8.11	0.02	-13.88	-1.19	0.29	0.00	-14.80
107	11.05	-8.88	-13.88	-1.19	0.29	0.00	-5.90
108	6.45	5.36	-9.48	-1.19	1.73	0.00	-14.30
109	11.33	-6.64	-9.48	-1.19	1.73	0.00	-2.30
110	9.52	3.93	-9.58	-1.19	0.00	0.00	-14.70
111	11.04	-8.37	-9.58	-1.19	0.00	0.00	-2.40
112	3.77	8.73	-3.28	-1.19	0.00	0.00	-13.20
113	6.71	0.73	-3.28	-1.19	0.00	0.00	-5.20
114	7.30	3.33	-10.88	-1.19	0.00	0.00	-15.40
115	9.19	-3.37	-10.88	-1.19	0.00	0.00	-8.70
116	6.91	4.43	-8.58	-1.19	0.00	0.00	-14.20
117	8.40	-3.57	-8.58	-1.19	0.00	0.00	-6.20
118	7.10	5.93	-7.78	-1.19	0.01	0.00	-14.90
119	9.01	-0.87	-7.78	-1.19	0.01	0.00	-8.10
120	8.79	-1.98	-14.38	-1.19	0.29	0.00	-13.30
121	10.02	-7.18	-14.38	-1.19	0.29	0.00	-8.10
122	1.05	6.59	-14.38	-1.19	11.96	0.00	-10.20
123	1.98	4.26	-14.38	-1.19	9.63	0.00	-10.20
124	6.35	-3.55	-14.38	-1.19	1.53	0.00	-10.50
125	7.18	-4.87	-14.38	-1.19	0.00	0.00	-10.70
126	7.40	-4.40	-14.38	-1.19	0.67	0.00	-10.50
127	7.52	-3.22	-14.38	-1.19	1.15	0.00	-11.20
128	6.92	0.66	-14.38	-1.19	4.03	0.00	-12.20
129	4.39	2.48	-14.38	-1.19	6.05	0.00	-12.00
130	3.88	3.63	-14.38	-1.19	7.20	0.00	-12.00
131	3.93	8.08	-14.38	-1.19	8.45	0.00	-15.20
132	3.12	11.40	-14.38	-1.19	12.77	0.00	-14.20

Table 4.5 – NPLS1 identification results for data partition R. Namely, *Fac* number, explained variance and percentage of error for calibration and test partitions and top 5 and top 1 data points, respectively.

Enc.	Number of *Fac*	PLS Explained Variance for Calibration Partition (%)	PLS Explained Variance for Test Partition (%)	Error of Highest Protein Titer Prediction (top 5 - %)	Error of Highest Protein Titer Prediction (top 1 - %)
1	3	65.22	22.98	-35.97	-7.62
2	3	62.77	15.56	-44.14	-33.57
3	4	75.12	28.20	-39.28	-19.74
4	4	72.84	17.77	-34.38	-29.09
5	3	71.20	29.02	-45.00	-21.82
6	3	58.41	0.93	-54.33	-41.25

Table 4.6 – NPLS1 identification results for data partition E33. Namely, *Fac* number, explained variance and percentage of error for calibration and test partitions and top 5 and top 1 data points, respectively.

Enc.	Number of *Fac*	PLS Explained Variance for Calibration Partition (%)	PLS Explained Variance for Test Partition (%)	Error of Highest Protein Titer Prediction (top 5 - %)	Error of Highest Protein Titer Prediction (top 1 - %)
1	2	50.36	26.76	-53.49	-16.12
2	1	18.63	22.31	-63.08	-48.89
3	3	63.30	51.32	-56.62	3.53
4	1	41.65	-100.54	-119.89	-110.14
5	3	66.33	49.32	-35.42	-0.35
6	2	47.08	-120.66	-140.80	-133.54

Table 4.7 – NPLS1 identification results for data partition E67. Namely, *Fac* number, explained variance and percentage of error for calibration and test partitions and top 5 and top 1 data points, respectively.

Enc.	Number of *Fac*	PLS Explained Variance for Calibration Partition (%)	PLS Explained Variance for Test Partition (%)	Error of Highest Protein Titer Prediction (top 5 - %)	Error of Highest Protein Titer Prediction (top 1 - %)
1	5	95.65	-527.06	-63.30	-42.21
2	6	97.15	-482.43	-60.23	-35.17
3	5	95.19	-555.74	-50.94	-24.46
4	2	84.20	-366.98	-172.01	-148.24
5	3	90.23	-147.00	-39.61	21.50
6	2	85.42	-359.35	-162.54	-146.12

Table 4.8 – NPLS2 identification results for data partition R. Namely, *Fac* number, explained variance and percentage of error for calibration and test partitions and top 5 and top 1 data points, respectively.

Enc.	Number of *Fac*	PLS Explained Variance for Calibration Partition (%)	PLS Explained Variance for Test Partition (%)	Error of Highest Protein Titer Prediction (top 5 - %)	Error of Highest Protein Titer Prediction (top 1 - %)
1	1	9.41	0.34	-96.53	-99.48
2	1	15.27	6.53	-92.53	-121.33
3	1	13.09	5.20	-96.80	-116.33
4	3	29.43	10.58	-80.94	-134.60
5	2	22.82	10.51	-82.54	-136.13
6	1	11.22	4.76	-96.93	-104.08

Table 4.9 – NPLS2 identification results for data partition E33. Namely, *Fac* number, explained variance and percentage of error for calibration and test partitions and top 5 and top 1 data points, respectively.

Enc.	Number of *Fac*	PLS Explained Variance for Calibration Partition (%)	PLS Explained Variance for Test Partition (%)	Error of Highest Protein Titer Prediction (top 5 - %)	Error of Highest Protein Titer Prediction (top 1 - %)
1	1	8.91	6.46	-98.09	-98.40
2	2	12.30	10.99	-97.43	-114.44
3	1	11.09	8.85	-99.33	-111.33
4	1	16.64	15.29	-88.00	-114.22
5	1	13.77	12.42	-93.61	-120.14
6	1	9.37	8.15	-97.67	-101.70

Table 4.10 – NPLS2 identification results for data partition E67. Namely, *Fac* number, explained variance and percentage of error for calibration and test partitions and top 5 and top 1 data points, respectively.

Enc.	Number of *Fac*	PLS Explained Variance for Calibration Partition (%)	PLS Explained Variance for Test Partition (%)	Error of Highest Protein Titer Prediction (top 5 - %)	Error of Highest Protein Titer Prediction (top 1 - %)
1	1	10.14	5.69	-97.48	-97.82
2	1	15.60	7.45	-96.54	-117.85
3	1	14.31	5.55	-98.46	-114.08
4	2	53.89	-815.65	-76.66	-191.75
5	1	18.83	5.33	-91.93	-126.43
6	1	10.50	7.46	-97.22	-101.83

Table 4.11 – Hybrid model TM+NPLS1 identification results for data partition E33.

Enc.	Number of *Fac*	PLS Explained variance (%)	Hyb. Explained variance (%)	PLS Explained variance (%)	Hyb. Explained variance (%)
		Calibration Partition		Test Partition	
1	2	36.63	79.81	-3.60	55.02
2	2	28.95	77.36	12.42	61.98
3	1	12.50	72.12	15.16	63.17
4	6	70.83	90.71	52.27	79.28
5	2	34.18	79.03	10.16	61
6	1	10.91	71.62	12.43	61.99
TM	-	-	68.14	-	56.59

Table 4.12 – Hybrid model TM+NPLS1 identification results for data partition E67.

Enc.	Number of *FAC*	PLS Explained variance (%)	Hyb. Explained variance (%)	PLS Explained variance (%)	Hyb. Explained variance (%)
		Calibration Partition		Test Partition	
1	2	49.29	92.83	-7.27	-75.40
2	1	33.32	90.57	-3.41	-69.09
3	2	47.39	92.56	-5.03	-71.75
4	1	28.63	89.91	4.88	-55.54
5	1	39.26	91.41	-1.51	-65.99
6	1	35.36	90.86	-2.40	-67.44
TM	-	-	85.86	-	-63.51

Table 4.13 – Hybrid model TM+NPLS1+NPLS2 identification results for data partition E33.

Enc.	Number of *FAC*	PLS Explained variance (%)	Hyb. Explained variance (%)	PLS Explained variance (%)	Hyb. Explained variance (%)
		Calibration Partition		Test Partition	
TM+NPLS1 (ENC. 4)	6	70.83	90.71	52.26	79.28
1	1	5.44	91.21	2.31	79.75
2	1	6.25	91.28	3.21	79.94
3	1	5.74	91.24	2.89	79.87
4	1	6.01	91.27	3.81	80.07
5	1	6.52	91.31	4.27	80.16
6	1	5.71	91.23	2.32	79.75
TM	-	-	68.14	-	56.59

Table 4.14 – Hybrid model TM+NPLS1+NPLS2 identification results for data partition E67.

Enc.	number of *FAC*	PLS Explained variance (%)	Hyb. Explained variance (%)	PLS Explained variance (%)	Hyb. Explained variance (%)
		Calibration Partition		Test Partition	
TM+NPLS1 (ENC. 4)	1	28.63	89.91	4.88	-55.54
1	8	58.95	95.85	-32.13	-105.51
2	2	19.28	91.85	0.27	-55.10
3	2	19.68	91.89	0.33	-55.02
4	3	33.20	93.25	-16.18	-80.71
5	2	18.97	91.82	1.34	-53.44
6	4	41.64	94.11	-8.11	-68.15
TM	-	-	85.86	-	-63.51

4.7. References

1. Marchisio,M. and Stelling,J. (2008) Computational design of synthetic gene circuits with composable parts. *Bioinformatics*, **24**, 1903–1910.

2. 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.

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. and Voigt,C. (2010) 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. Cebe,R. and Geiser,M. (2006) Rapid and easy thermodynamic optimization of the 5'-end of mRNA dramatically increases the level of wild type protein expression in *Escherichia coli*. *Protein Expr Purif*, **45**, 374–380.

10. Racle,J., Overney,J. and Hatzimanikatis,V. (2012) A computational framework for the design of optimal protein synthesis. *Biotechnol. Bioeng.*, **109**, 2127–2133.

11. Carothers,J.M., Goler,J.A., Juminaga,D. and Keasling,J.D. (2011) Model-Driven Engineering of RNA Devices to Quantitatively Program Gene Expression. *Science*, **334**, 1716–1719.

12. 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.

13. Tavares,L.G., Lopes,H.S., Lima,C.R.E. (2008) A Comparative Study of Machine Learning Methods for Detecting Promoters in Bacterial DNA Sequences. *International Conference on Intelligent Computing,* **5227,** 959–966.

14. Li,G., Leong,T. and Zhang,L. (2004). Translation initiation sites prediction with mixture gaussian models. In: Jonassen,I. and Kim,J., Algorithms in Bioinformatics: 4th International Workshop, pp. 338–349, Springer-Velrag, Berling Heidelberg.

15. Li,S.M., Wakefield,J. and Self,S. (2008) A transdimensional Bayesian model for pattern recognition in DNA sequences. *Biostatistics*, **9**, 668–685.

16. Zuo,Y.C. and Li,Q.Z. (2010) The hidden physical codes for modulating the prokaryotic transcription initiation. *Phys. A Stat. Mech. its Appl.*, **389**, 4217–4233.

17. Jonsson,J., Norberg,T., Carlsson,L., Gustafsson,C. and Wold,S. (1993) Quantitative Sequence-Activity Models (QSAM)-Toolsfor Sequence Design. *Nucleic Acids Res*, **21**, 733–739.

18. 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.

19. Ran,W. and Higgs,P.G. (2012) Contributions of speed and accuracy to translational selection in bacteria. *PLoS One*, **7**, e51652.

20. Stosch, M.v., Carinhas, N. and Oliveira, R. (2014) Hybrid Modeling for Systems Biology: Theory and Practice. In: Benner P, Findeisen R, Flockerzi D, Reichl U and Sundmacher K. Large-Scale Networks in Engineering and Life Sciences. pp. 367–388. Springer, Heidelberg. 10.1007/978-3-319-08437-4_7.

21. 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.

22. 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.

23. Dirks,R.M., Bois,J.S., Schaeffer,J.M., Winfree,E. and Pierce,N. (2007) Thermodynamic Analysis of Interacting Nucleic Acid Strands. *SIAM Rev.*, **49**, 65–88.

24. Nandy,A. (2006) Mathematical descriptors of DNA sequences: development and applications. *Arkivoc*, **2006**, 211–238.

25. Bro,R. (1996) Multiway calibration. Multilinear PLS. *J. Chemom.*, **10**, 47–61.

26. Andersson,C.A. and Bro,R. (2000) The N-way Toolbox for MATLAB. *Chemom. Intell. Lab. Syst.*, **52**, 1–4.

27. Li,B., Morris,J. and Martin,E.B. (2002) Model selection for partial least squares regression. *Chemom. Intell. Lab. Syst.*, **64**, 79–89.

28. Stosch,M.v., 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.

29. Stosch,M.v., Davy,S., Francois,K., Galvanauskas,V., Hamelink,J.M., Luebbert,A., Mayer,M., Oliveira,R., O'Kennedy,R., Rice,P., *et al.* (2014) Hybrid modeling for quality by design and PAT-benefits and challenges of applications in biopharmaceutical industry. *Biotechnol. J.*, **9**, 719–726.

30. Stosch,M.v., 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.

31. 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**, 1–8.

Chapter 5

Conclusions and Future work

5.1. General conclusions

The main achievements in Synthetic Biology have been accomplished through the implementation of new biological functions using synthetic genetic circuits. Such circuits are built by assembling Synthetic Biological Parts (SBPs). However, given the many possible circuit topologies that can perform equivalent functions, the different SBPs that have similar outputs and the degrees of freedom to design and fine-tune SBPs, the design and implementation of synthetic circuits is nontrivial.

Given the underlying complexity, before implementing a synthetic circuit experimentally, it is necessary to investigate the network topology and chose the appropriate SBPs. To this end, several System Biology approaches were used to bridge the gap between the cellular behavior and a specific DNA sequence. Namely, through the implementation software packages that take advantage of SBPs repositories, and their kinetic parameters to simulate the circuit behavior (1). However, even when the selected circuit topology is adequate for the problem at hand, the SBPs performance still varies greatly, depending on the local genetic environment.

To target this problem, in the last years, advances in high throughput screening were made. This enabled a faster screening and selection of a circuit with the desired behavior from a library of circuits created based on the available SBPs for a given chassis. Given the high amount of data being generated, there is the opportunity for the development of improved models that can predict more accurately the SBPs behavior in different contexts and enable the creation of more complex circuits without the need for exhaustive experimental screenings.

In this thesis, different problems related with the design and fine-tuning of SBPs were targeted. Such problems can appear throughout the development of a synthetic genetic circuit, namely: *de novo* design of SBPs, together with their modularity and orthogonality test, interpretation of their most important features, fine-tuning of SBPs and development of new approaches for the improvement of current SBPs design tools.

In Chapter 2, a systematic study of the effect of up to three point mutations on a eukaryotic core promoter strength was presented. In this case, the target was *P. pastoris AOX1* core promoter, where adjacent mutations of three nucleotides each were examined. To the best of our knowledge, this study represents the highest resolution experiment to this date. Overall, these results show that this promoter was remarkably robust, withstanding most mutations without a signific ant change in its expression. Also, we were able to identify key core promoter regions that, when mutated, cause a change in expression (*e. g.* protein binding regions – TATA box – or regions with some biological importance – downstream of transcriptional start site and 5' Untranslated Region (5'UTR) next to the protein start codon). Altogether, these results show the high tolerance towards few mutations (up to three point mutations) that yeast core promoters are able to endure, supporting regulatory models of degenerate regulatory motifs or redundant design.

In Chapter 3, the generation of completely synthetic core promoters, on the same expression system, was addressed. The design was based on data from a genome wide analysis of natural *S. cerevisiae* core promoters (2). In particular, these synthetic core promoters were developed taking into consideration four main factors: nucleotide occurrence frequency along the promoter, TATA box location, important motifs location and frequency and nucleosome affinity. The generated promoters were then screened in different contexts, to test their modularity and orthogonality. Overall, this work demonstrated the feasibility of a multi factor rational synthetic core promoter design (not only limited to one factor – *e. g.* nucleosome affinity (2) – nor based on random or rational diversification of natural core promoters). It was also shown its applicability as general engineering tool for gene expression fine-tuning. Due to their sequence diversity and independence of natural sequences, similarly designed synthetic core promoters may become valuable tools for Synthetic Biology and metabolic engineering applications in other eukaryotic organisms.

Lastly, on Chapter 4, a new SBPs design method is presented. In this chapter we focused on the *Escherichia coli* expression system. More specifically, we focused on the Ribosome Binding Site (RBS) as the target SBP to modulate protein production. The presented method aimed to improve the prediction capabilities of the current state of the art mechanistic model by adding information using a machine learning procedure. The main motivation for this approach is that data sparsity is a major limitation when developing Quantitative Sequence-Activity Models (QSAM) with sufficient predictive power for model-based design of synthetic SBPs. In this study, we have investigated the use of hybrid semiparametric systems, which join mechanistic knowledge with information derived from data for QSAM. All in all, this study was a first step towards the demonstration of the potential of hybrid modeling in Synthetic Biology. It is clearly shown that, for the 5'UTR QSAM problem in *E. coli*, the current state of the art mechanistic model can be improved by coupling it with a non-parametric model, in a hybrid manner.

5.2. Future work

From the results presented in this thesis, three main future research routes can be taken:

i. On its own, the *AOX1* core promoter mutations screening study seems to be insufficient to draw general rules on the effect of similar mutations on others promoters. That is, it is unlikely that it would be possible to predict how a similar mutation would affect other promoters only based on the type of mutation, its location and its effect on this promoter expression. Even so, we might speculate that most core promoters in *Pichia pastoris*, other yeasts and even eukaryotes evolved to be able to, in the same manner as *AOX1* core promoter, withstand up to three mutations on non-essential *loci* without a significant change in expression. Therefore, when using a promoter in several genetic contexts, it might be useful to test similar core promoter modifications to target the desired SBP expression. A similarly developed data set for others SBPs of interest might be used to adjust their functionality within a narrow interval, given that, as in this case, most of the mutations had a non-significant effect on expression.

Additionally, testing the cumulative effects of those mutations that increase or decrease the promoter strength may constitute a promising approach to expand the fine-tune expression window in this particular case. These results could also be used to enlighten the eventual cooperative effect of such mutations.

With the accumulation of acquired data (similar to the one described previously), it may be possible to create a general procedure to fine-tune expression using rationally directed mutations, rather than random mutations followed by high throughput screening, as most studies use nowadays.

ii. In alternative to the SBPs developed by diversification of a natural sequence, it was shown that it is possible to develop *de novo* SBPs (core promoters). The main question that remains unanswered is why the synthetic core promoters, that in principle only interact with general transcription factors, did not function with constitutive CRMs, but showed good functionality when controlled by the inducible ones. Interestingly, it is possible that it is related with an unknown regulatory mechanism.

To try to answer this question, it could be firstly investigated if the remaining synthetic promoters (not tested with a constitutive CRM) are also non-functional in these conditions. If some of the promoters show some expression, a bioinformatics analysis would, in principle, allow the identification of common features between the functional synthetic and the natural constitutive core promoters. In a second phase, the motifs found to be fundamental for gene expression with constitutive CRMs could be substituted by a neutral sequence. It would be expected, if these motifs are indeed fundamental for gene expression in these conditions, an expression disruption after this change. Such disruption would be similar to the one observed in Chapter 3 when TATA box was mutated Fig. 3.9.

In parallel, it would be interesting to show, in an applied metabolic engineering project, the advantages of using such promoters, in comparison with the use of a limited set of natural promoters in *P. pastoris*.

iii. In Synthetic Biology, many other problems (in addition to the ones considered in this thesis) could benefit from a hybrid modeling framework. The complex eukaryotic organisms' mechanistic models embody fewer details from the real mechanism as compared to the models of simpler prokaryotic cells. For instance, the nucleosome occupancy has been one of the key features to be included in a model for transcription initiation in *S. cerevisiae* (3). We envision that models such as this one (3), but not limited to, could highly benefit from the hybrid semiparametric modeling. Overall, hybrid models could be used to develop and improve the SBPs design methods and also to assist in the circuit topology choice, targeting different objective functions, such as reduced circuit noise or increased robustness.

Overall, this Ph. D. thesis targeted the main problems that occur prior to the synthetic genetic circuit implementation: improvement of the current SBPs design tools, development of *de*

novo SBPs, their characterization and how to rationally fine-tune them. The results and methods presented here aim at reducing the time that takes to develop a synthetic genetic circuits, by the means of offering alternative completely synthetic SBPs that might facilitate the cloning procedure. Moreover, the circuit development efficiency can be improved by the enhancement of the available predictive models quality. In this way, we presented here the first hybrid model for the design of SBPs. As SBPs data become more and more available, the development of models that take it into account will be essential.

5.3. References

1. Hill,A.D., Tomshine,J.R., Weeding,E.M.B., Sotiropoulos,V. and Kaznessis,Y.N. (2008) SynBioSS: the synthetic biology modeling suite. *Bioinformatics*, **24**, 2551–2553.

2. Lubliner,S., Keren,L. and Segal,E. (2013) Sequence features of yeast and human core promoters that are predictive of maximal promoter activity. *Nucleic Acids Res.*, **41**, 5569–5581.

3. 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**, 1–8.