Data-driven modelling of protein synthesis

A sequence perspective

Gritsenko, Alexey

DOI
10.4233/uuid:064f0a35-5d76-42e8-a1ad-3afb5916dd3c

Publication date
2017

Document Version
Final published version

Citation (APA)
Gritsenko, A. (2017). *Data-driven modelling of protein synthesis: A sequence perspective*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:064f0a35-5d76-42e8-a1ad-3afb5916dd3c

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

DATA-DRIVEN MODELLING OF PROTEIN SYNTHESIS

A SEQUENCE PERSPECTIVE

DATA-DRIVEN MODELLING OF PROTEIN SYNTHESIS

A SEQUENCE PERSPECTIVE

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op dinsdag 22 maart 2017 om 15:00 uur

door

Alexey Alexeevich GRITSENKO

Master of Science in Computer Science,
geboren te Barnaul, kraj Altaj, Rusland.

Dit proefschrift is goedgekeurd door de

promotor: Prof. dr. ir. M.J.T. Reinders
promotor: Prof. dr. ir. D. de Ridder

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. M.J.T. Reinders	Technische Universiteit Delft
Prof. dr. ir. D. de Ridder	Wageningen University
Onafhankelijke leden:	
Prof. dr. J. van der Oost	Wageningen University
Prof. dr. J.T. Pronk	Faculteit Technische Natuurwetenschappen, Technische Universiteit Delft
Prof. dr. B. Snel	Universiteit Utrecht
Dr. P.-B.A.C. 't Hoen	Leids Universitair Medisch Centrum
Dr. M. Depken	Faculteit Technische Natuurwetenschappen, Technische Universiteit Delft
Prof. dr. L. Wessels	Nederlands Kanker Instituut, Amsterdam en Technische Universiteit Delft, reservelid

This work was supported by the Platform Green Synthetic Biology (PGSB) consortium and the Kluyver Centre for Genomics of Industrial Fermentation, subsidiaries of the Netherlands Genomics Initiative (NGI). Part of this work was conducted at the Weizmann Institute of Science, Rehovot, Israel.

Keywords: translation, protein synthesis, sequence modelling, sequence analysis, cap-independent translation

Printed by: Ipskamp Printing, Enschede

Front & Back: Cover designed by Cindy dos Santos & Jens Schneider.

ISBN 978-94-028-0559-8

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

I must not fear.
Fear is the mind-killer.
Fear is the little-death that brings total obliteration.
I will face my fear.
I will permit it to pass over me and through me.
And when it has gone past I will turn the inner eye to see its path.
Where the fear has gone there will be nothing. Only I will remain.

Bene Gesserit Litany Against Fear
from *Dune* by Frank Herbert

CONTENTS

SUMMARY

Recent advances in DNA sequencing, synthesis and genetic engineering have enabled the introduction of choice DNA sequences into living cells. This is an exciting prospect for the field of industrial biotechnology, which aims at using microorganisms to produce foods, beverages, pharmaceuticals and fine- and bulk chemicals in a sustainable fashion. Biotechnologists often achieve this by genetically engineering these microorganisms to introduce novel production pathways using genes found in other strains or species. However, detailed understanding of gene expression regulation remains elusive, especially at the level of *translation*; thus, when it comes to writing DNA to express proteins at user-specified levels, we are still miles away.

Second generation DNA sequencing technologies have made it easy and affordable to reconstruct the genomes of industrially relevant microbes, thus providing better reference sequences for genetic engineering. However, technological limitations allow for reconstructing only parts of the entire genomes unambiguously, thus requiring additional *scaffolding* steps to obtain genome-length reconstructions. We propose a method that improves genome scaffolding by integrating heterogeneous sources of information on genome contiguity. These methods improve the quality of genome reconstructions at the cost of a limited number of additional errors.

The ease and affordability of DNA sequencing has also led to the development of a number of biological assays which exploit sequencing, among which the *ribosome profiling* assay. This assay allows for unprecedented examination of the process of protein synthesis by recording positions of actively translating ribosomes across thousands of living cells. We employed these data to develop data-driven models of *Saccharomyces cerevisiae* protein synthesis. A relatively simple model was used to re-design genes for heterologous expression; a second, more complex model yielded insights into the process of translation. Our models suggest that protein synthesis is limited at the stage of initiation, and that codon translation rates are not determined by tRNA levels alone, and appear to be sequence context-dependent.

Finally, the combination of DNA synthesis and sequencing offers the possibility to perform high-throughput *in vivo* assays to study the effect of user-designed sequences. We used this approach to study translation initiation at *Internal Ribosome Entry Sites* (IRESs). We identified short sequence elements predictive of IRES activity in viruses and humans, and obtained insights into the effect of element sequence, multiplicity and position on IRES activity. We propose a high-level architecture of viral and cellular IRESs, and offer a mechanistic explanation for differences between IRES architectures of different virus types.

SAMENVATTING

Recente ontwikkelingen in de genetische modificatie en in het aflezen en synthetiseren van DNA hebben het mogelijk gemaakt om gekozen sequenties in levende cellen te introduceren. Dit levert spannende mogelijkheden op voor de industriële biotechnologie, die tot doel heeft micro-organismen te gebruiken om voeding, dranken, geneesmiddelen en stoffen voor de fijn- en bulkchemie op duurzame wijze te produceren. Biotechnologen beogen dit vaak te bereiken door micro-organismen genetisch te modificeren, om nieuwe productiepaden te introduceren op basis van genen die in andere stammen of species zijn gevonden. Een gedetailleerd begrip van de regulering van genexpressie ontbreekt echter nog, in het bijzonder waar het gaat om *translatie*, en dus zijn we nog ver verwijderd van het schrijven van DNA zodanig dat we eiwitten op gewenste niveaus kunnen produceren.

De tweede generatie van de technologie om DNA af te lezen heeft het makkelijk en betaalbaar gemaakt om genomen van industrieel interessante micro-organismen te reconstrueren en daarmee betere referentiesequenties te krijgen voor genetische modificatie. Technologische beperkingen zorgen er echter voor dat genomen slechts in een aantal delen kunnen worden afgelezen, zodat er nog zogenaamde *scaffolding* ("steigerbouw") plaats moet vinden om sequenties van genoom-lengte te reconstrueren. Wij stellen een methode voor om de scaffolding van een genoom te verbeteren door heterogene informatiebronnen over contiguïteit te integreren. Deze methode verbetert de kwaliteit van genoomreconstructies, ten koste van een klein aantal additionele fouten.

Het gemak en de betaalbaarheid waarmee DNA kan worden afgelezen heeft ook geleid tot de ontwikkeling van een aantal biologische analyses die hier gebruik van maken, waaronder *ribosoomprofilering* (*ribosome profiling*). Deze analyse maakt het mogelijk om het proces van eiwitsynthese in ongekend detail te bestuderen, door de posities van actief translerende ribosomen te meten in duizenden levende cellen. We gebruiken deze meetgegevens om data-gedreven modellen van eiwitsynthese in *Saccharomyces cerevisiae* te ontwikkelen. Een relatief eenvoudig model is gebruikt om genen te herontwerpen voor heterologe expressie; een tweede, meer complex model gaf inzicht in het proces van translatie zelf. Ons model suggereert dat eiwitsynthese gelimiteerd wordt in de initiatiefase, en dat translatiesnelheden van codons niet alleen worden bepaald door tRNA niveaus, maar ook afhankelijk lijken van de sequentiecontext van de codons.

Tenslotte maakt de combinatie van DNA synthese en aflezen het mogelijk om metingen op grote schaal *in vivo* uit te voeren, om het effect van door een gebruiker ontworpen sequenties te meten. Deze aanpak hebben we gebruikt om initiatie van translatie in zogenaamde *Internal Ribosome Entry Sites* (IRESs) te bestuderen. We vonden korte stukken sequentie die IRES activiteit in mensen en virussen voorspellen, en kregen inzicht in het effect van de sequentie, aantal en positie van IRES elementen

op hun activiteit. We stellen een globale architectuur voor van virale en cellulaire IRES elementen, en geven een mechanistische verklaring voor het verschil tussen IRES architecturen in verschillende virustypes.

1

INTRODUCTION

1

Biological research and its bioinformatic challenges are driven by the introduction of new measurement and genetic engineering technologies. During the past decade, advances in DNA reading (DNA sequencing) and writing (DNA synthesis) have resulted in a continuing cost reduction of DNA sequencing and *de novo* DNA synthesis. The sharp decrease in sequencing costs prompted novel methods for interrogating previously inaccessible cellular mechanisms. This is revolutionising biotechnology by providing the tools necessary for production of user-designed proteins at user-specified levels through rational design methodologies. However, when it comes to rational design of synthetic sequences, we still struggle with detemining the exact message to write using these tools.

This thesis supports the ongoing adoption of DNA writing technologies in systems biology and biotechnology research. It describes methods for constructing models of protein synthesis that yield novel insights into the regulation of this mechanism, and could be used to guide rational design of synthetic sequences with desired regulatory properties. It also describes methods for improving genome sequence reconstructions obtained using current DNA sequencing technologies, which facilitates genetic engineering efforts required for downstream expression of designer DNA sequences.

1.1. READING DNA: A DATA-RICH ERA OF BIOLOGICAL SCIENCES

The first genome-scale biological datasets started appearing in the 1990's and 2000's and came from several independent directions: DNA microarrays for measuring relative expression levels and genomic copy number aberrations [1–3]; protein-protein interaction (PPI) measurements [4]; and first generation DNA sequencing [5]. At their prime, these technologies and their variants generated vast amounts of measurement data, and were widely used in research.

Introduction of the first generation sequencing technologies in 1970's [5] marked the beginning of a new age in biology, in which reading DNA sequences of selected genes and entire genomes became possible. This trend culminated with the 19-year long and an estimated $3bn Human Genome Project, which concluded in the early 2000's with the publication of the human genome [6]. But it was not until 2005-2007 [7–11] and the advent of second generation sequencing technologies, which super-exponentially decreased the costs of sequencing by introducing new chemistry and dramatically elevating sequencing instruments' throughput, that the data-rich era truly began. Nowadays, the costs of sequencing an entire human genome are approaching $1k [12]. This 300,000-fold reduction in costs not only turned genome sequencing into an accessible research tool, but generally made DNA sequencing a standard readout mechanism for high-throughput screens and assays. This prompted the development of a plethora of "-seq" counterparts of microarray-based measurements, and novel applications.

Today, genome-scale assays enabled by high-throughput sequencing can be found in virtually all areas of functional genomics (see Soon *et al.* [13], Pachter [14] for an extensive list). They include measurements of RNA levels (RNA-seq, Mortazavi *et al.* [15]); examination of mRNA alternative polyadenylation sequences (3′-seq, Lianoglou

et al. [16]); analysis of protein-DNA, protein-RNA and RNA-RNA interactions (ChIP-seq, iCLIP-seq and CLASH-seq, Johnson *et al.* [17], König *et al.* [18], Helwak *et al.* [19]); measurements of chromatin structure and accessibility (e.g., Hi-C and ATAC-seq, Lieberman-Aiden *et al.* [20], Buenrostro *et al.* [21]); determination of genome replication order (Repli-seq, Hansen *et al.* [22]); measurements of RNA structure (PARS-seq, [23]); measurements of locations of actively translating ribosomes (ribo-seq, Ingolia *et al.* [24]); and many more, including a growing number of single-cell analyses [25–27]. Second generation sequencing has also been used in combination with genome editing techniques to devise high-throughput screens for studying the architecture of transcriptional and translational regulation [28–31].

DNA SEQUENCING TECHNOLOGIES

Sanger sequencing, the first generation of sequencing technologies, is based on the chain-termination method [5]. It is characterised by low-throughput and the ability to "read" relatively large DNA molecules. Modern Sanger sequencing generates reads of 400 – 900 bases [32], which are suitable for *de novo* sequencing of small DNA molecules. However its low-throughput makes Sanger sequencing prohibitively expensive for most other applications. This shortcoming was addressed approximately 30 years later with the independent introduction of several second generation technologies [8, 10, 33–35] characterised by massively parallel sequencing through DNA synthesis. Out of these technologies, Illumina is currently the most widely used sequencing platform. Although its first instruments generated reads of only 35 bases, its modern chemistry is characterised by read lengths of up to 300 bases and the lowest cost per base in its class [32], which makes it the method of choice for high-throughput assays and screens.

The field of DNA sequencing is currently experiencing the rise of another, third, generation of sequencing technologies. The 3rd generation sequencing platforms are characterised by real-time single-molecule sequencing, and, with their current chemistry, produce high-error rate reads that are tens of thousands of bases long [36, 37]. Given their read lengths, 3rd generation platforms are particularly suited for *de novo* genome sequencing [38, 39].

COMPUTATIONAL CHALLENGES OF DNA SEQUENCING

Hand in hand with the introduction of DNA sequencing came the first computational challenges of reconstructing genomes from sequenced DNA fragments, and of comparing genomes and sequences to each other. As sequencing technologies evolved, and second and third generations of DNA sequencing machines became more mature, the computational challenges have also changed [40–43]. This thesis only describes computational challenges specific to the second generation sequencing technologies, as they were most prominent at the time when thesis work was carried out.

Whole-genome sequencing (WGS) is an important tool in genome engineering. First, availability of a high quality reference genome is a prerequisite for most genome editing efforts; and second, sequenced genomes can be used to learn what sequences we should write to achieve desired phenotypes. For these reasons, WGS is often employed to obtain reference genome sequences of organisms employed in biotechnology and industrial microbiology.

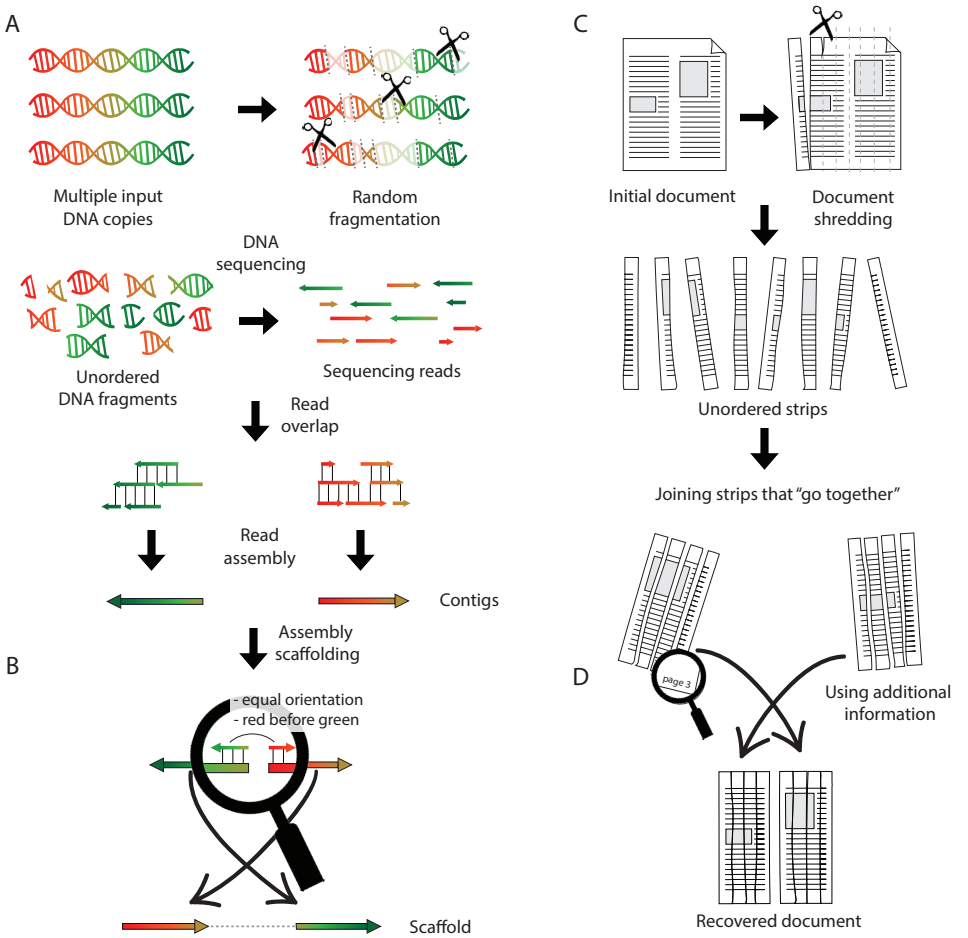

Figure 1.1: A schematic representation of shotgun sequencing, genome assembly and scaffolding. (**A**) Multiple copies of the genome are randomly fragmented to create short DNA fragments that can be read by sequencing machines. Some fragments (bleak) are lost in this process. Fragmentation causes the order and orientation relationships between fragments to be lost. Once sequenced, the redundant information from fragmenting multiple identical genome copies is used to reconstruct the original genome sequence from overlapping reads. However, due to the fragments lost during fragmentation, only parts of the genome (contigs) can be reconstructed. (**B**) The problem of recovering shredded documents from unordered strips is a helpful analogy for understanding genome reconstruction. It can be solved by unambiguously joining those strips that go together. In this example only parts of the document corresponding to each of the two columns can be recovered because column order is lost during shredding. (**C**) To improve the assembly, contigs can be further joined into longer gapped scaffolds. This requires the use of additional information on the relative contig order, orientation and distance constraints, which often comes from paired reads (inside the magnifying glass). Optimisation of the contig order, distance and orientation to satisfy these constraints produces the sought scaffolds. (**D**) In the document shredding analogy additional information, such as the position of page numbers, can be used to correctly re-order the recovered document parts.

GENOME ASSEMBLY

WGS aims at reading the entire genome of an organism, i.e., all the molecules (chromosomes or plasmids) that it carries. However, because most (first and second generation) sequencing technologies can only read sequences that are substantially shorter than the millions and billions of nucleotides composing microbial and mammalian genomes, a method called *shotgun sequencing* is often employed to sequence longer fragments [44]. In shotgun sequencing many copies of the genome are randomly fragmented into smaller molecules, which can be (partially) read, as shown in Fig. 1.1A. When the genome is fragmented, all information regarding the location and strand of the genome, from which fragments originate, is lost. So once these fragments are read, the resulting reads need be put together to form the original genome like strips of a shredded document (see Fig. 1.1B). Such a document could be reconstructed by joining strips that "go together". The shredder model is illustrative for the problem of genome reconstruction, where reads need to be joined into longer sequences to form the genome. However, because in shotgun sequencing multiple genome copies are fragmented simultaneously, one can decide whether two reads belong together based on their sequence overlap. The process of repeated joining of overlapping sequences and reads into longer contiguous sequences (contigs) forms the basis of *genome assembly* [45].

ASSEMBLY SCAFFOLDING

Unfortunately, due to repeats in the genome (identical sentences in a shredded document), read errors (unreadable letters on the shredded strips) and uneven genome coverage (lost strips), the read extension process inevitably becomes ambiguous and cannot continue indefinitely. Contigs resulting from the assembly step may belong to one or more chromosomes, can come from any of the two strands of the genome, and may not even cover the entire genome. To improve the assembly further, so-called scaffolds may be constructed by joining contigs from the same DNA strand into longer (gapped) sequences in the correct order in a process called *scaffolding* [45]. This process relies on additional information about contig order, distance and orientation (whether two contigs come from the same DNA strand, or opposite strands), which would allow extending contigs beyond ambiguities encountered in the assembly step (see Fig. 1.1C and D).

Additional information for scaffolding can be obtained from a variety of sources. For example, *paired end* and *mate pair* information on read pairs can be used, which provides relative orientation and approximate distance for pairs of reads originating from the same piece of fragmented DNA [41, 46]. Read pairs are a particularly popular source of additional scaffolding information because they can easily be generated with standard DNA sequencing protocols. However, information from related genomes, or restriction maps, can also be used [47, 48]. The scaffolding problem is particularly important for genomes assembled from shorter reads generated by the 2nd generation sequencing technologies, as they yield highly fragmented assemblies [49].

Due to the relatively low complexity of microbial genomes, second generation sequencing quickly became the technology of choice for *de novo* microbial sequencing. However, its adoption further aggravated the challenge of improving the resulting

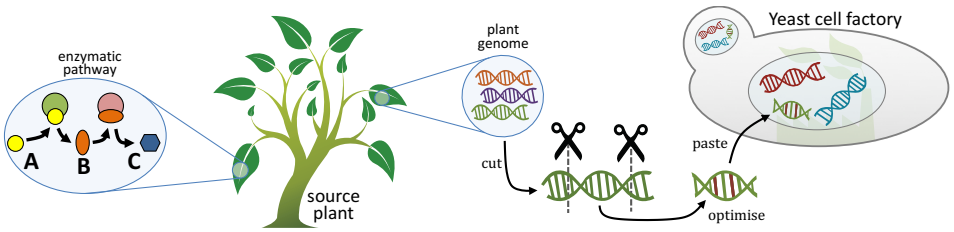

Figure 1.2: Schematic example of a plant enzyme introduced into a yeast genome. One of the plant genes (green; left) involved in the enzymatic conversion of A → B → C is isolated from the plant genome (green chromosome; centre), cut out using "molecular scissors", optimised for expression in yeast and "pasted" into the yeast genome (right).

fragmented short-read assemblies. Our involvement in the *de novo* sequencing of the *Saccharomyces cerevisiae* CEN.PK 113-7D, a laboratory yeast strain commonly used in industrial biotechnology research [50], prompted us to develop GRASS, one of the first approaches for scaffolding such assemblies. As described in Chapter 2, GRASS is a generic assembly scaffolder based on a computational model, that can integrate any type of scaffolding information, and is combined with an efficient optimisation strategy. Since the publication of our approach, a number of assembly and scaffolding algorithms have been proposed [51–57], with different underlying models, assumptions and optimisation strategies; some allowing to combine different types of scaffolding information. However, to our knowledge, in 2010 few standalone scaffolders were available that could make use of any type of scaffolding information.

1.2. WRITING DNA: A NEW FRONTIER

Independent from the introduction of Sanger sequencing in 1970s, important advances were made in recombinant DNA technology [58]. Sequence-specific DNA cutting using restriction enzymes, commercial DNA synthesis, *in vitro* DNA amplification and the first transgenic organisms were introduced at that time [59–62], and formed the foothold of modern biotechnology and synthetic biology. Recent advances in nuclease-based genetic engineering technologies (ZFNs, TALENs and CRISPR-Cas; Gilles and Averof [63]), and continuously decreasing costs of DNA synthesis [64] have allowed for *writing* DNA sequences spanning entire chromosomes [65].

MICROBIAL CELL FACTORIES

Biotechnology has a long history of using microorganisms for sustainable production of foods, pharmaceuticals, fuels and fine and bulk chemicals. Some famous examples of using microbes for their natural products include the use of fungus *Penicillium rubens* for production of antibiotic penicillin [66, 67], the use of yeast for beverage fermentation and the use lactic acid bacteria for cheese production [68, 69]. However, modern biotechnology is also used to engineer organisms for production of proteins and chemicals that they normally cannot synthesise. One of the first applications of recombinant DNA technology for this purpose dates back to 1978, when an *Escherichia*

coli containing the human insulin gene was engineered [70]. Nowadays, genetic engineering in biotechnology is used at a much larger scale to create *microbial cell factories*, which go through several enzymatic steps before producing the target chemical. This is achieved by introducing entire chemical pathways, often from higher eukaryotes, into the host microbes [71, 72] as illustrated in an example in Fig. 1.2.

The aim of microbial cell factory engineering is to make production of chemicals of interest sustainable and accessible, which would ultimately contribute to increasing quality of life and to economic growth. For example, recently, artemisinic acid, a precursor of a highly effective anti-malarian drug artemisinin, was produced at an industrial scale using genetically engineered yeast. Such production has the potential to substantially reduce the cost of artemisinin and make it available to people who need it the most [73, 74]. However, this milestone required almost a decade of strain engineering to make the production cost-effective and scalable.

The stage of improving cell factories for yield or robustness is common to production process engineering. It is usually accomplished by *metabolic engineering*, i.e., optimising cellular processes through genetic modification to increase production of a target substance. Genetic modifications in metabolic engineering can be introduced through laboratory evolution, random mutagenesis or by means of rational design. The latter often includes adjusting expression levels of pathway enzymes by replacing their promoters, modifying their genomic copy numbers or by changing coding sequences of those enzymes to increase their translation rates. This can be achieved by introducing recombinant or, when possible, synthetic DNA sequences into the microbial factories.

DATA-DRIVEN MODELS

Despite advances in DNA synthesis, the use of synthetic DNA in biotechnology and synthetic biology remains limited. Regardless of our ability to write DNA, we often do not know *what exactly to write*, as determining the sequence of synthetic DNA that would exert the desired regulatory effect (a version of the genotype-phenotype mapping problem) remains a challenging task.

High-throughput assays and screens generate data at a pace previously uncharacteristic for biology, which allows for employing modelling approaches from Statistics and machine learning (ML) for their analyses. These approaches have a long history of solving data-rich problems, and when applied to biological problems, can be used to construct predictive genotype-phenotype models (e.g., predicting promoter strength from its sequence; Lubliner *et al.* [75]) for guiding rational design of synthetic DNA sequences.

CLASSIFICATION AND REGRESSION

Classification and regression are supervised ML techniques that are used for assigning class labels (classification) or numeric values (regression) to objects based on their features [76]. They rely on constructing models (classifiers or regressors) based on a training set of objects with known labels, which are used to learn the unknown relationships between object features and labels. Ultimately, the trained models are used for predicting labels of new objects, and are *interpreted* to uncover object features most predictive of the labels.

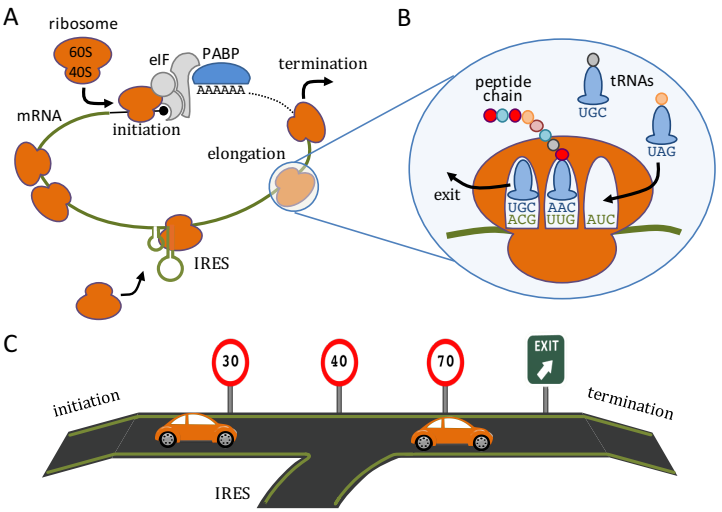

Figure 1.3: A high-level overview of translation and the "cars on a road" mental model helpful in understanding of this process. (**A**) Classically, eukaryotic translation requires capped (black filled circle) and polyadenylated mRNA (black A's) to be circularised through the interaction of the eukaryotic initiation factors (eIFs; grey), the poly-A tail and the poly(A)-binding protein (PABP; blue), which prompts the recruitment of the ribosome (orange) through its 40S subunit. Recruited ribosomes bind to the 5′ untranslated region (UTR) of the mRNA (black solid line) and move in the direction of the 3′ end until they encounter the protein-coding region (thick green line) and initiate translation. Ribosomes synthesise the encoded protein during the elongation phase, and terminate once they reach the end of the protein-coding region (black dashed line). In some cases ribosomes can be recruited to regions other than the 5′ UTR through internal ribosome entry sites (IRESs; clover leaf structure shown in green). (**B**) In the elongation phase the ribosome repeatedly grows the peptide chain one amino acid at a time (coloured circles) by matching codons (ribonucleotide triplets) on the mRNA against their complementary anticodons on the tRNAs (light blue). (**C**) For simplicity one can think of ribosomes attached to mRNA as cars entering a road (initiation), driving on it (elongation) and exiting it (termination). Speed limits for regions of such road would correspond to various codons and their translation speeds, whereas ramps merging into the main road would be equivalent to translation initiation via an IRES.

These techniques enjoyed successful applications in numerous fields [77–79], including biology [80], where they have also been applied to the analysis of high-throughput functional genomics data. For example, ML allowed for recognising transcription start sites, promoters, enhancers, splice sites and for determining nucleosome positioning [81–84]. Additionally, it helped to gain insight into the molecular mechanisms involving these genetic elements and processes. In Chapters 3 and 5 of this thesis we too employ machine learning techniques to analyse high-throughput functional genomics data, and to improve our understanding of the molecular mechanisms of protein synthesis.

THE PROCESS OF PROTEIN SYNTHESIS

Protein synthesis is a crucial cellular mechanism, entailing the translation of DNA-encoded genetic information into biomolecules central to virtually all cellular processes - proteins. A holistic understanding of protein synthesis has fundamental

scientific significance. It also finds important applications in health and disease and biotechnology, such as understanding the effects of synonymous mutations on high-level disease phenotypes [85], or solving the problem of gene optimisation for expression in a new host [86]. Despite its centrality and a relatively old age, translation regulation is a still field of active research with ongoing debates about determinants of translation initiation and elongation.

A detailed description of eukaryotic translation can be found in Hinnebusch and Lorsch [87] and Dever and Green [88]. In a simplistic view, the three-phase process of translation begins with the recruitment of ribosomes at the 5′ end of capped and circularised mRNA molecules. Recruitment of the ribosome to the 5′ untranslated region of the mRNA involves the interaction of several eukaryotic translation initiation factors, the Poly(A)-binding protein and the 40S ribosomal subunit, which form a complex together. Once assembled at the 5′ of the mRNA, the complex starts scanning it in the direction of the 3′ end for the start of the protein-coding region to recruit the 60S ribosomal subunit, finalise ribosome assembly, and initiate translation (see Fig. 1.3A). Next, in the elongation phase, the ribosome repeatedly grows the nascent peptide chain by decoding each codon it encounters using a suitable aminoacyl-tRNA molecule as an adapter, and adding the corresponding amino acid to the growing chain (see Fig. 1.3B). This process stops once the ribosome encounters a stop codon and the peptide chain is released to fold into its three-dimensional conformation and become a protein. Leaving biological complexity aside for a moment, one can think of ribosomes translating an mRNA as cars on a single-lane road with consecutive regions and speed regimes of this road corresponding to codons with their specific elongation rates (see Fig. 1.3C). In this analogy, translation initiation and termination are equivalent to entering and exiting the road.

However, despite the relative simplicity of this above process, the exact mechanistic details of its individual steps remain largely unknown, including the exact rates of translation initiation or elongation, and their RNA sequence determinants. Owing to this knowledge gap and the difficulty of measuring rate parameters directly, existing computational models of translation often make significant simplifying assumptions about the process of translation [89–92].

CODON OPTIMISATION

It is generally accepted that synonymous codons, i.e., codons translated to the same amino acid, are translated at different rates. Moreover, these rates, believed to be determined mainly by the abundance of tRNAs recognising them [93, 94], may differ between organisms, as does tRNA abundance. For microbial cell factories this means that a gene that is efficiently translated in one organism may be translated slowly in its new host. For this reason genes are often “recoded” prior to synthesis and heterologous expression in a way that would maximise their translation rate, but retains the original amino acid sequence (see Fig. 1.2). This process is called *codon optimisation*.

Despite being commonly used, codon optimisation (CO) remains largely an empirical technique due to the limited understanding of the mechanistic details of the process it optimises. Consequently, it is reported to increase protein expression of an optimised gene in some cases; and to have no effect on expression, or to reduce protein

solubility or enzymatic activity in others [95–97]. Despite the complex and multifactorial nature of mechanisms of translation regulation [98], virtually all existing CO approaches focus on a single aspect of the optimised sequence. Typically, the extent to which codon usage of the optimised gene matches that of a reference set of the hosts highly expressed genes, thought to be efficiently translated, is minimised [99–101]. The latter is often quantified using the Codon Adaptation Index (CAI, Sharp and Li [102]) or a similar *ad hoc* measure [103–105].

In Chapter 3 we introduce a data-driven codon optimisation approach that does not explicitly model the process of translation, but rather attempts to capture features predictive of efficient translation using ML. In our approach, instead of arbitrarily choosing a single aspect or measure for optimising the sequence, we employ regression to learn the relationship between multiple sequence features and its total protein production from *ribosome profiling* measurements [24, 106] for native *Saccharomyces cerevisiae* genes. We then use the learned model to navigate the space of possible optimised sequences and choose the one that maximises model prediction. Unfortunately, in a follow-up experimental validation of our approach we discovered that it improved enzyme activity of an optimised synthetic test gene relative to its wild type to a lesser extent than a CAI-based method did, suggesting that our approach was unable to fully capture sequence determinants of translation efficiency. We briefly describe the experimental validation procedures and potential issues of our approach in an *addendum* to Chapter 3.

RIBOSOME PROFILING

Recently, a new high-throughput measurement technique, called *ribosome profiling*, was proposed [24, 106]. It allows for previously unavailable genome-wide measurements of the exact locations of actively translating ribosomes *in vivo*. The core of ribosome profiling consists of (i) the ribo-seq high-throughput assay, which measures positions of translating ribosomes; and (ii) RNA-seq used for measuring mRNA transcript abundances.

Briefly, through the addition of the chemical cycloheximide and the use of low temperatures, ribo-seq achieves a situation when ribosomes are frozen in place on the transcripts that they were translating. Transcripts with bound ribosomes are then digested, leaving only 28nt – 30nt fragments bound by the ribosomes, which can be reverse-transcribed and sequenced using second generation sequencing technologies. When mapped back the genome, sequenced reads yield a snapshot of locations of actively translating ribosomes from many cells. Mapped reads also yield ribosome density profiles for every translated gene. Density changes along the profiles can be interpreted as changes in local elongation speed, where slower and faster regions respectively have higher and lower *normalised* density (see Fig. 1.4A; busy and free). Continuing the analogy of cars on a road, ribosome profiling essentially yields a view of how busy roads are, akin to modern navigation software (see Fig. 1.4B). These data were used to study differences in translation efficiency between yeast species [107], to provide evidence of short peptide translation in $5'$ untranslated regions [24], to demonstrate prevalence of stop codon read-through in *Drosophila melanogaster* [108], to study ribosome pausing [109–113], to derive yeast codon elongation rates [111, 114] and even to construct whole-genome models of protein translation [89, 92].

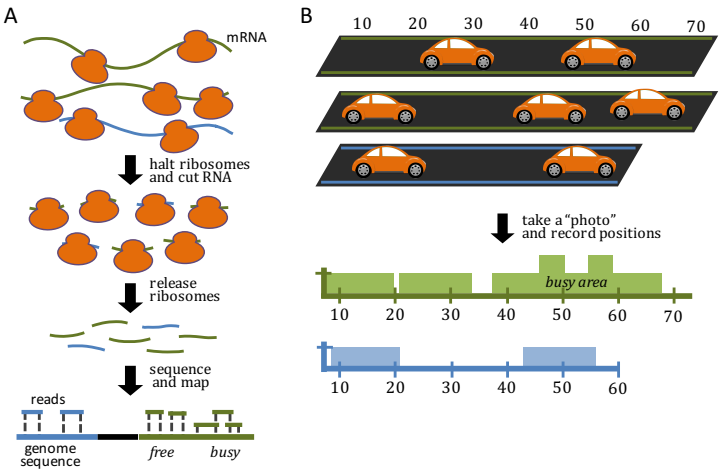

Figure 1.4: A brief outline of ribo-seq high-throughput assay employed in ribosome profiling. (**A**) To determine the locations of actively translating ribosomes (orange), they are first halted and cross-linked to mRNA transcripts they are translating (green and blue strings). The transcripts are then subjected to nuclease treatment, which removes all RNA that is not protected by bound ribosomes. After releasing the ribosomes the remaining footprints (short green and blue strings) can be sequenced using high-throughput sequencing, and mapped to the genome to recover ribosome positions and determine fast (sparse) and slow (dense) translation regions. (**B**) In the framework of the "cars on a road" analogy introduced earlier, performing ribo-seq is equivalent to taking photographs of roads (halting the ribosomes), calculating car positions on the photos, and accumulating position information across several photos of the same road (footprint mapping) to determine busy areas.

Whole-genome modelling of protein translation is a computationally challenging task. To facilitate it, existing approaches [89, 92] either assumed that codon elongation rates are known, and used ribosome profiling data only to find gene-specific initiation rates; or neglected situations when one ribosome would block elongation of another ribosome on the same transcript (ribosome queueing). In Chapter 4 we propose a modelling framework that combines strengths of existing models, while making no *a priori* assumptions about model parameters (elongation and initiation rates). Our framework unites an explicit ribosome movement model, that supports ribosome queueing, with a data-driven approach to find its parameters by fitting model simulations on to the ribosome profiling data. Or, using the cars on a road analogy, our approach aims at learning what the speed limits on roads are without knowing how to read the speed limit signs written in a foreign language (see Fig. 1.3), just by looking at how busy the roads are on average.

CAP-INDEPENDENT TRANSLATION INITIATION

One notable exception to the described simplistic view of the protein synthesis process (Section 1.2) is translation initiation that does not require the 5′ mRNA cap structure, and can directly recruit ribosomes to inner regions of the mRNA. RNA elements responsible for this mechanism of initiation are called *Internal Ribosome Entry Sites* (IRESs; see

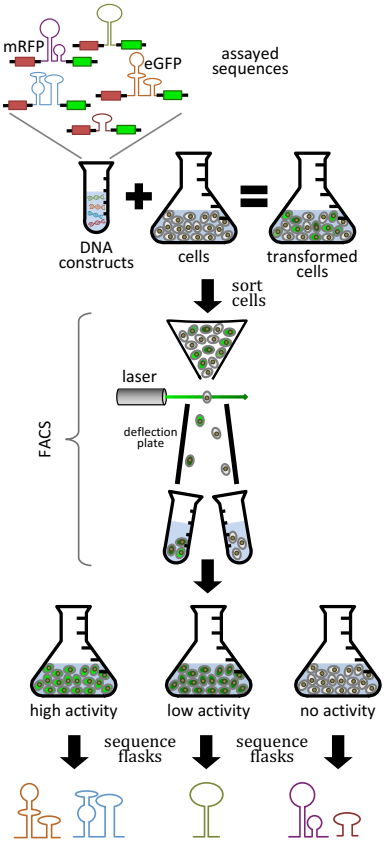

Figure 1.5: Simplified outline of the IRES activity high-throughput assay from Weingarten-Gabbay *et al.* [31]. Assayed sequences (RNA loops; dark green, purple, blue, orange and red) are inserted in between the mCherry red fluorescent protein (mRFP, red) and the enhanced green fluorescent protein (eGFP, green) so that if an assayed sequence can function as an IRES, the eGFP protein will be produced and will turn the cells green. These DNA constructs are integrated into cells to obtain transformed cells that contain at most one construct. A fluorescence activated cell sorter (FACS) is used to iteratively separate transformed cells into populations that have a varying amounts of eGFP (high IRES activity, low IRES activity, and no IRES activity). Sequences that are responsible for these activities are then read out by sequencing each of the three populations individually.

Fig. 1.3A). In the "cars on a road" analogy (Fig. 1.3C), IRES-mediated ribosome recruiting can be viewed as cars joining the main road through a ramp connecting to it in the middle. IRESs were first discovered in positive-sense ssRNA viruses [115, 116]. Some of these viruses spend their entire replication cycle in the hosts cytoplasm and do not poses the 5′ cap, since capping occurs co-transcriptionally in the nucleus [117], and hence their translation has to occur in a cap-independent manner. Since then IRESs were discovered in other viruses, including HCV and HIV [118, 119], and in cellular transcripts [120]; and implicated in cell apoptosis and stress response [121]. The study of IRES-mediated translation regulation can have significant ramifications for

understanding and treatment of numerous diseases [122–124].

Unfortunately, relatively little is known about the mechanisms governing IRES-mediated ribosome recruitment [125]. This has to do with an apparent lack of common sequence or structure motifs shared by known cellular IRESs [120]. And, because experimental validation of potential IRESs is cumbersome and requires the use of bicistronic report constructs and multiple control experiments to rule out expression due to cryptic splicing or promoter activity [126], relatively few IRESs were known until recently (≈ 120 according to IRESite [127]; accessed on December 1, 2015). To alleviate this situation, in Weingarten-Gabbay *et al.* [31] we designed and performed a high-throughput IRES activity screen that is used to measure IRES activity for tens of thousands short sequences and increased the number of known IRESs 50-fold. In Fig. 1.5 we provide a shortened description of this screen, a complete description can be found in the original publication.

In Chapter 5 we describe work to exploit high-throughput IRES activity measurement data to uncover sequence determinants of IRES activity. We achieve this by constructing data-driven regression models that learn the relationship between IRES activity and RNA sequence and structure, and by interpreting the learned models afterwards.

1

1.3. CONTRIBUTIONS OF THIS THESIS

This thesis contributes to the ongoing transition from DNA reading to DNA writing philosophy in biotechnology and systems biology. First, by developing computational algorithms for scaffolding microbial genome assemblies to facilitate reconstruction of the reference genome sequences for downstream microbial host engineering, in Chapter 2. And second, by constructing data-driven models for understanding regulation of classical translation initiation and elongation in Chapter 4, and of cap-independent translation initiation in Chapter 5. In addition to providing novel insights into the corresponding regulatory mechanisms, these models have potential applications in guiding design of synthetic sequences for metabolic engineering efforts, cf. the methods discussed in Chapter 3.

In Chapter 2 we describe GRASS, an algorithm for improving genome assemblies through scaffolding. It facilitates obtaining high-quality reference genome sequences, a prerequisite for genetic engineering efforts. GRASS relies on a novel computational model, which combines the goals of finding the correct order, orientation and positions of assembled contigs in an intuitive way. This allows it to use a variety of information sources for constructing long high-quality scaffolds, which we demonstrated by applying it to short-read second generation sequencing assemblies of three bacterial genomes in situations when multiple sequencing datasets or related genomes were available.

Our venture into systems biology and algorithms for writing DNA sequences began with the challenge of optimising production of naringenin in recombinant yeast, which we sought to achieve by maximising expression of individual enzymes in the naringenin biosynthesis pathway by means of codon optimisation. In Chapter 3 we describe a simple data-driven approach for codon optimisation based on predicting the total protein production of a gene from its sequence. We used it to optimise genes from the naringenin biosynthesis pathway genes from the plant *Arabidopsis thaliana* [128] for expression in *Saccharomyces cerevisiae*. In a later experimental validation of one of the optimised genes, we discovered that it improved protein expression, albeit to a lesser extent than a traditional method did. We describe the experimental validation procedure in an *addendum* to Chapter 3, where we also discuss possible improvements of our codon optimisation strategy.

Having learned about the complexities of translation regulation and limitations of our codon optimisation approach through validation experiments, we sought to devise a whole-cell model of translation that would overcome these limitations and explicitly model the physical processes of translation initiation and elongation, while also learning model parameters from data. In Chapter 4 we present an approach for deriving data-driven models of translation from ribosome profiling measurements. In this work we developed an efficient simulation method for the physical process of translation, a framework for analysing ribosome profiling data, and an overall computational framework for fitting translation models on to this data. We applied this approach to learn models of *Saccharomyces cerevisiae* translation, which were used to study this process in the context of its rate-limiting steps, robustness to changes in codon elongation rates and in the context of codon optimisation. Our models indicated that codon elongation rates often deviate from values dictated by tRNA levels alone, suggesting that other factors are involved in determining these rates.

Finally, in Chapter 5 we describe a study of an alternative mechanism of translation initiation via the Internal Ribosome Entry Sites (IRESs), in which we developed sequence models of IRES activity using machine learning methods. Interpretation of the learned models highlighted similarities and differences between IRESs from different species and viral classes. Together, our models yield an high-level architecture of IRESs that suggests optimal mRNA binding site positions of IRES *trans*-acting factors (ITAFs), proteins involved in IRES-mediated translation initiation [129].

Overall, this thesis contributes to several aspects of cell factory engineering through (i) methods and analyses that improve our understanding of the process of translation regulation, and (ii) a method for improving genome assemblies. These two major contributions set the stage for further systems biology research and its applications in metabolic engineering through synthetic DNA design.

1

REFERENCES

[1] A. C. Pease, D. Solas, E. J. Sullivan, M. T. Cronin, C. P. Holmes, and S. Fodor, *Light-generated oligonucleotide arrays for rapid DNA sequence analysis,* Proceedings of the National Academy of Sciences **91**, 5022 (1994).

[2] D. Shalon, S. J. Smith, and P. O. Brown, *A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization.* Genome Research **6**, 639 (1996).

[3] S. Solinas-Toldo, S. Lampel, S. Stilgenbauer, J. Nickolenko, A. Benner, H. Döhner, T. Cremer, and P. Lichter, *Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances,* Genes, Chromosomes and Cancer **20**, 399 (1997).

[4] S. Fields and O. Song, *A novel genetic system to detect protein protein interactions,* Nature (1989).

[5] F. Sanger, *Determination of nucleotide sequences in DNA,* Bioscience Reports **1**, 3 (1981).

[6] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, *et al.*, *The sequence of the human genome,* Science **291**, 1304 (2001).

[7] S. Bennett, *Solexa ltd,* Pharmacogenomics **5**, 433 (2004).

[8] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, *et al.*, *Genome sequencing in microfabricated high-density picolitre reactors,* Nature **437**, 376 (2005).

[9] M. L. Metzker, *Emerging technologies in DNA sequencing,* Genome Research **15**, 1767 (2005).

[10] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, *et al.*, *Accurate whole human genome sequencing using reversible terminator chemistry,* Nature **456**, 53 (2008).

[11] R. Bumgarner, *Overview of DNA microarrays: types, applications, and their future,* Current Protocols in Molecular Biology , 22 (2013).

[12] K. A. Wetterstrand, *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP),* Online: `http://www.genome.gov/sequencingcosts/` (2013), updated: Jan 15, 2016.

[13] W. W. Soon, M. Hariharan, and M. P. Snyder, *High-throughput sequencing for biology and medicine,* Molecular Systems Biology **9**, 640 (2013).

[14] L. Pachter, *Bits of DNA: *Seq,* Online: `https://liorpachter.wordpress.com/seq/` (2013), updated: Nov 23, 2013.

[15] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, *Mapping and quantifying mammalian transcriptomes by RNA-Seq,* Nature Methods **5**, 621 (2008).

[16] S. Lianoglou, V. Garg, J. L. Yang, C. S. Leslie, and C. Mayr, *Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression,* Genes & Development **27**, 2380 (2013).

[17] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, *Genome-wide mapping of in vivo protein-DNA interactions,* Science **316**, 1497 (2007).

[18] J. König, K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D. J. Turner, N. M. Luscombe, and J. Ule, *iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution,* Nature Structural & Molecular Biology **17**, 909 (2010).

[19] A. Helwak, G. Kudla, T. Dudnakova, and D. Tollervey, *Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding,* Cell **153**, 654 (2013).

[20] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, *et al.*, *Comprehensive mapping of long-range interactions reveals folding principles of the human genome,* Science **326**, 289 (2009).

[21] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, *Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position,* Nature Methods **10**, 1213 (2013).

[22] R. S. Hansen, S. Thomas, R. Sandstrom, T. K. Canfield, R. E. Thurman, M. Weaver, M. O. Dorschner, S. M. Gartler, and J. A. Stamatoyannopoulos, *Sequencing newly replicated DNA reveals widespread plasticity in human replication timing,* Proceedings of the National Academy of Sciences **107**, 139 (2010).

[23] M. Kertesz, Y. Wan, E. Mazor, J. L. Rinn, R. C. Nutter, H. Y. Chang, and E. Segal, *Genome-wide measurement of RNA secondary structure in yeast,* Nature **467**, 103 (2010).

[24] N. T. Ingolia, S. Ghaemmaghami, J. R. Newman, and J. S. Weissman, *Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling,* Science **324**, 218 (2009).

[25] T. Nagano, Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E. D. Laue, A. Tanay, and P. Fraser, *Single-cell Hi-C reveals cell-to-cell variability in chromosome structure,* Nature **502**, 59 (2013).

[26] J. D. Buenrostro, B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, and W. J. Greenleaf, *Single-cell chromatin accessibility reveals principles of regulatory variation,* Nature **523**, 486 (2015).

[27] D. Grün, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, and A. van Oudenaarden, *Single-cell messenger RNA sequencing reveals rare intestinal cell types,* Nature **525**, 251 (2015).

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

[29] C. D. Arnold, D. Gerlach, C. Stelzer, Ł. M. Boryń, M. Rath, and A. Stark, *Genome-wide quantitative enhancer activity maps identified by STARR-seq,* Science **339**, 1074 (2013).

[30] J. A. Brophy and C. A. Voigt, *Antisense transcription as a tool to tune gene expression,* Molecular Systems Biology **12**, 854 (2016).

[31] S. Weingarten-Gabbay, S. Elias-Kirma, R. Nir, A. A. Gritsenko, N. Stern-Ginossar, Z. Yakhini, A. Weinberger, and E. Segal, *Systematic discovery of cap-independent translation sequences in human and viral genomes,* Science **351**, aad4939 (2016).

[32] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law, *Comparison of next-generation sequencing systems,* BioMed Research International **2012** (2012).

[33] V. Pandey, R. C. Nutter, and E. Prediger, *Applied biosystems solid™system: ligation-based sequencing,* in *Next Generation Genome Sequencing: Towards*

Personalized Medicine, edited by M. Janitz (Wiley Online Library, 2008) pp. 29–42.

[34] J. M. Rothberg, W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, J. H. Leamon, K. Johnson, M. J. Milgrew, M. Edwards, *et al.*, *An integrated semiconductor device enabling non-optical genome sequencing,* Nature **475**, 348 (2011).

[35] E. L. van Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes, *Ten years of next-generation sequencing technology,* Trends in Genetics **30**, 418 (2014).

[36] M. G. Ross, C. Russ, M. Costello, A. Hollinger, N. J. Lennon, R. Hegarty, C. Nusbaum, and D. B. Jaffe, *Characterizing and measuring bias in sequence data,* Genome Biology **14**, R51 (2013).

[37] T. Laver, J. Harrison, P. O'Neill, K. Moore, A. Farbos, K. Paszkiewicz, and D. Studholme, *Assessing the performance of the Oxford Nanopore Technologies MinION,* Biomolecular Detection and Quantification **3**, 1 (2015).

[38] K. B. Stadermann, B. Weisshaar, and D. Holtgräwe, *SMRT sequencing only de novo assembly of the sugar beet (Beta vulgaris) chloroplast genome,* BMC Bioinformatics **16**, 295 (2015).

[39] L. Faino, M. F. Seidl, E. Datema, G. C. van den Berg, A. Janssen, A. H. Wittenberg, and B. P. Thomma, *Single-Molecule Real-Time Sequencing Combined with Optical Mapping Yields Completely Finished Fungal Genome,* mBio **6**, e00936 (2015).

[40] J. R. Miller, S. Koren, and G. Sutton, *Assembly algorithms for next-generation sequencing data,* Genomics **95**, 315 (2010).

[41] D. Earl, K. Bradnam, J. S. John, A. Darling, D. Lin, J. Fass, H. O. K. Yu, V. Buffalo, D. R. Zerbino, M. Diekhans, *et al.*, *Assemblathon 1: a competitive assessment of de novo short read assembly methods,* Genome Research **21**, 2224 (2011).

[42] N. Nagarajan and M. Pop, *Sequence assembly demystified,* Nature Reviews Genetics **14**, 157 (2013).

[43] C.-S. Chin, D. H. Alexander, P. Marks, A. A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E. E. Eichler, *et al.*, *Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data,* Nature Methods **10**, 563 (2013).

[44] R. Staden, *A strategy of DNA sequencing employing computer programs,* Nucleic Acids Research **6**, 2601 (1979).

[45] M. C. Schatz, A. L. Delcher, and S. L. Salzberg, *Assembly of large genomes using second-generation sequencing,* Genome Research **20**, 1165 (2010).

[46] S. van Heesch, W. P. Kloosterman, N. Lansu, F.-P. Ruzius, E. Levandowsky, C. C. Lee, S. Zhou, S. Goldstein, D. C. Schwartz, T. T. Harkins, *et al.*, *Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing,* BMC Genomics **14**, 257 (2013).

[47] S. Saha and S. Rajasekaran, *Efficient and scalable scaffolding using optical restriction maps,* BMC Genomics **15**, S5 (2014).

[48] E. Bao, T. Jiang, and T. Girke, *AlignGraph: algorithm for secondary de novo genome assembly guided by closely related references,* Bioinformatics **30**, i319 (2014).

[49] O. Morozova and M. A. Marra, *Applications of next-generation sequencing technologies in functional genomics,* Genomics **92**, 255 (2008).

[50] J. F. Nijkamp, M. van den Broek, E. Datema, S. de Kok, L. Bosman, M. A. Luttik, P. Daran-Lapujade, W. Vongsangnak, J. Nielsen, W. H. Heijne, *et al.*, *De novo sequencing, assembly and analysis of the genome of the laboratory strain Saccharomyces cerevisiae CEN.PK113-7D, a model for modern industrial biotechnology,* Microbial Cell Factories **11**, 1 (2012).

[51] K. R. Bradnam, J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. A. Chapman, G. Chapuis, R. Chikhi, *et al.*, *Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species,* GigaScience **2**, 1 (2013).

[52] W. Xue, J.-T. Li, Y.-P. Zhu, G.-Y. Hou, X.-F. Kong, Y.-Y. Kuang, and X.-W. Sun, *L_RNA_scaffolder: scaffolding genomes with transcripts,* BMC Genomics **14**, 604 (2013).

[53] N. Kaplan and J. Dekker, *High-throughput genome scaffolding from in vivo DNA interaction frequency,* Nature Biotechnology **31**, 1143 (2013).

[54] G. G. Silva, B. E. Dutilh, T. D. Matthews, K. Elkins, R. Schmieder, E. A. Dinsdale, and R. A. Edwards, *Combining de novo and reference-guided assembly with scaffold_builder,* Source Code for Biology and Medicine **8**, 1 (2013).

[55] J. Lindsay, H. Salooti, I. Măndoiu, and A. Zelikovsky, *ILP-based maximum likelihood genome scaffolding,* BMC Bioinformatics **15**, 1 (2014).

[56] K. Sahlin, F. Vezzi, B. Nystedt, J. Lundeberg, and L. Arvestad, *BESST-Efficient scaffolding of large fragmented assemblies,* BMC Bioinformatics **15**, 1 (2014).

[57] M. Boetzer and W. Pirovano, *SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information,* BMC Bioinformatics **15**, 1 (2014).

[58] P. Berg and J. E. Mertz, *Personal reflections on the origins and emergence of recombinant DNA technology,* Genetics **184**, 9 (2010).

[59] H. O. Smith and K. Welcox, *A restriction enzyme from Hemophilus influenzae: I. purification and general properties,* Journal of Molecular Biology **51**, 379 (1970).

[60] T. J. Kelly and H. O. Smith, *A restriction enzyme from Hemophilus influenzae: Ii. base sequence of the recognition site,* Journal of Molecular Biology **51**, 393 (1970).

[61] K. Itakura, T. Hirose, R. Crea, A. D. Riggs, H. L. Heyneker, F. Bolivar, and H. W. Boyer, *Expression in Escherichia coli of a chemically synthesized gene for the hormone somatostatin,* Science **198**, 1056 (1977).

[62] D. Lubertozzi, *Life Since the Double Helix: 60 Years of Evolution in Biotechnology,* Online: `http://www.bioradiations.com/life-since-the-double-helix-60-years-of-evolution-in-biotechnology/` (2014), updated: Jan 14, 2014.

[63] A. F. Gilles and M. Averof, *Functional genetics for all: engineered nucleases, CRISPR and the gene editing revolution,* EvoDevo **5**, 1 (2014).

[64] R. Carlson, *Time for new DNA synthesis and sequencing cost curves,* Online: `http://www.synthesis.cc/2014/02/time-for-new-cost-curves-2014.html` (2014), updated: Feb 12, 2014.

[65] N. Annaluru, H. Muller, L. A. Mitchell, S. Ramalingam, G. Stracquadanio, S. M. Richardson, J. S. Dymond, Z. Kuang, L. Z. Scheifele, E. M. Cooper, *et al.*, *Total synthesis of a functional designer eukaryotic chromosome,* Science **344**, 55 (2014).

[66] S. Aldridge and J. Sturichio, *The Discovery and Development of Penicillin: 1928-1945,* (1999).

[67] J. Houbraken, J. C. Frisvad, and R. A. Samson, *Fleming?s penicillin producing strain is not Penicillium chrysogenum but P. rubens,* IMA Fungus: The Global Mycological Journal **2**, 87 (2011).

[68] M. L. Pasteur, *Études sur la bière, ses maladies, causes qui les provoquent, procédé pour la rendre inaltérable: avec une théorie nouvelle de la fermentation* (Gauthier-Villars, 1876).

[69] D. Das and A. Goyal, *Lactic acid bacteria in food industry,* in *Microorganisms in Sustainable Agriculture and Biotechnology* (Springer, 2012) pp. 757–772.

[70] I. S. Johnson, *Human insulin from recombinant DNA technology,* Science **219**, 632 (1983).

[71] N. Ferrer-Miralles, J. Domingo-Espín, J. L. Corchero, E. Vázquez, and A. Villaverde, *Microbial factories for recombinant pharmaceuticals,* Microbial Cell Factories **8**, 17 (2009).

[72] J. Du, Z. Shao, and H. Zhao, *Engineering microbial factories for synthesis of value-added products,* Journal of Industrial Microbiology & Biotechnology **38**, 873 (2011).

[73] D.-K. Ro, E. M. Paradise, M. Ouellet, K. J. Fisher, K. L. Newman, J. M. Ndungu, K. A. Ho, R. A. Eachus, T. S. Ham, J. Kirby, *et al.*, *Production of the antimalarial drug precursor artemisinic acid in engineered yeast,* Nature **440**, 940 (2006).

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

[75] S. Lubliner, I. Regev, M. Lotan-Pompan, S. Edelheit, A. Weinberger, and E. Segal, *Core promoter sequence in yeast is a major determinant of expression level,* Genome Research , gr (2015).

[76] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006).

[77] E. Cambria and B. White, *Jumping NLP curves: a review of natural language processing research,* Computational Intelligence Magazine, IEEE **9**, 48 (2014).

[78] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kégl, and D. Rousseau, *The Higgs boson machine learning challenge,* in *NIPS 2014 Workshop on High-energy Physics and Machine Learning,* Vol. 42 (2014) p. 37.

[79] N. Emanet, H. R. Öz, N. Bayram, and D. Delen, *A comparative analysis of machine learning methods for classification type decision problems in healthcare,* Decision Analytics **1**, 1 (2014).

[80] M. W. Libbrecht and W. S. Noble, *Machine learning applications in genetics and genomics,* Nature Reviews Genetics (2015).

[81] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.-P. Z. Wang, and J. Widom, *A genomic code for nucleosome positioning,* Nature **442**, 772 (2006).

[82] N. D. Heintzman, R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. Van Calcar, C. Qu, K. A. Ching, *et al.*, *Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome,* Nature Genetics **39**, 311 (2007).

[83] D. Kleftogiannis, P. Kalnis, and V. B. Bajic, *DEEP: a general computational framework for predicting enhancers,* Nucleic Acids Research , gku1058 (2014).

[84] A. B. Rosenberg, R. P. Patwardhan, J. Shendure, and G. Seelig, *Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences,* Cell **163**, 698 (2015).

[85] Z. E. Sauna and C. Kimchi-Sarfaty, *Understanding the contribution of synonymous mutations to human disease,* Nature Reviews Genetics **12**, 683 (2011).

[86] T. E. Quax, N. J. Claassens, D. Söll, and J. van der Oost, *Codon bias as a means to fine-tune gene expression,* Molecular Cell **59**, 149 (2015).

[87] A. G. Hinnebusch and J. R. Lorsch, *The mechanism of eukaryotic translation initiation: new insights and challenges,* Cold Spring Harbor Perspectives in Biology **4**, a011544 (2012).

[88] T. E. Dever and R. Green, *The elongation, termination, and recycling phases of translation in eukaryotes,* Cold Spring Harbor Perspectives in Biology **4**, a013706 (2012).

[89] M. Siwiak and P. Zielenkiewicz, *A comprehensive, quantitative, and genome-wide model of translation,* PLoS Computational Biology **6**, e1000865 (2010).

[90] S. Reuveni, I. Meilijson, M. Kupiec, E. Ruppin, and T. Tuller, *Genome-scale analysis of translation elongation with a ribosome flow model,* PLoS Computational Biology **7**, e1002127 (2011).

[91] J. Racle, J. Overney, and V. Hatzimanikatis, *A computational framework for the design of optimal protein synthesis,* Biotechnology and Bioengineering **109**, 2127 (2012).

[92] P. Shah, Y. Ding, M. Niemczyk, G. Kudla, and J. B. Plotkin, *Rate-limiting steps in yeast protein translation,* Cell **153**, 1589 (2013).

[93] K. Fredrick and M. Ibba, *How the sequence of a gene can tune its translation,* Cell **141**, 227 (2010).

[94] A. Dana and T. Tuller, *The effect of tRNA levels on decoding times of mRNA codons,* Nucleic Acids Research **42**, 9171 (2014).

[95] C. Gustafsson, S. Govindarajan, and J. Minshull, *Codon bias and heterologous protein expression,* Trends in Biotechnology **22**, 346 (2004).

[96] G. L. Rosano and E. A. Ceccarelli, *Rare codon content affects the solubility of recombinant proteins in a codon bias-adjusted Escherichia coli strain,* Microbial Cell Factories **8**, 41 (2009).

[97] E. Angov, *Codon usage: nature's roadmap to expression and folding of proteins,* Biotechnology Journal **6**, 650 (2011).

[98] S. A. Shabalina, N. A. Spiridonov, and A. Kashina, *Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity,* Nucleic Acids Research **41**, 2073 (2013).

[99] J. Zhou, W. J. Liu, S. W. Peng, X. Y. Sun, and I. Frazer, *Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability,* Journal of Virology **73**, 4972 (1999).

[100] W. J. Liu, K.-N. Zhao, F. G. Gao, G. R. Leggatt, G. J. Fernando, and I. H. Frazer, *Polynucleotide viral vaccines: codon optimisation and ubiquitin conjugation enhances prophylactic and therapeutic efficacy,* Vaccine **20**, 862 (2001).

[101] N. A. Burgess-Brown, S. Sharma, F. Sobott, C. Loenarz, U. Oppermann, and O. Gileadi, *Codon optimization can improve expression of human genes in Escherichia coli: A multi-gene study,* Protein Expression and Purification **59**, 94 (2008).

[102] P. M. Sharp and W.-H. Li, *The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications,* Nucleic Acids Research **15**, 1281 (1987).

[103] M. dos Reis, R. Savva, and L. Wernisch, *Solving the riddle of codon usage preferences: a test for*

translational selection, Nucleic Acids Research **32**, 5036 (2004).

[104] G. Cannarozzi, N. N. Schraudolph, M. Faty, P. von Rohr, M. T. Friberg, A. C. Roth, P. Gonnet, G. Gonnet, and Y. Barral, *A role for codon order in translation dynamics,* Cell **141**, 355 (2010).

[105] A. Roth, M. Anisimova, and G. M. Cannarozzi, *Measuring codon usage bias,* in *Codon evolution: mechanisms and models*, edited by G. M. Cannarozzi and A. Schneider (Oxford University Press, New York, 2012) Chap. 13, pp. 189–217.

[106] N. T. Ingolia, *Ribosome profiling: new views of translation, from single codons to genome scale,* Nature Reviews Genetics **15**, 205 (2014).

[107] C. J. McManus, G. E. May, P. Spealman, and A. Shteyman, *Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast,* Genome Research **24**, 422 (2014).

[108] J. G. Dunn, C. K. Foo, N. G. Belletier, E. R. Gavis, and J. S. Weissman, *Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster,* eLife **2**, e01179 (2013).

[109] G.-W. Li, E. Oh, and J. S. Weissman, *The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria,* Nature **484**, 538 (2012).

[110] C. A. Charneski and L. D. Hurst, *Positively charged residues are the major determinants of ribosomal velocity,* PLoS Biology **11** (2013).

[111] L. F. Lareau, D. H. Hite, G. J. Hogan, and P. O. Brown, *Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments,* eLife **3**, e01257 (2014).

[112] C. G. Artieri and H. B. Fraser, *Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation,* Genome Research **24**, 2011 (2014).

[113] C. J. Woolstenhulme, N. R. Guydosh, R. Green, and A. R. Buskirk, *High-Precision Analysis of Translational Pausing by Ribosome Profiling in Bacteria Lacking EFP,* Cell Reports **11**, 13 (2015).

[114] J. Gardin, R. Yeasmin, A. Yurovsky, Y. Cai, S. Skiena, and B. Futcher, *Measurement of average decoding rates of the 61 sense codons in vivo,* eLife **3**, e03735 (2014).

[115] J. Pelletier and N. Sonenberg, *Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA,* Nature **334**, 320 (1988).

[116] S. Jang, H. Kräusslich, M. Nicklin, G. Duke, A. Palmenberg, and E. Wimmer, *A segment of the 5′ nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation,* Journal of Virology **62**, 2636 (1988).

[117] A. R. Jurado, D. Tan, X. Jiao, M. Kiledjian, and L. Tong, *Structure and function of pre-mRNA 5′-end capping quality control and 3′-end processing,* Biochemistry **53**, 1882 (2014).

[118] C. H. Herbreteau, L. Weill, D. Décimo, D. Prévôt, J.-L. Darlix, B. Sargueil, and T. Ohlmann, *HIV-2 genomic RNA contains a novel type of IRES located downstream of its initiation codon,* Nature Structural & Molecular Biology **12**, 1001 (2005).

[119] I. N. Shatsky, S. E. Dmitriev, I. M. Terenin, and D. Andreev, *Cap-and IRES-independent scanning mechanism of translation initiation as an alternative to the concept of cellular IRESs,* Molecules and Cells **30**, 285 (2010).

[120] A. A. Komar and M. Hatzoglou, *Cellular IRES-mediated translation: the war of ITAFs in pathophysiological states,* Cell Cycle **10**, 229 (2011).

[121] M. Holčík and N. Sonenberg, *Translational control in stress and apoptosis,* Nature Reviews Molecular Cell Biology **6**, 318 (2005).

[122] M. Holčík, *Targeting Translation for Treatment of Cancer-A Novel Role for IRES?* Current cancer drug targets **4**, 299 (2004).

[123] M. Stoneley and A. E. Willis, *Cellular internal ribosome entry segments: structures, trans-acting factors and regulation of gene expression,* Oncogene **23**, 3200 (2004).

[124] A. A. Komar and M. Hatzoglou, *Exploring Internal Ribosome Entry Sites as Therapeutic Targets,* Frontiers in Oncology **5** (2015).

[125] L. Balvay, R. S. Rifo, E. P. Ricci, D. Decimo, and T. Ohlmann, *Structural and functional diversity of viral IRESes,* Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms **1789**, 542 (2009).

[126] A. A. Komar and M. Hatzoglou, *Internal ribosome entry sites in cellular mRNAs: The mystery of their existence,* Journal of Biological Chemistry **280**, 23425 (2005).

[127] M. Mokrejš, T. Mašek, V. Vopálenskỳ, P. Hlubuček, P. Delbos, and M. Pospíšek, *IRESite - a tool for the examination of viral and cellular internal ribosome entry sites,* Nucleic Acids Research **38**, D131 (2010).

[128] F. Koopman, J. Beekwilder, B. Crimi, A. van Houwelingen, R. D. Hall, D. Bosch, A. van Maris, J. T. Pronk, and J.-M. Daran, *De novo production of the flavonoid naringenin in engineered Saccharomyces cerevisiae,* Microbial Cell Factories **11**, 155 (2012).

[129] H. King, L. Cobbold, and A. Willis, *The role of IRES trans-acting factors in regulating translation initiation,* Biochemical Society Transactions **38**, 1581 (2010).

2

SCAFFOLDING NEXT-GENERATION SEQUENCING ASSEMBLIES

Alexey A. GRITSENKO, Jurgen F. NIJKAMP, Marcel J.T. REINDERS and Dick DE RIDDER

This chapter has been published in *Bioinformatics* **28**, 1429 (2012) [1].

2

ABSTRACT

Motivation: The increasing availability of second-generation *high-throughput sequencing* (HTS) technologies has sparked a growing interest in *de novo* genome sequencing. This in turn has fuelled the need for reliable means of obtaining high-quality draft genomes from short-read sequencing data. The millions of reads usually involved in HTS experiments are first assembled into longer fragments called *contigs*, which are then scaffolded, i.e., ordered and oriented using additional information, to produce even longer sequences called *scaffolds*. Most existing scaffolders of HTS genome assemblies are not suited for using information other than paired reads to perform scaffolding. They use this limited information to construct scaffolds, often preferring scaffold length over accuracy, when faced with the tradeoff.

Results: We present GRASS (GeneRic ASsembly Scaffolder) - a novel algorithm for scaffolding second-generation sequencing assemblies capable of using diverse information sources. GRASS offers a mixed-integer programming formulation of the contig scaffolding problem, which combines contig order, distance and orientation in a single optimisation objective. The resulting optimisation problem is solved using an *Expectation-Maximization* (EM) procedure and an unconstrained binary quadratic programming approximation of the original problem. We compared GRASS to existing HTS scaffolders using Illumina paired reads of three bacterial genomes. Our algorithm constructs a comparable number of scaffolds, but makes fewer errors. This result is further improved when additional data, in the form of related genome sequences, are used.

2.1. INTRODUCTION

High-throughput sequencing (HTS) technologies, such as Illumina (Illumina, Inc., San Diego, CA), 454 (Roche Applied Science, Penzberg, Germany) and SOLiD and IonTorrent (Life Technologies, Carlsbad, CA) produce millions of short DNA reads of typical lengths of 36-500 bp at low cost, making them attractive for *de novo* sequencing applications. With the aid of assembly algorithms [2–4], short reads can be joined together into longer sequences called *contigs*. However, contigs are typically shorter than the sequenced DNA molecules, as genomic repeat regions longer than the read length cannot be unambiguously assembled using the read sequences alone. Scaffolding, the process of using additional data to place contigs in the right order, orientation and at the right distance in longer (gapped) supercontigs called *scaffolds*, is a crucial step in obtaining high quality draft genome sequences.

Paired reads (mate pair or paired end reads, depending on the sequencing protocol), i.e., reads of known relative orientation, order and approximate physical distance, are often used for scaffolding. Additional information, including reference sequences of related organisms, restriction maps [5] and RNA-seq data, can be used to derive more accurate contig placement [6, 7], thereby reducing the cost of finishing experiments and allowing for more reliable downstream analyses. However, most existing scaffolding algorithms are not able to utilise such information for scaffolding. To our knowledge, only Bambus [7] and SOPRA [8] can make use of additional data sources, although the latter was not originally designed for this purpose.

Generally, the *Contig Scaffolding Problem* (CSP) is finding a linear ordering and orientation of contigs that minimises the number of unsatisfied scaffolding constraints. These constraints are derived from the available data through translation of the inherent distance, order and orientation constraints onto the contigs. The derived constraints can be mutually exclusive, which makes the problem of minimising the number of unsatisfied constraints NP-hard [9, 10]. Consequently, practical scaffolding algorithms only approximately solve this problem: Bambus [7] separately finds contig orientation and order and uses greedy heuristics to remove inconsistent constraints; SSPACE [11] greedily extends scaffolds using a heuristic stopping criterion; and SOPRA [8] uses an iterative procedure to identify a subset of contigs with consistent scaffolding constraints. Notable exceptions are OPERA [12] and the MIP Scaffolder [13], which simplify the CSP by dropping types of constraints. OPERA implements an algorithm for finding an exact CSP solution without minimum contig distance constraints; the MIP Scaffolder [13] couples a *Mixed-Integer Programming* (MIP) formulation of the contig scaffolding problem that does not enforce order constraints with an algorithm heuristically dividing the original problem into subproblems to be solved exactly.

We propose a novel GeneRic ASembly Scaffolding (GRASS) algorithm that can be applied to any type of scaffolding information. Our work is similar to Salmela *et al.* [13], as we propose a MIP formulation of the scaffolding problem. However, we combine contig orientation, order and distance in a single quadratic optimisation objective. Similar to Dayarian *et al.* [8], we employ an iterative procedure to select a consistent subset of contigs. However, we apply an expectation-maximization strategy to maximise the objective function that identifies inconsistent constraints rather than contigs, thereby retaining more scaffolding information.

We implemented the algorithm in C++ and tested it on *de novo* assemblies of paired read data for the bacteria *Eschrichia coli*, *Pseudoxanthomonas suwonensis*, and *Pseudomonas syringae* and compared it to the SSPACE, OPERA and MIP scaffolders. GRASS produces a competitive number of scaffolds with fewer scaffolding errors, particularly when combining various sources of scaffolding information.

2.2. METHODS

DATA REPRESENTATION

Scaffolding constraints on contig distance, order and orientation are derived from the data in a manner depending on the data type. For example, the known *relative* orientation, *relative* order and approximate distance of paired reads that map to different contigs can be translated into relative contig orientation, order and approximate contig distance by taking mapping orientations and positions into account; similarly, physical distance, relative order and orientation of two contigs mapping to the same reference sequence can be translated into corresponding constraints. However, different data types eventually define the same type of pair-wise contig constraints, which can be conveniently represented as arcs (i.e., directed edges) $l_j = (a_{l_j}, b_{l_j}) \in E$ of weight ω_{l_j} in a digraph $G = (V, E)$ defined over the set of contigs V [7, 10, 12]. The weight can be chosen to reflect information source importance and consistency. A relative order r_{l_j}, relative orientation e_{l_j} and approximate distance suggested by the pair-wise constraints, are then associated with every arc l_j. The approximate distance is recorded as mean μ_{l_j} and its standard deviation σ_{l_j}. This form is a natural choice for capturing variation in contig distances derived from the paired read insert size. It is also suitable for scaffolding constraints without (reliable) distance estimates, for example constraints derived from paired RNA-seq data of an organism with abundant intron splicing, or by mapping contigs to genome of a distant relative. Such constraints can use a large σ_{l_j} to reflect the uncertainty in the data source. We refer to l_j, its importance weight ω_{l_j}, and the corresponding contig pair-wise constraints as a *contig link*, and to G as the *contig link graph*. For succinct notation, for every contig link constraints are recorded as

- $e_{l_j} = \begin{cases} 0, & a_{l_j} \text{ and } b_{l_j} \text{ are from different strands} \\ 1, & a_{l_j} \text{ and } b_{l_j} \text{ are from the same strand} \end{cases}$
- $r_{l_j} = \begin{cases} 0, & a_{l_j} \text{ follows } b_{l_j} \\ 1, & b_{l_j} \text{ follows } a_{l_j} \end{cases}$ given that a_{l_j} has forward orientation.

This abstract definition is illustrated in Fig. 2.1. It allows capturing any combination of contig order, distance and orientation, including constraints derived from paired end reads, mate pair reads, and contig mapping.

CONTIG LINK BUNDLING AND EROSION

We create a single contig link for every available piece of evidence (e.g., pair of reads) and by default set its importance weight to one (a parameter adjustable per information source). For high coverage HTS data this procedure creates a large number of links. Contig link bundling is used to reduce the number of links, and thereby the complexity

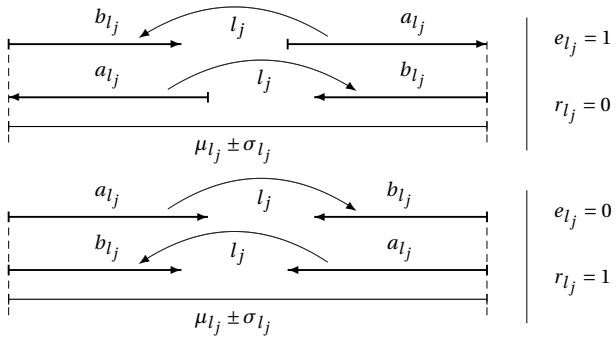

Figure 2.1: Examples of contig links l_j between contigs a_{l_j} and b_{l_j} and their corresponding relative orientation (e_{l_j}), relative order (r_{l_j}) and distance ($\mu_{l_j} \pm \sigma_{l_j}$) constraints.

of the problem. For every ordered pair of contigs (u, v), arcs $(u, v) \in E$ that agree on contig distance, order and orientation are combined into one or more contig links as in Huson *et al.* [10]. The weight of a link after bundling is equal to the sum of weights of links bundled together to create it. Our definition of contig links permits having links that agree on all constraints, yet cannot be bundled together because they are oppositely directed in G. To enable bundling of such links, we re-set r_{l_j} relative to one of the end points of l_j to make sure that all links connecting a pair of contigs have the same directionality. Finally, contig links with importance weight smaller than a predefined erosion threshold e are removed from the graph. This assumes that erroneous links are rare.

OPTIMISATION FORMULATION

We present a *mixed-integer quadratic programming* (MIQP) formulation of the contig scaffolding problem. Our formulation is equivalent to the traditional one (minimise the number of unsatisfied constraints, Huson *et al.* [10]), but uses slack variables as continuous measures of the extent to which each order and orientation constraint is satisfied. This allows for uncertain data, yielding less trustworthy constraints, to be accurately exploited in the scaffolding process. A number of optimisation variables are associated with every contig and contig link. We maximise an objective function f of these variables subject to scaffolding constraints expressed as linear optimisation constraints. The function reaches its maximum value when all distance, order and orientation constraints are satisfied. Each valid collection of the optimisation variable values forms a solution to the optimisation problem. These values are sufficient to puzzle contigs into scaffolds. For every contig c_i, where $i = 1, \ldots, n$, the following variables are defined as illustrated in Fig. 2.2

- $t_i = \begin{cases} 0, & c_i \text{ comes from the forward strand of the scaffold} \\ 1, & c_i \text{ comes from the reverse strand of the scaffold} \end{cases}$
  is used to define contig orientation in the scaffold.
- $x_i \in \mathbb{R}^+$ corresponds to the $5'$ position of c_i in the scaffold (when input contigs and the constructed scaffold are viewed as having a $5'$ to $3'$ orientation).

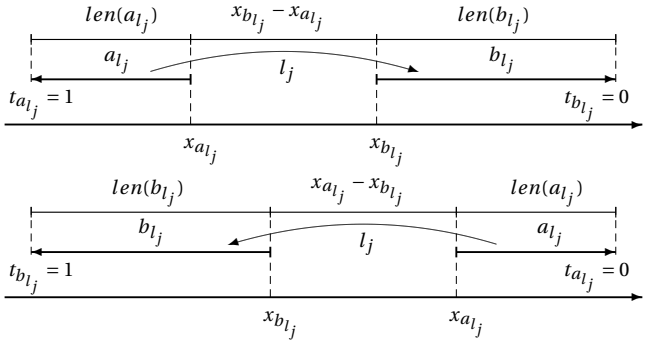

Figure 2.2: Optimisation variables $x_{a_{l_j}}$, $x_{b_{l_j}}$, $t_{a_{l_j}}$ and $t_{b_{l_j}}$ associated with contigs. Example for $e_{l_j} = 0 \wedge r_{l_j} = 0$.

Naturally, x_i should be an integer variable, but it is relaxed to simplify the optimisation problem and is rounded to the nearest integer when the solution is converted into scaffold nucleotide sequences. Additionally, with every link l_j, $j = 1,\ldots,m$ the following variables are associated:

- **Slack variables** for distance constraints, $\xi_{l_j} = \{\vec{\xi}_{l_j}, \overleftarrow{\xi}_{l_j}\} \in \mathbb{R}^+ \times \mathbb{R}^+$, and order constraints, $\Delta_{l_j} = \{\vec{\Delta}_{l_j}, \overleftarrow{\Delta}_{l_j}\} \in \mathbb{R}^+ \times \mathbb{R}^+$, for forward ($t_{a_{l_j}} = 0$) and reverse ($t_{a_{l_j}} = 1$) orientations of the contig pair respectively. By design these variables reflect the degree to which the corresponding constraints are violated. They are penalised in the optimisation objective f.

- **Switch variables** for distance constraints, $\alpha_{l_j} \in \{0,1\}$, and order constraints, $\beta_{l_j} \in \{0,1\}$ (0, constraint is disabled; 1, enabled) used for disabling contig link constraints with high penalties.

As distance and order constraints are influenced by the orientation, different slack variables are required for both orientations. We omit orientation arrows above slacks ξ_{l_j} and Δ_{l_j} when the contig pair orientation is not important, or is clear from the context.

Contig links impose scaffolding constraints, which can be modelled as MIQP optimisation constraints. We demonstrate here how such constraints can be derived from paired read data; the same type of constraints can be derived in a similar way from other sources of scaffolding information (for example, see section 2.3).

Distance constraints are expressed as:

$$\frac{|d(a_{l_j}, b_{l_j}) - \mu_{l_j}|}{\sigma_{l_j}} \leq \xi_{l_j}, \tag{2.1}$$

where $d(a_{l_j}, b_{l_j})$ is the distance between contigs a_{l_j} and b_{l_j}, and ξ_{l_j} is a distance slack variable. This inequality captures uncertainty in the distance by measuring the difference with the mean in standard deviations. We derive contig distance $d(a_{l_j}, b_{l_j})$ from the paired read insert size as the gap size plus the contig lengths. The calculation

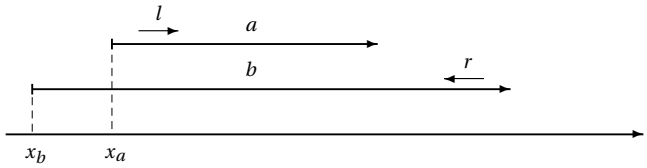

Figure 2.3: Contigs a and b are not in the order predicted by mapped paired reads l and r, although the paired reads are in the correct order.

then depends on the order and orientation of contigs connected by l_j. It can be fixed by assuming that the contigs have relative orientation and order suggested by l_j. For example, for the case of $(e_{l_j} = 0 \wedge r_{l_j} = 0)$ shown in Fig. 2.2, the distance expression depends on contig pair orientation through $t_{a_{l_j}}$:

$$d(a_{l_j}, b_{l_j}) = x_{a_{l_j}} - x_{b_{l_j}} + len(a_{l_j}) + len(b_{l_j}), \quad t_{a_{l_j}} = 0$$
$$d(a_{l_j}, b_{l_j}) = x_{b_{l_j}} - x_{a_{l_j}} + len(a_{l_j}) + len(b_{l_j}), \quad t_{a_{l_j}} = 1.$$

Combined with (2.1) the following constraints are obtained:

$$\begin{cases} x_{a_{l_j}} - x_{b_{l_j}} \leq & \sigma_{l_j}\vec{\xi}_{l_j} + \mu_{l_j} - len(a_{l_j}) - len(b_{l_j}) \\ x_{a_{l_j}} - x_{b_{l_j}} \geq & -\sigma_{l_j}\vec{\xi}_{l_j} + \mu_{l_j} - len(a_{l_j}) - len(b_{l_j}) \\ x_{b_{l_j}} - x_{a_{l_j}} \leq & \sigma_{l_j}\overleftarrow{\xi}_{l_j} + \mu_{l_j} - len(a_{l_j}) - len(b_{l_j}) \\ x_{b_{l_j}} - x_{a_{l_j}} \geq & -\sigma_{l_j}\overleftarrow{\xi}_{l_j} + \mu_{l_j} - len(a_{l_j}) - len(b_{l_j}) \end{cases}, \tag{2.2}$$

where different slack variables are used for the two contig pair orientations. The expressions for other combinations of e_{l_j} and r_{l_j} are derived similarly.

Order constraints are derived from read order constraints (i.e., if c_j follows c_i, then they should not overlap and c_j must be upstream of c_i), which additionally can be relaxed. The relaxation is necessary because (i) assembled contigs may overlap [7]; (ii) in some cases the order constraints on data are not valid when extended to contigs, as illustrated in Fig. 2.3. Translating order constraints into optimisation constraints as

$$\begin{cases} x_{a_{l_j}} - x_{b_{l_j}} \geq & -len(b_{l_j}) \cdot \vec{\Delta}_{l_j}, \quad t_{a_{l_j}} = 0 \\ x_{b_{l_j}} - x_{a_{l_j}} \geq & -len(a_{l_j}) \cdot \overleftarrow{\Delta}_{l_j}, \quad t_{a_{l_j}} = 1 \end{cases} \tag{2.3}$$

(formulas shown for $e_{l_j} = 0 \wedge r_{l_j} = 0$) discourages overlaps while still allowing the order constraint to be violated when $\Delta_{l_j} > 1$. These slack variables are weighed by the length of the downstream contig to allow measuring them on a single scale. As for the distance optimisation constraints, it is assumed that the relative contig orientation is correct.

Orientation constraints are modelled in the optimisation objective function, which is designed to attain larger values when more orientation constraints are satisfied. The

function is given by a polynomial

$$g(t) = \sum_{j=1,\dots,m}^{e_{l_j}=0} q_{a_{l_j} b_{l_j}} \omega_{l_j} + \sum_{j=1,\dots,m}^{e_{l_j}=1} (1 - q_{a_{l_j} b_{l_j}}) \omega_{l_j},$$

where $q_{ab} = t_a + t_b - 2t_a t_b \equiv \begin{cases} 0, & a \text{ and } b \text{ are equally oriented} \\ 1, & \text{otherwise} \end{cases}$. It is equal to the sum of weights of contig links with satisfied orientation and serves as a basis for the optimisation objective that is further penalised proportionally to slack variables.

Slack penalties. The distance and order constraints are added to the optimisation problem through slack variable penalisation. The penalty is proportional to the importance weight of the corresponding contig link and to the value of the slack variable. To avoid situations when a low-weight violated constraint results in a large penalty, a maximum penalty of half of the importance weight is enforced, after which the constraint is considered disabled. Doing this has the additional benefit of equalising the influence of order and distance constraints. To this end we penalise as follows

$$\frac{\omega_{l_j}}{2} \cdot \frac{\min(\xi_{l_j}, S_\xi)}{S_\xi}, \tag{2.4}$$

where ξ_{l_j} is chosen as $\vec{\xi}_{l_j}$ or $\overleftarrow{\xi}_{l_j}$, according to the contig pair orientation and S_ξ is the maximum slack threshold (after which the slack is disabled). Because the expression $\min(\xi, S_\xi)$ is not suitable for direct use in a MIP, it is unrolled using the switch variables as $[\alpha_{l_j} \xi_{l_j} + (1 - \alpha_{l_j}) S_\xi]$. Similar penalties with variables Δ_{l_j} and β_{l_j}, and maximum slack threshold S_Δ are used for the order constraints. We set $S_\xi = 6$ (i.e., six standard deviations), as in Gao *et al.* [12], Li and Durbin [14]; and $S_\Delta = 1$, as at this value of slack the physical order constraint is not satisfied anymore. Further, only the slacks for the appropriate contig pair orientation have to be penalised. This is achieved by penalising $(1 - t_{a_{l_j}})\vec{\xi}_{l_j} + t_{a_{l_j}} \overleftarrow{\xi}_{l_j}$ in place of ξ_{l_j} in (2.4). This expression "chooses" which slack variable to penalise depending on the contig pair orientation. Finally, the constraints have to be penalised only when they are meaningful (i.e., the relative contig orientation e_{l_j} is assumed to be satisfied). The resulting function looks as follows:

$$\begin{aligned} h(t, \alpha, \xi, S_\xi) = &\sum_{j=1,\dots,m}^{e_{l_j}=0} q_{a_{l_j} b_{l_j}} \frac{\omega_{l_j}}{2 S_\xi} \left[(1 - t_{a_{l_j}}) \vec{\xi}_{l_j} + t_{a_{l_j}} \overleftarrow{\xi}_{l_j} \right] + \\ &+ \sum_{j=1,\dots,m}^{e_{l_j}=1} (1 - q_{a_{l_j} b_{l_j}}) \frac{\omega_{l_j}}{2 S_\xi} \left[(1 - t_{a_{l_j}}) \vec{\xi}_{l_j} + t_{a_{l_j}} \overleftarrow{\xi}_{l_j} \right]. \end{aligned}$$

Expansion of this function leads to a fourth degree polynomial, containing only terms that consist purely of binary variables, or one continuous and up to three binary variables. To construct a MIQP formulation, using the big-M formulation [15], these terms can be replaced by a single new auxiliary variable each at the expense of

introducing new optimisation constraints.

Putting it all together. We maximise

$$f(x,t,\alpha,\beta,\xi,\Delta) \equiv g(t) - h(t,\alpha,\xi,S_\xi) - h(t,\beta,\Delta,S_\Delta),$$

s.t. constraints (2.2) and (2.3) are satisfied. Here $g(t)$ is maximised for orientation, $h(t,\alpha,\xi,S_\xi)$ is minimised for orientation and distance, and $h(t,\beta,\Delta,S_\Delta)$ is minimised for orientation and order, in a single optimisation objective. Given the NP-hard nature of MIPs and the large number of binary variables in the proposed formulation, this problem becomes intractable even for small numbers of contigs.

PROBLEM SPLITTING

We tackle this intractability with an *expectation-maximisation* (EM) -like procedure.

The maximisation step assumes the contig orientations are known (i.e., t_i and q_{ab} are fixed). Knowing t_i allows us to choose the slack variables ($\vec{\xi}_{l_j}$ or $\overleftarrow{\xi}_{l_j}$, and $\vec{\Delta}_{l_j}$ or $\overleftarrow{\Delta}_{l_j}$) depending on the contig pair orientations, and to select contig links with satisfied relative orientation before the optimisation problem is constructed, significantly reducing the number of optimisation constraints and the complexity of the optimisation problem:

$$\begin{aligned} f(x,\alpha,\beta,\xi,\Delta) &= g - h(\alpha,\xi,S_\xi) - h(\beta,\Delta,S_\Delta) \\ g = \sum \omega_{l_j} \equiv \text{const}, \quad h(\alpha,\xi,S_\xi) &= \sum \min(\xi_{l_j}, S_\xi) \cdot \tfrac{\omega_{l_j}}{2S_\xi} \end{aligned} \quad . \tag{2.5}$$

This *fixed optimisation problem*, however, is still NP-hard due to the binary variables α_{l_j} and β_{l_j} involved in expansion of the min terms. We obtain an approximate solution to this problem by first exactly solving its continuous relaxation, choosing α_{l_j} and β_{l_j} according to the slack values in the relaxation solution and finally, re-solving the problem with these values fixed. The relaxation is obtained by replacing $h(\alpha,\xi,S_\xi)$ by $h(\xi,S_\xi) = \frac{1}{2S_\xi}\sum \omega_{l_j}\xi_{l_j}$ in (2.5). This eliminates all binary variables, allowing the use of efficient optimisation algorithms [16]. The solution for the relaxed problem gives us optimal values for slacks ξ_{l_j} and Δ_{l_j}, which are used to choose α_{l_j} and β_{l_j} as

$$\alpha_{l_j} = \begin{cases} 0, & \xi_{l_j} > S_\xi \\ 1, & \xi_{l_j} \leq S_\xi \end{cases}, \quad \beta_{l_j} = \begin{cases} 0, & \Delta_{l_j} > S_\Delta \\ 1, & \Delta_{l_j} \leq S_\Delta \end{cases},$$

and allows us to re-solve problem (2.5). The rationale behind is that, since the majority of link information is assumed to be correct, large slack values will be associated with incorrect constraints that have to be disabled. The total penalty for l_j is memorised (initially set to zero) for use in the expectation step as

$$\Theta_{l_j} \leftarrow \frac{\min(\xi_{l_j}, S_\xi)}{2S_\xi}\omega_{l_j} + \frac{\min(\Delta_{l_j}, S_\Delta)}{2S_\Delta}\omega_{l_j}.$$

The expectation step is used to obtain the expected contig orientations t_i, which maximise the objective function for the *previously observed* penalties. Consider the

MIQP problem when penalties associated with the links are known (i.e., Δ_{l_j}, ξ_{l_j}, α_{l_j} and β_{l_j} are fixed), and the optimal contig orientation is sought. In this problem, when a contig link is enabled, its weight is penalised by the associated slack Θ_{l_j}. We can, therefore, consider an equivalent problem where all slacks are zero and link weights are modified as $\tilde{\omega}_{l_j} \leftarrow \omega_{l_j} - \Theta_{l_j}$. The problem is then to maximise

$$f(t) \equiv g(t) = \sum_{j=1,\ldots,m}^{e_{l_j}=0} q_{a_{l_j} b_{l_j}} \tilde{\omega}_{l_j} + \sum_{j=1,\ldots,m}^{e_{l_j}=1} (1 - q_{a_{l_j} b_{l_j}}) \tilde{\omega}_{l_j} \tag{2.6}$$

free of any constraints. This is an *unconstrained binary quadratic programming* (UBQP) problem [17], the problem of maximising a function $c(t) = t^t C t$, where x is a binary vector of length n and C is an $n \times n$ real matrix. Consider a vector of orientations $t \in \{0,1\}^n$ and a matrix C of size n. Starting from a zero matrix, $C = (c_{ij})$ can be obtained by updating it for every link $l_j = (a, b)$ as

$$c_{aa} \leftarrow (-1)^{e_{l_j}} \tilde{\omega}_{l_j} + c_{aa}, \qquad c_{bb} \leftarrow (-1)^{e_{l_j}} \tilde{\omega}_{l_j} + c_{bb}$$
$$c_{ab} \leftarrow (-1)^{e_{l_j}+1} \cdot 2\tilde{\omega}_{l_j} + c_{ab}.$$

The functions $f(t)$ and $c(t)$ will then differ by a constant and, therefore, reach maxima for the same t. Solving a UBQP is known to be an NP-hard, but well-studied problem with efficient heuristic algorithms available [18–20]. Thus, the UBQP formulation of the problem is preferred over (2.6) for obtaining values of t_i.

The EM steps are iterated while contig orientations change. The algorithm can be viewed as an iterative UBQP approximation of the original MIQP problem. In practice, it converges to a solution within 7 iterations.

SCAFFOLD EXTRACTION AND POST-PROCESSING

Repeat contigs in the contig link graph G are connected by ambiguous links, hindering a confident positioning in scaffolds. In a pre-processing step, we detect such contigs using a modification of the A-statistic [21] proposed by Zerbino [22], and prevent their incorporation in scaffolds by removing all links from G incident to them. The connected components of G correspond to separate subproblems, which are solved independently.

After optimisation, each solution tuple (x, t, α, β) and corresponding subgraph G' are converted into one or more scaffolds. First, contig links with disabled constraints (i.e., $\alpha_{l_j} = 0 \vee \beta_{l_j} = 0$) are removed from G' to minimise the chance of incorrectly incorporating contigs in the same scaffold. Every connected component of the resulting G' is used to construct a single nucleotide sequence. Contigs are processed in order of their downstream end coordinates. The left end of the first contig is put at the start of the sequence; every new contig is added to the scaffold such that the gap between two consecutive contigs is preserved. When consecutive contigs are predicted to overlap (i.e., have a negative gap size), the new contig is pushed upstream to eliminate the overlap.

Because resolving contig overlaps in this way potentially leads to erroneous sequence reconstruction, we also explore an optional post-processing approach that performs global sequence alignment on consecutive contigs to find the best overlap. Global alignment is performed using a divide-and-conquer version of the Needleman-Wunsch

algorithm [23]. Algorithm implementation from the NCBI C++ Toolkit was used [24]. For every consecutive pair of contigs predicted to have a gap of μ bp, all gap sizes of at most $d = 100$ bp away from the predicted value are examined. Negative gap sizes indicate overlaps. For each gap size g, global alignment of overlapping contig ends is performed (match score of $p_{\text{match}} = 2$, mismatch penalty of $p_{\text{mismatch}} = -3$). The best gap size is then chosen based on the alignment score S and proximity to the predicted gap size μ by maximising

$$\frac{S}{g \cdot p_{\text{match}}} \cdot \frac{d - |g - \mu|}{d}. \tag{2.7}$$

With the (mis)match scores chosen as above, this expression takes values in $[-1.5; 1]$. Due to computational complexity only overlaps of no more than 1500 bp are considered (gap sizes with longer overlaps are assigned a score of -1). The decision to join two contigs, to leave a gap between them or to split the scaffold is then made:

- If none of the considered gap sizes suggest overlaps, the two contigs are positioned in a scaffold with a gap of μ bp.
- If value of expression (2.7) for the chosen gap size g passes a quality threshold of 0.8, the contigs are positioned to have an overlap of g bp. The overlap is replaced with the alignment consensus sequence, where mismatches are masked with unknown nucleotides.
- If the chosen gap size does not pass the quality threshold and is shorter than 50 bp, the two contigs are positioned successively one following another with no overlap.
- Finally, if the chosen gap size suggests a longer overlap, the currently constructed scaffold is split into two with a new scaffold starting from a contig that was predicted to lie upstream.

In principle, the proposed post-processing step with scaffold splitting allows for construction of more accurate scaffolds compared to the naïve scaffold extraction. We refer to the combination of GRASS and post-processing as GRASS+.

EVALUATION CRITERIA

Similar to assemblies, scaffolds are evaluated based on accuracy and contiguity. Scaffold accuracy can be assessed by comparing scaffolds to available reference sequences. We adopted the evaluation criteria from Dayarian *et al.* [8], Gao *et al.* [12] and counted the number of scaffold breakpoints, i.e., consecutive contig pairs in the scaffold that do not agree with the reference on contig distance, order or orientation. We perform local alignment of scaffolds to the reference and count the number of breakpoints within each scaffold. Two consecutive alignments are counted as a breakpoint if any of these hold: (i) they align to two different chromosomes in the reference; (ii) their relative orientations in the scaffold and in the reference do not match; (iii) their relative orders in the scaffold and in the reference do not match; (d) the difference in distance in the scaffold and in the reference is larger than Δ. We used $\Delta = 10$ kbp and $\Delta = 500$ bp to asses contig distance correctness at low and high resolution respectively.

MUMmer [25] was used to align scaffolds to references. Best hits for each position in the scaffold were computed. Only hits with at least 90 aligned bases (alignment length × alignment identity), were taken into account. In practice, very few alignments do not pass this cutoff. The alignments are also used to calculate the percentage of the scaffold bases and the reference bases that are aligned [13]. These numbers capture scaffold accuracy and completeness.

Finally, scaffold completeness and contiguity are captured as in sequence assembly, calculating total length of all scaffolds, number of scaffolds, maximum scaffold length and the N50 statistic.

2.3. IMPLEMENTATION

GRASS source code is available under the GNU GPL v3 license. It was developed in C++ and tested on Linux. GRASS consists of *linker* and *scaffolder* modules. The linker takes contigs and the available information sources as input and produces linking and coverage data, which is then used by the scaffolder module. It filters out repeat contigs and uses the remaining data to produce scaffolds. Scaffolds are output both as lists of contigs with assigned coordinates and orientations, and as linear FASTA sequences with gaps.

PAIRED READ DATA PROCESSING

To obtain contig links from paired read data, the linker module performs single-end mapping of the reads to contigs. The algorithm used for mapping depends on the data type: BWA [14] for Illumina reads, NovoAlign (http://www.novocraft.com/) for 454 data. The aligners are set to output all mapping locations, including non-unique hits, as a SAM file [26], which is then converted to BAM for further processing. This process is applied to each paired read library.

Read alignments are preprocessed to remove read pairs with low-quality and ambiguous alignments. As a rule, only unique hits with no mismatches and minimum read length of 30 bp are kept. The filtered alignments are then scanned for paired reads that align to different contigs. Each such read pair mapping is used to create a single contig link with distance, order and orientation constraints derived from the mapping and the given read pairing method (i.e., paired ends or mate pairs). The BamTools API [27] is used for filtering and processing read alignments.

RELATED GENOME DATA PROCESSING

An available reference sequence, such as the genome of a related organism, can be used for guiding the scaffolding process. For this purpose, contigs are aligned to the reference sequence. For every contig, a position in the reference sequence is obtained from contig tiling constructed from local alignments using MUMmer. Contig links are then created for every pair of consecutive contigs aligning to the same reference sequence, with relative orientation and order derived from the tiling. To capture alignment quality, weights for links $l_j = (a_{l_j}, b_{l_j})$ are set to $I_{a_{l_j}} \times I_{b_{l_j}} \times C_{a_{l_j}} \times C_{b_{l_j}} \times W$, where $I_{a_{l_j}}$ and $I_{b_{l_j}}$ are alignment identities, $C_{a_{l_j}}$ and $C_{b_{l_j}}$ are alignment coverages reported by MUMmer for the corresponding contigs, and $W > 0$ is a weight assigned to the reference sequences as

Table 2.1: Available datasets. NCBI/EBI accession numbers are given for reference sequences and read sets. In all cases reads were produced by the Illumina sequencing platform.

	E. coli		*P. suwonensis*		*P. syringae*
Genome size	4.64 Mbp		3.42 Mbp		6.09 Mbp
Reference	NC_000913.2		CP002446.1		NC_007005.1
Dataset	SRR001665	SRR001666	SRR097515	SRR191848	ERR005143
Read count	$2\times10,408,224$	$2\times7,047,668$	$2\times23,960,004$	$2\times19,789,425$	$2\times3,551,133$
Read length	36 bp	36 bp	76 bp	76 bp	36 bp
Coverage	160×	107×	709×	824×	38×
Insert size	216 ± 10	488 ± 18	189 ± 17	189 ± 17	401 ± 33

a scaffolding information source. This procedure is applied for each available reference sequence to create links, which are then used together in the optimisation.

OPTIMISATION PROBLEM SOLUTION

The EM procedure proposed for solving the MIQP formulation of the contig scaffolding problem splits it into a continuous *linear programming* (LP) problem, and an UBQP problem. Although more efficient algorithms for solving UBQPs are available [18, 20], a memetic algorithm from Merz and Katayama [19] was chosen for ease of implementation. Usually, contig link graphs are sparse due to the linear scaffold structure that they encompass. Memetic algorithms improve individual solutions through local search, which in turn is well-suited for smooth search landscapes (as in the case of sparse contig link graphs). Graph sparsity is further exploited by implementing sparse matrix operations as in Merz and Katayama [19].

We use the C++ Concert API for the CPLEX Optimiser [28] to solve LPs. CPLEX is freely available for academic use.

2.4. RESULTS AND DISCUSSION

EXPERIMENTAL SETUP

We have evaluated GRASS on *de novo* HTS assemblies of three bacterial genomes: *Escherichia coli* K12, substr. MG1655; *Pseudoxanthomonas suwonensis* 11-1; and *Pseudomonas syringae* B728a. For these organisms, finished genome sequences and HTS data from resequencing experiments are available. Presence of a finished genome sequence allows for reliably evaluating the algorithm and comparing it to other scaffolders in a *de novo* setup. This is achieved by using the reference sequence only in scaffold evaluation (thus not as an additional information source in the scaffolding process). The available test data is summarised in Table 2.1. Insert size and coverage were obtained from paired read mapping using BWA and BEDTools [29].

Velvet [2] was used to assemble reads into contigs. All assemblies had a coverage cutoff of 6 and were not scaffolded by the assembler. Only contigs longer than 150 bp were kept. Repeat resolution was disabled (i.e., no expected coverage was provided). For each organism, the k-mer length was chosen by performing assemblies for various k and choosing one based on assembly contiguity, length, percentage of mapped single reads,

and percentage of properly paired reads [14] (Suppl. Tables 2.6,2.7 and 2.8). For *E. coli*, *P. suwonensis* and *P. syringae*, $k = 31$, $k = 59$ and $k = 23$ were chosen respectively. This way of choosing k reflects real-life *de novo* assembly scenarios, yielding a realistic algorithm evaluation. Final assemblies are characterised in Tables 2.2–2.4.

COMPARISON TO OTHER SCAFFOLDERS

We compared GRASS to SSPACE, MIP and OPERA scaffolders. Where required, insert size estimates from Table 2.1 were used. Tables 2.2–2.4 show evaluation metrics calculated for these scaffolders and the available test data. Unless stated otherwise, all scaffolders were run with default parameter settings. BWA was used to map reads to scaffolds and produces SAM files required by MIP. As in Salmela *et al.* [13], at most two mismatches were allowed in read mapping. For SSPACE and OPERA, reads were aligned with Bowtie [30] using scripts provided with the scaffolders.

GRASS used an erosion cutoff of 4 (although better results can be obtained by tuning this parameter) and coverage estimates obtained from exact mapping of the reads to the assembly contigs. The latter is available from output of the linker module.

The SSPACE maximum distance parameter was set to 6 standard deviations for each paired library. Libraries were input in order of increasing insert size.

The MIP Scaffolder was also provided with coverage estimates computed from exact read mapping. Following the original publication, we tried different filtering parameters (ω, p) and chose those which gave the highest N50 value. Settings $(36, 0.8)$, $(70, 0.4)$ and $(50, 0.6)$ were selected for the *E. coli*, *P. suwonensis* and *P. syringae* data respectively. Maximum partition sizes were set to 100 for the *E. coli* scaffolds and 50 for the *P. suwonensis* and *P. syringae* scaffolds. Maximum and minimum insert sizes were chosen by adding and subtracting 6 standard deviations to the mean insert size.

OPERA does not allow using multiple read sets. It was applied to each read library separately, and in the case of *P. suwonensis*, also to a join of the available read sets, as they have the same insert size. The minimum contig length was set to 150 bp, i.e., the contig length cutoff parameter used in Velvet. We used the default PET parameter value whenever possible and increased it to the minimum value that allowed OPERA to finish without triggering a timeout abort. Cutoff values 6 and 7 were used for the *E. coli* dataset; cutoffs 27, 5 and 5 were used for the *P. suwonensis* dataset; and 11 was used for the *P. syringae* dataset (values are given in the order of the experiments in Tables 2.2–2.4).

SOPRA was applied to assembly graphs produced by Velvet. However, when used with parameters chosen in accordance to the manual provided, SOPRA produced highly fragmented scaffolds compared to results from Salmela *et al.* [13]. To allow for a fair comparison, its results were not taken into account.

As a scaffolder, Velvet was provided with mean insert size and standard deviation for each library. The data was reassembled with repeat resolution (expected coverage estimated automatically) and scaffolding turned on. Its performance was used as a baseline over which all scaffolders improved on *P. syringae* data and only SSPACE and GRASS improved on *E. coli* and *P. suwonensis* data.

Tables 2.2–2.4 contain the results. Note that the minimum number of breakpoints is one, due to the circular structure of bacterial genome. Breakpoints at $\Delta = 10$ kb and $\Delta = 500$ bp differ only slightly, suggesting that gap lengths are estimated with high precision.

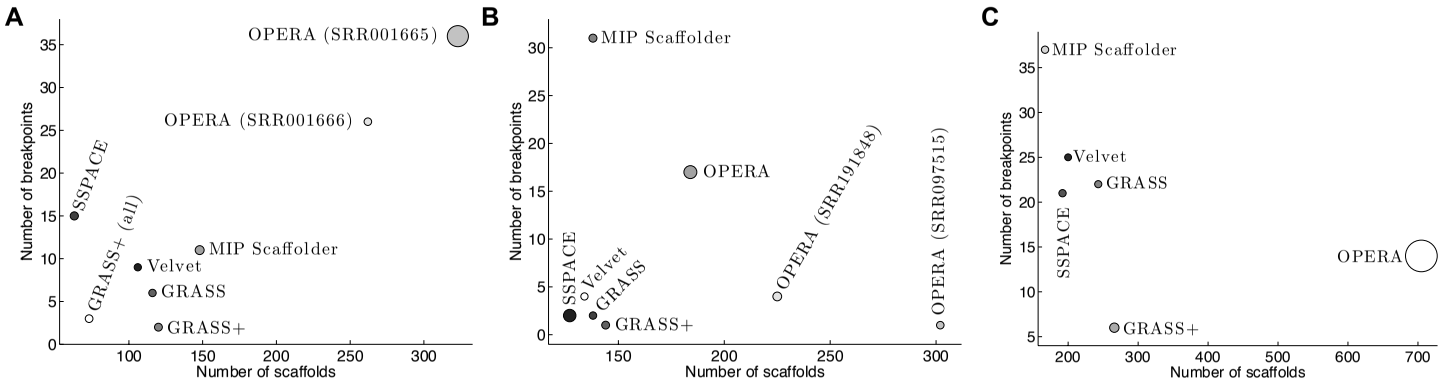

Figure 2.4: Scaffold accuracy and contiguity tradeoff for the (**A**) *E. coli*, (**B**) *P. suwonensis* and (**C**) *P. syringae* data. Marker size indicates scaffolding running time in minutes ranging between 8 sec and 72 mins, exact numbers are given in Suppl. Table 2.9. GRASS+ using paired reads and two related genomes is shown in A as "GRASS+ (all)".

SSPACE produced the longest scaffolds for *E. coli*. It also produced the smallest number of scaffolds for *E. coli* and *P. suwonensis*. The longest scaffolds and the smallest number of scaffolds on the *P. syringae* dataset are achieved by the MIP Scaffolder. Similar scaffold and reference coverage percentages were achieved by all scaffolders. However, GRASS+ has the smallest number of breakpoints for all considered organisms. Additionally, for the case of *P. suwonensis*, GRASS constructed the longest scaffolds and GRASS+ produced breakpoint-free scaffolds while providing a 2-fold reduction in the number of contigs. Scaffolds produced by the MIP Scaffolder and OPERA are either very fragmented or have a large number of breakpoints.

When constructing scaffolds, scaffolding algorithms balance between scaffold contiguity and scaffold accuracy. This tradeoff is captured in Fig. 2.4 by plotting the number of breakpoints (at $\Delta = 10$ kbp) against the number of scaffolds. A good scaffolder would be located in the lower left corner of such a plot. In many cases, GRASS combines a smaller number of breakpoints with a small number of scaffolds, compared to other scaffolders. The MIP Scaffolder and SSPACE can achieve smaller numbers of scaffolds, but at the cost of (much) larger numbers of breakpoints. Clearly, GRASS and SSPACE represent two possible choices of scaffolding algorithms, with GRASS being more accurate with respect to the number of breakpoints and SSPACE constructing longer scaffolds. This behaviour of the two algorithms is consistent over all datasets.

We also measured scaffolding running times, these are depicted in Fig. 2.4 using marker size. Exact numbers, as well as read mapping running times are available in Suppl. Table 2.9. Like most scaffolders, GRASS spends a majority of its time on read alignment, making running times of different scaffolders comparable and running time of the core scaffolding part of GRASS on the considered datasets negligible. Based on simulation results, we do not expect computation to become a bottleneck for large genomes. Nevertheless, to reduce computational load it is always possible to split the contig graph into graphs of manageable size by increasing the erosion parameter e.

USING ADDITIONAL INFORMATION

To demonstrate the ability of GRASS to utilise various scaffolding information sources, we used two related genomes (see Fig. 2.5) to help scaffold the *E. coli* assembly: DH10B and BW2952. These genomes were used individually, together and in combination with paired reads. When combining several information sources, care has to be taken in choosing the weights W_r and the erosion threshold parameter e. In individual genome experiments, $W = 100$ and $e = 80$ were chosen to remove links derived from low-quality alignments. In the experiment using only two related genomes (thus no links derived from paired read data) a higher weight was given to the more closely related strain: $e = 70$ and $W_{\text{DH10B}} = 80$, $W_{\text{BW2952}} = 100$ were used for the DH10B and BW2952 strains correspondingly. For experiments combining a single genome with paired reads, $W = 10$ and $e = 4$ were chosen. Finally, $W_{\text{DH10B}} = W_{\text{BW2952}} = 3$ and $e = 5$ were used in the experiment combining all data (including the paired read constraints) to emphasise use of links supported by at least two information sources. When used in the experiment, paired read link weights were set to 1. A standard deviation of 3000 bp was used for links derived from related genomes.

Interestingly, using just related genomes GRASS constructs a smaller number of

Table 2.2: Contiguity and accuracy statistics of the initial assembly of *E. coli* and its scaffolds. Results with the smallest number of breakpoints or scaffolds are shown in bold.

Scaffolder	Breakpoints Δ = 10 kbp	Breakpoints Δ = 500 bp	Number of scaffolds	N50	Maximum length, bp	Total length, bp	Reference covered	Scaffolds covered
Velvet contigs	1	1	481	19,872	73,062	4,535,181	97.44%	99.79%
Velvet scaffolds	9	10	106	171,726	312,219	4,561,490	97.98%	99.74%
SSPACE	15	16	**63**	178,023	374,265	4,547,685	97.79%	99.70%
GRASS	6	6	116	117,964	267,989	4,546,975	97.53%	99.55%
GRASS+	**2**	**2**	120	112,254	268,030	4,546,640	97.53%	99.55%
MIP Scaffolder	11	12	148	89,070	221,548	4,546,430	97.54%	99.59%
OPERA (SRR001665)	36	38	323	32,799	131,842	4,544,447	97.52%	99.67%
OPERA (SRR001666)	26	28	262	37,330	126,797	4,556,203	97.52%	99.42%

Table 2.3: Contiguity and accuracy statistics of the initial assembly of *P. suwonensis* and its scaffolds.

Scaffolder	Breakpoints Δ = 10 kbp	Breakpoints Δ = 500 bp	Number of scaffolds	N50	Maximum length, bp	Total length, bp	Reference covered	Scaffolds covered
Velvet contigs	1	1	303	26,043	90,572	3,394,128	99.01%	99.90%
Velvet scaffolds	4	5	134	57,614	153,169	3,395,148	99.03%	99.78%
SSPACE	2	2	**127**	60,526	151,961	3,388,872	99.09%	99.99%
GRASS	2	2	138	62,908	152,258	3,394,155	99.02%	99.91%
GRASS+	**1**	**1**	144	53,211	151,938	3,389,098	99.02%	99.91%
MIP Scaffolder	31	32	138	52,743	115,278	3,390,104	99.03%	99.93%
OPERA	17	18	184	45,559	186,349	3,413,751	99.01%	99.34%
OPERA (SRR097515)	**1**	**1**	302	26,053	90,582	3,397,028	99.01%	99.81%
OPERA (SRR191848)	4	4	225	34,214	90,582	3,397,065	99.02%	99.84%

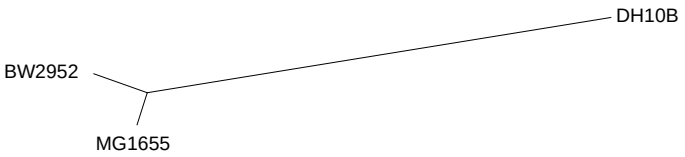

Figure 2.5: Phylogenetic tree showing evolutionary distance between the *E. coli* MG1655 strain and two related strains. Genome sequences were obtained from GenBank.

scaffolds than when only paired reads are used. Table 2.5 shows, however, that this is achieved at the expense of scaffold accuracy: besides having an increased number of breakpoints, scaffolds constructed based on related genomes alone have a high total assembly length and, as a consequence, a low scaffold coverage. The higher than anticipated total assembly length is due to differences in contig distances (i.e., physical distances obtained by aligning contigs to a genome sequence) between the MG1655 strain and the related strains. This is also the reason for the large differences observed between breakpoints at $\Delta = 10$ kbp and $\Delta = 500$ bp: while relative order and orientation have been preserved for large parts of the genomes of the considered strains, the exact physical distances have not. This situation is partially alleviated when information from the two genomes is combined, because (i) consistent links (derived from the two genomes) get higher weights after link bundling, and (ii) the more closely related strain BW2952 was given a higher weight. In this case GRASS is able to further reduce the number of scaffolds without introducing new breakpoints.

Combining paired read data with information from individual related genomes allows for construction of a smaller number of scaffolds with fewer breakpoints than when using these data individually. The results vary between repeated runs of the algorithm, due to inconsistencies between linking information provided by paired reads and related genomes, combined with the stochastic nature of the optimisation strategy used for solving the MIQP formulation. Depending on the intermediate solutions found, different contig links are disabled in the optimisation process, leading to different final solutions and, thereby to different scaffolds. Table 2.5 hence shows a range of scaffold and breakpoint counts, and other results as averages over five repeated runs. This variability is smaller when all data is combined, since a "voting" approach can be implemented by setting W and e in such a way that all links supported by only a single information source have low weights and are ignored. Using all available information, GRASS reduced the number of scaffolds by 40% compared to just using paired reads, at the expense of introducing a single new breakpoint. The increase in the number of breakpoints is not surprising, as the *de novo* scaffolding information is augmented with links derived for a different (related) organism. The best result on combined data is shown in Fig. 2.4A.

2.5. CONCLUSION

We presented GRASS, a generic scaffolding algorithm suitable for combining multiple information sources, as well as GRASS+, incorporating a post-processing scaffolding

Table 2.4: Contiguity and accuracy statistics of the initial assembly of *P. syringae* and its scaffolds.

Scaffolder	Breakpoints Δ = 10 kbp	Breakpoints Δ = 500 bp	Number of scaffolds	N50	Maximum length, bp	Total length, bp	Reference covered	Scaffolds covered
Velvet contigs	1	1	1,560	8,599	46,055	5,902,217	96.41%	99.78%
Velvet scaffolds	25	27	200	122,286	683,615	6,012,535	97.78%	99.37%
SSPACE	21	26	192	87,996	520,403	5,946,936	96.61%	99.09%
GRASS	22	25	243	85,493	618,916	5,931,679	96.57%	99.38%
GRASS+	**6**	**7**	266	77,945	460,726	5,945,096	96.56%	99.01%
MIP Scaffolder	37	47	**167**	94,327	279,875	5,943,358	96.58%	99.17%
OPERA	14	14	705	18,108	76,357	5,950,236	96.58%	99.17%

Table 2.5: Contiguity and accuracy statistics of *E. coli* scaffolds obtained with GRASS+ using additional data. The "≈" sign indicates mean values over 10 repeated runs in cases, when variation was observed.

Reads used	DH10B used	BW2952 used	Breakpoints Δ = 10 kbp	Breakpoints Δ = 500 bp	Number of scaffolds	N50	Maximum length, bp	Total length, bp	Reference covered	Scaffolds covered
yes	no	no	**2**	**2**	120	112,254	268,030	4,546,640	97.53%	99.55%
no	yes	no	10	66	105	425,724	1,948,314	5,047,825	97.51%	89.65%
	no	yes	4	65	90	843,564	1,099,102	4,773,879	97.52%	92.64%
	yes	yes	6	70	81	612,889	1,315,367	4,763,935	97.52%	94.80%
yes	yes	no	3-6	40-49	72-120	≈ 273,503	≈ 850,450	≈ 4,804,124	≈ 97.52%	≈ 94.99%
	no	yes	2-7	51-55	**67-80**	≈ 497,383	≈ 1,077,789	≈ 4,569,001	97.53%	≈ 99.06%
	yes	yes	3	44-46	71-73	≈ 363,105	≈ 988,508	≈ 4,583,534	97.53%	≈ 98.75%

step. Its use was demonstrated by scaffolding genomes based on paired read data and information in related genome sequences, both individually and combined. GRASS achieves the best results when all available scaffolding information is used, as this allows conflicting information from a single source to be ignored when the majority of sources do not support it. Such a mode of operation is supported by the possibility of choosing weights for the individual information sources, combined with the contig link erosion threshold.

We compared GRASS to a number of state-of-the-art scaffolders (SSPACE, MIP and OPERA) on three datasets. GRASS constructs the most accurate scaffolds on all datasets, while keeping the number of scaffolds low. Only SSPACE consistently produces lower numbers of scaffolds, but these are significantly less accurate. The accuracy/contiguity tradeoff displayed by GRASS puts it in a unique niche compared to existing scaffolders.

The current implementation of GRASS supports the use of paired read information and related genomes for scaffolding. However, the algorithm is not limited to any particular set of information sources. We will extend GRASS to allow use of other sources, such as optical restriction maps, RNA-seq and EST data.

2.A. SUPPLEMENTARY INFORMATION

SEQUENCE ASSEMBLY

To select the k-mer length for *de novo* genome assembly using Velvet we tried different values of k and calculated length and accuracy statistics for the resulting assemblies. We measured the number of contigs, maximum contig length, the N50 statistic and total assembly length to get a feel of assembly completeness and contiguity. We also measured *coverage* as percentage of reads mapping to the genome, and *accuracy* as the percentage of paired reads with proper pairing (as defined by BWA, [14]). To measure accuracy and coverage, single- and paired-end mapping of the reads to the assembled contigs was performed using BWA. Tables 2.6, 2.8 and 2.7 show these statistics for different k for *E. coli*, *P. syringae* and *P. suwonensis* assemblies correspondingly.

PHYLOGENETIC TREE CONSTRUCTION

The phylogenetic tree for *E. coli* strains MG1655, BW2952 and DH10B was constructed using the SplitsTree 4 package [31] and the coverage distance function from [32]. Genome alignments were obtained using MUMmer [25] with settings from [33].

SCAFFOLDER RUNNING TIME

Scaffolding and mapping running times were measured for all experiments. This data is presented in Table 2.9. Scaffolding time for Velvet and mapping time for SSPACE have been calculated from the programs' output. Preprocessing of reads prior to mapping and post-processing of the mapper's output was counted as mapping time.

REFERENCES

[1] A. A. Gritsenko, J. F. Nijkamp, M. J. Reinders, and D. de Ridder, *GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies,* Bioinformatics **28**, 1429 (2012).

[2] D. R. Zerbino and E. Birney, *Velvet: algorithms for de novo short read assembly using de Bruijn graphs,* Genome Research **18**, 821 (2008).

[3] J. R. Miller, A. L. Delcher, S. Koren, E. Venter, B. P. Walenz, A. Brownley, J. Johnson, K. Li, C. Mobarry, and G. Sutton, *Aggressive assembly of pyrosequencing reads*

Table 2.6: *E. coli* assembly statistics for different k-mer lengths of Velvet. Assembly for the chosen k is highlighted.

k	Contigs	N50	Maximum	Total length	Coverage	Accuracy
19	4,180	1,621	9,259	4,505,092	91.66%	72.38%
21	1,485	5,466	40,066	4,516,751	93.86%	87.63%
23	951	9,181	41,213	4,521,870	94.47%	90.93%
25	722	12,114	55,230	4,527,423	94.83%	92.51%
27	581	15,644	73,054	4,529,084	95.00%	93.50%
29	512	18,358	71,241	4,531,657	95.16%	94.00%
31	481	19,872	73,062	4,535,181	95.26%	94.21%
33	586	15,104	62,943	4,541,512	95.38%	93.75%
35	11,079	445	2,853	4,245,608	82.96%	36.44%

with mates, Bioinformatics **24**, 2818 (2008).

[4] Y. Peng, H. C. Leung, S.-M. Yiu, and F. Y. Chin, *IDBA–a practical iterative de Bruijn graph de novo assembler,* in *Research in Computational Molecular Biology* (Springer, 2010) pp. 426–440.

[5] N. Nagarajan, T. D. Read, and M. Pop, *Scaffolding and validation of bacterial genome assemblies using optical restriction maps,* Bioinformatics **24**, 1229 (2008).

[6] W. J. Kent and D. Haussler, *Assembly of the working draft of the human genome with GigAssembler,* Genome Research **11**, 1541 (2001).

[7] M. Pop, A. Phillippy, A. L. Delcher, and S. L. Salzberg, *Comparative genome assembly,* Briefings in Bioinformatics **5**, 237 (2004).

[8] A. Dayarian, T. P. Michael, and A. M. Sengupta, *SOPRA: Scaffolding algorithm for paired reads via statistical optimization,* BMC Bioinformatics **11**, 1 (2010).

[9] J. D. Kececioglu and E. W. Myers, *Combinatorial algorithms for DNA sequence assembly,* Algorithmica **13**, 7 (1995).

[10] D. H. Huson, K. Reinert, and E. W. Myers, *The greedy path-merging algorithm for contig scaffolding,* Journal of the ACM (JACM) **49**, 603 (2002).

[11] M. Boetzer, C. V. Henkel, H. J. Jansen, D. Butler, and W. Pirovano, *Scaffolding pre-assembled contigs using SSPACE,* Bioinformatics **27**, 578 (2011).

[12] S. Gao, W.-K. Sung, and N. Nagarajan, *Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences,* Journal of Computational Biology **18**, 1681 (2011).

[13] L. Salmela, V. Mäkinen, N. Välimäki, J. Ylinen, and E. Ukkonen, *Fast scaffolding with small independent mixed integer programs,* Bioinformatics **27**, 3259 (2011).

[14] H. Li and R. Durbin, *Fast and accurate short read alignment with Burrows–Wheeler transform,* Bioinformatics **25**, 1754 (2009).

[15] G. L. Nemhauser and L. A. Wolsey, *Integer and combinatorial optimization* (Wiley-Interscience, 1988).

[16] G. B. Dantzig, *Linear programming and extensions* (Princeton university press, 1998).

[17] J. E. Beasley, *Heuristic algorithms for the unconstrained binary quadratic programming problem,* London, England (1998).

[18] Y. Nesterov *et al.*, *Quality of semidefinite relaxation for nonconvex quadratic optimization* (Université Catholique de Louvain. Center for Operations Research and Econometrics (CORE), 1997).

[19] P. Merz and K. Katayama, *Memetic algorithms for the unconstrained binary quadratic programming problem,* BioSystems **78**, 99 (2004).

[20] P. M. Pardalos, O. A. Prokopyev, O. V. Shylo, and V. P. Shylo, *Global equilibrium search applied to the unconstrained binary quadratic optimization problem,* Optimisation Methods and Software **23**, 129 (2008).

[21] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert, K. A. Remington, *et al.*, *A whole-genome assembly of Drosophila,* Science **287**, 2196 (2000).

[22] D. R. Zerbino, *Genome assembly and comparison using de Bruijn graphs,* Ph.D. thesis, University of Cambridge (2009).

[23] D. S. Hirschberg, *A linear space algorithm for computing maximal common subsequences,* Communications of the ACM **18**, 341 (1975).

[24] D. Vakatov, *The NCBI C++ Toolkit Book,* Online: `http://ncbi.github.io/cxx-toolkit/` (2004).

[25] A. L. Delcher, A. Phillippy, J. Carlton, and S. L. Salzberg, *Fast algorithms for large-scale genome alignment and comparison,* Nucleic Acids Research **30**, 2478 (2002).

[26] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, *et al.*, *The sequence alignment/map format and SAMtools,* Bioinformatics **25**, 2078 (2009).

[27] D. W. Barnett, E. K. Garrison, A. R. Quinlan, M. P. Strömberg, and G. T. Marth, *BamTools: a C++ API and toolkit for analyzing and managing BAM files,* Bioinformatics **27**, 1691 (2011).

[28] ILOG CPLEX, *High-performance software for mathematical programming and optimization,* Online: `http://www.ilog.com/products/cplex` (2005).

[29] A. R. Quinlan and I. M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features,* Bioinformatics **26**, 841 (2010).

[30] B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, *et al.*, *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome,* Genome Biology **10**, R25 (2009).

[31] D. H. Huson and D. Bryant, *Application of phylogenetic networks in evolutionary studies,* Molecular Biology and Evolution **23**, 254 (2006).

[32] S. R. Henz, D. H. Huson, A. F. Auch, K. Nieselt-Struwe, and S. C. Schuster, *Whole-genome prokaryotic phylogeny,* Bioinformatics **21**, 2329 (2005).

Table 2.7: *P. syringae* assembly statistics for different *k*-mer lengths of Velvet.

k	Contigs	N50	Maximum	Total length	Coverage	Accuracy
19	5,059	1,892	12,464	5,846,661	84.47%	64.51%
21	1,926	7,024	42,317	5,886,062	86.70%	80.78%
23	1,560	8,599	46,055	5,902,217	87.20%	82.93%
25	1,990	5,977	24,056	5,930,228	87.55%	81.27%
27	3,829	2,623	13,478	5,946,020	87.32%	72.76%
29	8,825	865	8,433	5,592,074	81.59%	45.63%
31	5,523	343	2,676	1,755,054	28.42%	6.57%
33	57	500	3,166	21,040	1.10%	0.61%
35	15	244	448	3,588	0.24%	0.04%

Table 2.8: *P. suwonensis* assembly statistics for different *k*-mer lengths of Velvet.

k	Contigs	N50	Maximum	Total length	Coverage	Accuracy
21	798	178	672	148,597	1.16%	0.70%
23	3,640	194	609	724,989	6.90%	6.73%
25	6,457	222	900	1,451,717	15.79%	16.58%
27	8,045	264	1,273	2,084,930	25.28%	28.08%
29	8,538	313	1,793	2,522,405	32.97%	37.82%
31	8,306	385	2,421	2,846,252	39.62%	46.66%
33	7,520	482	3,595	3,069,871	45.06%	54.49%
35	6,391	635	3,505	3,220,911	49.30%	61.05%
37	5,270	857	5,770	3,321,047	52.65%	66.44%
39	3,978	1,223	7,233	3,371,436	55.17%	70.96%
41	2,939	1,706	11,487	3,396,276	56.95%	74.35%
43	2,039	2,721	16,786	3,407,475	58.35%	77.06%
45	1,435	3,959	16,772	3,408,865	59.12%	78.75%
47	1,020	5,818	23,722	3,408,282	59.68%	79.90%
49	697	9,367	36,131	3,405,741	60.05%	80.72%
51	537	12,638	46,479	3,402,802	60.21%	81.10%
53	427	16,065	64,878	3,400,488	60.33%	81.40%
55	351	19,866	87,700	3,399,187	60.42%	81.60%
57	308	24,193	87,698	3,396,963	60.49%	81.74%
59	303	26,043	90,572	3,394,128	60.47%	81.74%
61	309	24,862	90,573	3,392,147	60.46%	81.73%
63	301	24,005	78,697	3,386,612	60.46%	81.74%
65	334	21,764	78,707	3,380,022	60.38%	81.63%
67	380	17,029	78,569	3,372,389	60.26%	81.44%
69	462	13,262	74,778	3,363,394	60.10%	81.18%
71	648	9,303	54,433	3,351,627	59.81%	80.67%
73	1,088	5,308	22,390	3,338,680	59.36%	79.71%
75	4,214	933	13,128	3,082,996	53.13%	68.00%

[33] A. F. Auch, H.-P. Klenk, and M. Göker, *Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs,* Standards in Genomic Sciences **2**, 142 (2010).

Table 2.9: Scaffolder and mapping running time. For *E. coli* "(all)" denotes usage of paired reads and related genomes of *E. coli* strains DH10W and BW2952 for scaffolding.

Dataset	Scaffolder	Mapping time, min	Scaffolding time, min	Total time, min
E. coli	Velvet	N/A	8 sec	8 sec
	SSPACE	2 m 48 sec	1 m 7 sec	3 m 11 sec
	GRASS	29 m 55 sec	23 sec	30 m 18 sec
	GRASS+	29 m 55 sec	53 sec	30 m 48 sec
(all)	GRASS+	47 m 16 sec	40 sec	47 m 56 sec
	MIP Scaffolder	68 m 49 sec	2 m 2 sec	70 m 52 sec
SRR001665	OPERA	21 m 11 sec	27 m 45 sec	48 m 56 sec
SRR001666	OPERA	27 m 49 sec	30 sec	28 m 19 sec
P. suwonensis	Velvet	N/A	13 sec	13 sec
	SSPACE	5 m 8 sec	7 m 22 sec	12 m 3 sec
	GRASS	139 m 59 sec	23 sec	140 m 23 sec
	GRASS+	139 m 59 sec	45 sec	140 m 44 sec
	MIP Scaffolder	95 m 37 sec	1 m 1 sec	96 m 37 sec
	OPERA	125 m 28 sec	8 m 19 sec	133 m 47 sec
SRR097515	OPERA	74 m 56 sec	25 sec	75 m 22 sec
SRR191848	OPERA	75 m 32 sec	1 m 53 sec	77 m 25 sec
P. syringae	Velvet	N/A	1 sec	1 sec
	SSPACE	1 m 6 sec	27 sec	1 m 33 sec
	GRASS	13 m 20 sec	15 sec	13 m 35 sec
	GRASS+	13 m 20 sec	3 m 7 sec	16 m 27 sec
	MIP Scaffolder	9 m 19 sec	27 sec	9 m 46 sec
	OPERA	10 m 38 sec	72 m 22 sec	83 m 1 sec

3

CODON OPTIMISATION THROUGH PREDICTIVE MODELLING

Alexey A. GRITSENKO, Marcel J.T. REINDERS and Dick DE RIDDER

This chapter has been published in *Pattern Recognition in Bioinformatics* (2013) pp. 159–171 [1].

3

ABSTRACT

Given recent advances in synthetic biology and DNA synthesis, there is an increasing need for carefully engineered biological parts (e.g. genes, promoter sequences or enzymes) and circuits. However, forward engineering approaches are thus far rarely used in biology due to lack of detailed knowledge of the biological mechanisms. We describe a framework that enables forward engineering in biology by constructing models predictive of properties of interest, then inverting and using these models to design biological parts.

We demonstrate the applicability of the proposed framework on the problem of codon optimisation, concerned with optimising gene coding sequences for efficient translation. Results suggest that our data-driven codon optimisation (DECODON) method simultaneously considers the effects multiple translation mechanisms to produce optimal sequences, in contrast to existing codon optimisation techniques.

3.1. INTRODUCTION

In biotechnology, microorganisms such as yeast are genetically engineered for improved production of foods, beverages, fuels and pharmaceuticals. Recent advances in synthetic biology and dropping cost of DNA synthesis have led to a growing need for methods to engineer biological parts (promoter regions, gene *coding sequences* (CDSs) and even entire enzymes) with specific properties. Whereas in many engineering disciplines optimisation techniques are routinely used to design such parts (e.g., aircraft wings [2]), in synthetic biology this is not yet the case. This stems from a lack of fundamental biological knowledge on the processes in which these parts are involved.

For some problems, this limitation can be overcome by constructing predictive models for properties of biological parts (e.g., promoter strength, mRNA translation rate or enzyme activity) and inverting the constructed models to design biological parts with desired properties. A successful use of such a "black-box" modelling approach would enable forward engineering in areas of biology where detailed knowledge of the underlying processes is unavailable. We showcase the use of our proposed framework on the problem of codon optimisation, in which a gene coding sequence is changed to obtain a desired translation rate of the mRNA into protein while keeping the amino acid sequence intact.

The degeneracy of the genetic code manifests itself in the differential use of synonymous codons in different organisms and different genes in the same organism. It has been long noticed that organisms preferentially use just one or two codons out of a family of codons translated into the same amino acid. This preference, termed *codon usage bias* (CUB), is more pronounced in highly expressed genes, which sometimes exclusively use only the preferred codons. For this reason it is believed that in unicellular organisms, such as baker's yeast *Saccharomyces cerevisiae* and the bacterium *Escherichia coli*, the codon bias of a gene is related to its translation rate [3]. Over the years numerous methods (called *indices*) summarising the degree of CUB of a gene in a single number have been proposed and have been demonstrated to correlate with intracellular mRNA and protein levels [4].

These correlations have been used in a process called *codon optimisation* to modify gene CDSs such that their translation rate is maximised, by introducing synonymous codon substitutions which increase one of the codon indices [5]. Codon optimisation is routinely applied in biotechnology to overexpress genes for heterologous protein production and heterologous pathway expression [6]. However, CUB only partially explains the difference in translation rates among genes. Although the precise mechanisms influencing gene translation rates are not known, there is evidence suggesting that codon pair usage, tRNA recycling [7], mRNA secondary structure [8], adaptation to an organisms tRNA pool, mRNA untranslated regions (UTRs) and protein amino acid charge [8] may influence translation initiation and elongation rates. The relative influence of these factors on translation is not understood, making it difficult to combine them in a single codon optimisation strategy. To our knowledge only Maertens *et al.* [9] have successfully combined multiple codon optimisation objectives, by equally weighting them.

We present DECODON (data-driven codon optimization), an approach to codon optimisation that combines multiple optimisation objectives in a data-driven way by

constructing a regression model. We use *Support Vector Regression* (SVR) [10] to predict *ribosome density*, a measure related to translation rate, based on coding sequence features of *S.cerevisiae* genes. We then invert this predictor by using it inside a genetic algorithm to optimise gene CDSs for desired ribosome density.

3.2. MATERIALS AND METHODS

DATASET

To our knowledge no datasets with direct measurements of translation rates are available. However, Ingolia *et al.* [11] performed genome-scale measurements of average *ribosome density*, defined as the number sequencing reads originating from parts of mRNA molecules covered by ribosomes in all mRNA copies of a particular gene, divided by the length of the gene transcript. Ribosome density is indicative of translation rate, as genes with higher densities are expected to produce more protein per copy of mRNA.

The number of gene mRNA copies per cell depends on its transcription rate and the stability of its mRNA. Although the relationship is poorly understood, the latter may be influenced by the secondary structure of the mRNA, which can differ between synonymous (i.e., encoding the same peptide) versions of a gene. In order to take the potential influence of coding sequence on the transcript levels into account, we propose to directly (i.e., without normalising by the mRNA *read density*) use ribosome density as a measure of gene translation rate.

Yeast gene CDSs were obtained from the Saccharomyces Genome Database and the matching 5′- and 3′-UTR sequences were obtained from Nagalakshmi *et al.* [12] and Yassour *et al.* [13] (preference given to the former in cases when the two studies were not in agreement). The resulting dataset contains of 5,048 yeast genes, each associated with coding and UTR sequences and a measured ribosome density.

SEQUENCE FEATURES

In order to construct a predictor of ribosome density from gene sequences a number of candidate sequence-based features identified from the literature have been computed for each gene in the dataset. These features were then used in a multivariate regression training step. Selected candidate features (Table 3.1) include a subset of existing codon bias indices (13 features); protein indices and protein properties (12 features); and nucleotide, codon and amino acid composition features (122 features). Prior to training, features as well as the ribosome density to be predicted were standardised to zero mean and unit variance.

REGRESSION MODEL TRAINING

ϵ-SVR [17] has been chosen as a regression method as it supports nonlinear regression through the use of kernels, allowing for complex models, and because efficient training algorithms are available. SVR relies on the choice of several parameters, including the cost parameter C, the error in sensitivity ϵ, the regression kernel and its parameters. Often, due to the lack of a theoretical framework for choosing these parameters, a grid search approach is used to find a combination of parameters that minimises the regression error. This training procedure, if performed inside cross-validation (CV),

Table 3.1: Sequences-based features used as initial input for regression model training. CF and SF respectively stand for the number of candidate features in the feature group and the number of features selected for the final ribosome density predictor. Description of codon indices can be found in Cannarozzi and Schneider [4].

Name	Description	SF	CF
CAI	*Codon Adaptation Index* measures the extent to which a gene is composed of codons from the highly expressed genes.	0	1
tAI	*tRNA Adaptation Index* measures the extent to which a gene consists of codons recognised by abundant tRNAs. It is computed for the full CDS and its first 14, 17 and 19 codons (tAI, tAI_{14}, tAI_{17} and tAI_{19} respectively) [8].	3	4
N_c	*Effective number of codons* estimates the number of uniformly used codons that would produce the CUB observed in a gene.	0	1
D_{ncu}	*Distance to native codon usage* [14] measures the difference between codon usage of a gene and the overall codon usage of the organism.	1	1
E_w	*Weighted sum of relative entropy* measures the degree of deviation from equal usage of synonymous codons using the Shannon entropy.	1	1
CPB	*Codon Pair Bias* score [15] is computed as the sum of log-ratios of observed and expected codon pair counts.	0	1
TPI_2	*tRNA Pairing Index* measures the extent of potential tRNA re-use during gene translation.	1	1
F_{op}	For computing the *Frequency of optimal codons*, optimal codons were chosen as corresponding to the most abundant tRNA species.	1	1
RCBS	*Relative codon usage bias* measures codon usage difference of a gene with respect to the its nucleotide composition.	0	1
P_1	Mean number of non-specific tRNA interactions per elongation cycle.	1	1
prot	Protein hydrophobicity, aromaticity, aliphatic and instability indices.	3	4
Q_{port}	Protein net charge, isoelectric point and weight.	3	3
Q_{side}	Mean amino acid side chain charge computed for the full protein and its first 4, 11, 15 and 40 amino acids [8].	0	5
len	Lengths of the CDS, the 5′- and the 3′-UTR regions.	3	3
nuc	Nucleotide and dinucleotide frequencies of the CDS regions.	7	20
GC_{15}	GC-content computed for the first 15 codons of the CDS	1	1
RSCU	*Relative Synonymous Codon Usage* is computed for each codon (except ATG) as the ratio between the observed number of its occurrences and the mean number of occurrences for codons encoding the same amino acid.	41	63
codon^2	tAI and CAI weights of the second codon in the CDS (denoted tAI^2 and CAI^2).	2	2
amino	Amino acid frequencies.	6	21
ΔG	Gibson free energy for mRNA secondary structures predicted by the Vienna RNA package [16]. It is computed for the 5′-/3′-UTR sequences; and the first 17, 34, and 53 codons of the CDS [8] with ($\Delta G_{\text{5′-UTR,CDS}_{17}}$, $\Delta G_{\text{5′-UTR,CDS}_{34}}$ and $\Delta G_{\text{5′-UTR,CDS}_{53}}$) and without ($\Delta G_{\text{CDS}_{17}}$, $\Delta G_{\text{CDS}_{34}}$ and $\Delta G_{\text{CDS}_{53}}$) 5′-UTR sequence	4	12

becomes computationally very expensive.

As a performance measure we calculate the coefficient of determination R^2. Normally this measure approaches 1 with increasing model complexity regardless of its validity and is therefore not suitable for assessing quality of complex (nonlinear, many features) models. However, if the coefficient of determination is computed using CV (denoted R^2_{CV}), it becomes a measure of the amount of variance in *unseen* data explained by the model. Similar to the coefficient of determination computed without CV, the

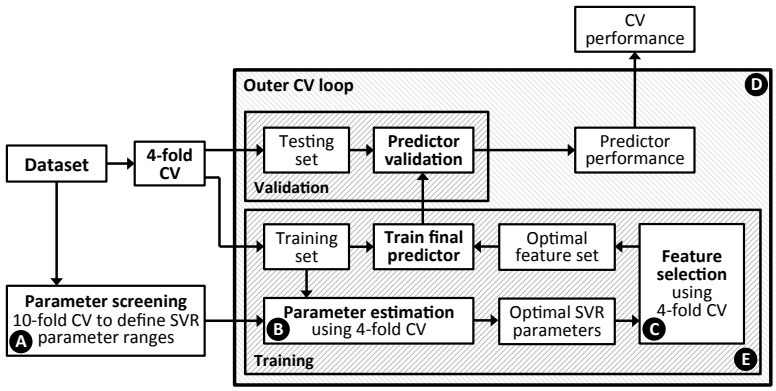

Figure 3.1: Predictor training and evaluation scheme (adapted from [18]). The full dataset is used to preselect SVR parameter ranges (block **A**) and evaluate the training protocol using CV (block **D**). Predictor training consists of parameter estimation (block **B**) used to find an optimal set of SVR parameters, for which feature selection is performed (block **C**). The optimal parameters and the selected features are used to train the final predictor which is evaluated on the testing set of the CV loop. The same training procedure (block **E**) is used to train the final predictors used for sequence optimisation on the *complete* dataset.

cross-validation R^2_{CV} approaches 1 as *generalisation* becomes better, but can be negative if the trained model explains less variance in unseen data than a constant model. We believe that R^2_{CV} is a suitable measure for assessing quality of nonlinear models and use it to optimise and assess performance of our regression models.

PARAMETER PRESELECTION

To keep the amount of computation tractable, we first *screened* the parameter space by training predictors with different parameter settings and assessing their coefficient of determination computed by 10-fold CV (R^2_{10CV}) on the complete dataset (Fig. 3.1, block A). Screening results (data not shown) indicated that the performance of RBF and polynomial kernels on the considered dataset is comparable, which led us to consider only polynomial kernels $K(u,v) = (\gamma \cdot \langle u,v\rangle + 1)^d$ with degrees $d = 2,3,4$ for the actual parameter selection stage. Based on the screening R^2_{10CV} results, ranges for parameters C, γ and ϵ were set to $\{1\} \cup \{0.001 \cdot 3^i\}$ for $i = 0,\ldots,6$.

PARAMETER ESTIMATION

The preselected parameter ranges were used to estimate optimal SVR parameter settings (Fig. 3.1, block B) in a grid search procedure. For each combination of parameters an SVR is trained and its R^2_{4CV} is computed to select a *single* combination of SVR parameter settings with the best performance. This combination is then used in the subsequent feature selection step.

FEATURE SELECTION

Feature selection was used to eliminate features that do not contribute to the model's generalisation capability. This also allowed for selecting a concise set of features

which can be interpreted biologically. While generally yielding good results, wrapper approaches to feature selection are computationally very demanding. To lower the computational load, backward feature elimination [19] was performed only on the SVR parameter settings obtained as discussed above (Fig. 3.1, block C). At every step of the feature elimination procedure, given n features, we computed R^2_{4CV} for n predictors trained on subsets of $n-1$ features (i.e., obtained by removing one of the features). A subset with the highest R^2_{CV} was then selected for the next step of the feature elimination procedure. After the procedure was complete, the number of features (and the corresponding subset) with the best performance was chosen. If multiple subsets gave optimal performance, the smallest one was selected. The selected features were used to train the final predictor on the available data (Fig. 3.1, block E).

TRAINING STRATEGY EVALUATION

In order to obtain an unbiased estimate of the predictor performance we used a second 4-fold CV loop (Fig. 3.1, block D) around the described parameter estimation and feature selection strategies. The R^2_{4CV} values computed in the outer CV loop are reported in Section 3.3 as estimates of predictor generalisation.

SEQUENCE OPTIMISATION

In order for the constructed predictor $y = f(x)$ to be useful for sequence optimisation, it first needs to be "inverted" such that it can be used to find sequences x that have the desired ribosome density $\check{y}$. Constructing the inverse function $x = f^{-1}(y)$ for SVR is impossible. Moreover, solving this function for a given $\check{y}$ would yield multiple non-synonymous sequences x, thereby presenting an additional problem of selecting the suitable sequences from a large pool of solutions. Instead we implicitly invert the predictor by searching through the space of sequences x_i synonymous to the original sequence x to find $\check{x}$ such that its predicted ribosome density $f(\check{x})$ is close to the desired $\check{y}$.

GENETIC ALGORITHM

The space of all nucleotide sequences synonymous to a given sequence x grows exponentially with the length of the sequence. Typically, it is too large to evaluate all possible x_i and requires an efficient search strategy to find (an approximation of) $\check{x}$ in a timely manner. Genetic algorithms (GAs), specifically tailored for large discrete optimisation problems, use computational equivalents of genetic crossover, mutation and selection concepts from biological systems to evolve a pool of potential solutions to a given optimisation problem. The problem of finding an $\check{x}$ whose predicted ribosome density $f(\check{x})$ is as close as possible to a desired level $\check{y}$ can be cast into an optimisation problem and tackled using GAs if $g(x) = |f(x) - \check{y}|$ is used as an objective to be minimised.

In practical applications, optimised gene sequences are synthesised and cloned into living cells in the wet lab. It is then required that the sequences do not contain certain motifs, such as restriction sites of enzymes used in cloning. This presents an optimisation constraint that has to be taken care of by the GA. Treating this constraint as an additional objective of minimising the number of undesired motifs present in the sequence allows to refrain from banning parts of the search space at the cost of

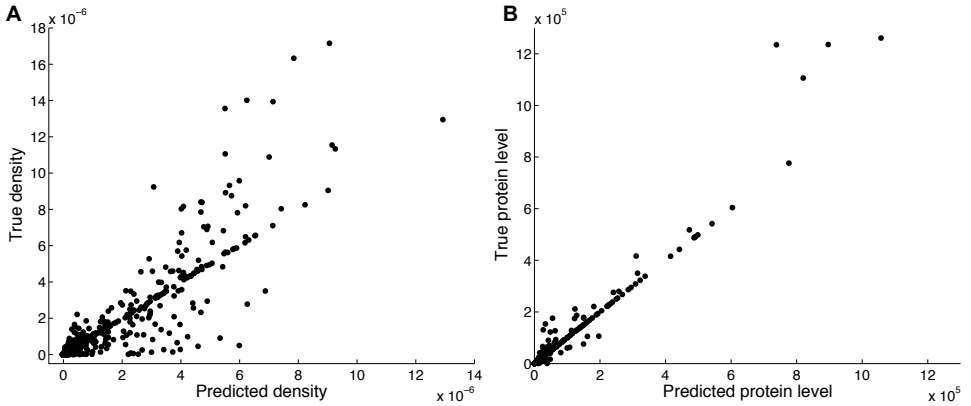

Figure 3.2: Predicted vs. true (**A**) ribosome density and (**B**) protein level plotted for *S.cerevisiae* genes.

casting the problem of finding $\check{x}$ into a multi-objective discrete optimisation problem with two objectives. If it exists, the solution to the original problem will then be among the non-dominated solutions (i.e., solutions that cannot be improved in both objectives simultaneously) of the multi-objective optimisation problem.

NSGA-II [20], a multi-objective GA, was chosen to solve the optimisation problem as previously it has been successfully applied to DNA sequence optimisation. It was implemented using multi-point crossover with a rate of 0.9; a mutation operator synonymously changing every sequence codon with probability $\frac{1}{n}$, where n is the number of degenerate codons in the sequence; and a binary tournament selection operator. For the genes optimised in this paper, the number of crossover points was set to 100.

3.3. RESULTS

REGRESSION MODEL

The cross-validation loop used to evaluate the regressor training strategy described in Section 3.2 gave an $R^2_{4CV} = 0.66 \pm 0.03$, suggesting that the proposed strategy produces regressors that generalise well on unseen data. This strategy was employed to train the *final* ribosome density predictor (shown in Fig. 3.2A) for use in codon optimisation on the complete dataset.

SELECTED FEATURES

The final predictor contained 78 features (Table 3.1, Fig. 3.3), including codon indices, protein features, sequence composition and mRNA structure features selected to best explain the data. While black-box predictors are generally hard to interpret in biological terms, the fact that a certain feature was selected in the final predictor suggests that the mechanism it describes could indeed be used by the translation machinery. In this way, selection of the tRNA Pairing Index (TPI_2) suggests presence in yeast of a tRNA recycling mechanism, in which outgoing tRNA molecules stay bound to the ribosome

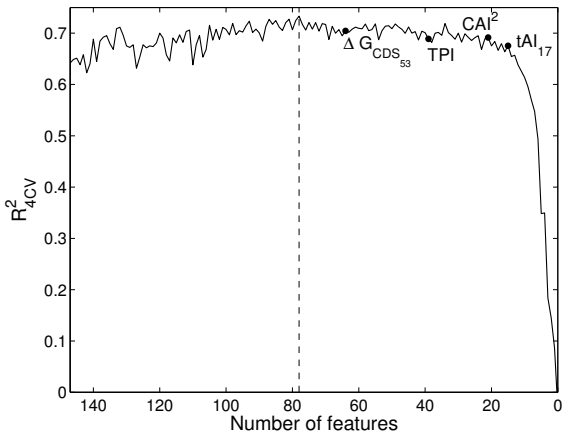

Figure 3.3: Cross-validated R^2_{4CV} for the backward feature elimination procedure during final predictor training. Features eliminated at a particular step are marked with black circles. The maximum R^2_{4CV} is achieved at 78 features (see Table 3.1).

to be recharged and reused in the course of translation [4]. Selection of the CAI^2 and tAI^2 features, describing respectively the extent to which the second codon of a gene is used in highly expressed genes of *S.cerevisiae* and its adaptation to the organisms tRNA pool, suggests that choice of the second codon influences ribosome density. Fredrick and Ibba [21] observe that the second codon is usually a highly frequently used codon that is translated more quickly, and speculate that this mechanism may be required for efficient recycling of the initiator tRNA.

Similarly, the selected tAI_{17}, tAI_{19}, and the $\Delta G_{5'\text{-UTR,CDS}_{17}}$, $\Delta G_{5'\text{-UTR,CDS}_{53}}$ and $\Delta G_{\text{CDS}_{53}}$ features suggest that the mechanism of slowly translated "ramp" in the beginning of the CDS [8] influences gene translation rate. It is believed that the role of this "ramp" is to generate space between translating ribosomes and thereby prevent ribosome collision [8, 21]. The same mRNA structure features also describe the accessibility of the 5′-UTR for translation initiation by the ribosome machinery, suggesting it as another *S.cerevisiae* mechanism influencing gene translation.

CODON OPTIMISATION

The final ribosome density predictor (Section 3.3) was used to optimise sequences of the genes *4CL* (*4-coumaric acid-CoA ligase*, 562 codons) and *PAL1* (*phenylalanine ammonia lyase*, 726 codons) involved in flavonoid biosynthesis [6]. The genes' cDNA, obtained from the plant *Arabidopsis thaliana*, was optimised using the described GA for *maximum* ribosome density. Based on preliminary experiments, optimisation was performed for 200 generations with a population size equal to the gene length in codons. An initial population was generated by back-translating genes from their amino acid sequences by choosing codons with probabilities proportional to their CAI weights.

Table 3.2: Sequence optimisation results for the *4CL* and *PAL1* genes. Predicted ribosome densities are shown for the plant cDNA, sequences codon-optimised using JCat [5] and sequences optimised using DECODON. The number of different codons and the fold increase in the predicted density are computed relative to the cDNA sequences.

	4CL			*PAL1*		
Type	Different codons	Predicted density	Fold inc.	Different codons	Predicted density	Fold inc.
cDNA	N/A	0.0000000090	1	N/A	0.0000000524	1
JCat	338 (60.14%)	0.0000101491	1128	414 (57.02%)	0.0000079718	152
DECODON	361 (64.23%)	0.0000201560	2240	444 (61.16%)	0.0000172657	329

The 5′- and 3′-UTR sequences were set based on the respective sequences of the GPD promoter and CYC1 terminator sequences used in the pAG416GPD yeast expression vector. The *SpeI* and *XhoI* restriction site sequences used for cutting the expression vector were treated as undesired motifs.

Table 3.2 shows that the predicted ribosome density of the optimised sequences is significantly higher than that of the plant cDNA. As a sanity check, we compared sequences optimised using our method DECODON to sequences optimised by JCat [5], a well-known codon optimisation tool that optimises sequences for high CAI. The constructed predictor also predicts a significant increase in ribosome density for the JCat-optimised sequences (Table 3.2), showing that the trained predictor agrees with the currently used codon optimisation methods. Note that the predicted ribosome density for the DECODON-optimised sequences is nearly two-fold higher than that of the JCat-optimised sequences.

SEQUENCE ANALYSIS

Compared to the cDNA sequences, the DECODON- and JCat-optimized versions have roughly the same number of codon substitutions. To highlight the specific differences between the sequences, we compared them to each other. It can be seen from Fig. 3.4 that codon usage in the DECODON sequences is more similar to that of the JCat-optimized genes than to that of the original sequences.

When optimised for maximum ribosome density, codon usage of the optimised sequences follows the “one amino acid - one codon” rule meaning that for each amino acid only a single (preferred) codon is used to encode it. The preferred codons in the genes optimised by DECODON mostly correspond to the codons with high CAI weights (the JCat- and density-optimised *4CL* and *PAL1* genes differ only in 126 and 150 codons respectively) with a few notable exceptions: (i) ACC is preferred for the amino acid threonine; (ii) GTC is preferred for valine; (iii) TGC is preferred for cysteine; and (iv) ATT is preferred for isoleucine.

The preference rules account for all but a few codon differences (underscored in Fig. 3.4) between the optimised sequences. These substitutions, when introduced in the sequences optimised using the “one amino acid - one codon rule”, influence codon indices and mRNA features ($\Delta G_{\text{CDS}_{53}}$ and $\Delta G_{5'\text{-UTR,CDS}_{53}}$), according to which the mRNA secondary structures at the 5′-UTR become less stable. This further suggests that the

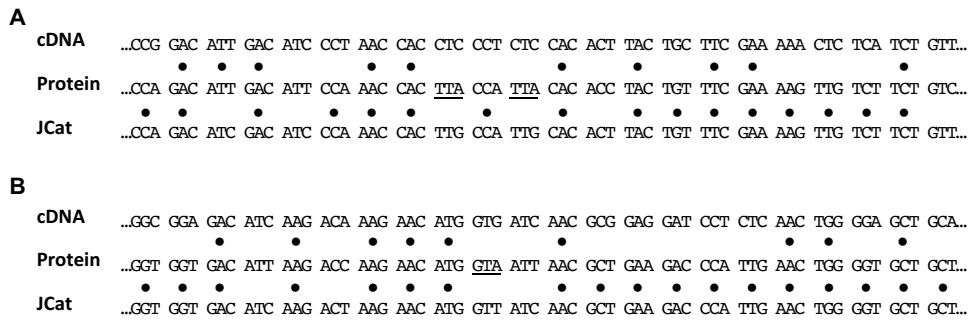

Figure 3.4: Comparison of codons 20 to 41 of sequences of the (**A**) *4CL* and (**B**) *PAL1* genes codon-optimised by JCat and DECODON for ribosome density. Matching codons are marked with black circles. Underscored codons are not explained by the "one amino acid - one codon" rule.

constructed predictor takes into account multiple translation mechanisms, even when used to optimise genes for maximum ribosome density.

APPLICABILITY TO OTHER DATASETS

To demonstrate the applicability of the framework proposed in this paper to different datasets, we used it to optimise codon use based on the predicted absolute protein level measurements of 756 proteins [22]. All the training steps (parameter preselection, training strategy evaluation and final predictor training) were repeated, yielding an cross-validation $R^2_{4CV} = 0.65 \pm 0.09$ and a final predictor with 138 features (Fig. 3.2B). This large number of features, explained by the relatively high variance in the R^2_{4CV} used for feature selection due to the limited size of the dataset, hampers further biological interpretation.

The *4CL* and *PAL1* gene sequences optimised for maximum protein levels using the constructed predictor show a "one amino acid - one codon"' rule behaviour similar to the density-optimised genes with several differences: (i) TGT is preferred for cysteine (as in JCat); (ii) ATC is preferred for isoleucine (as in JCat); and (iii) GCT and GCC are preferred for alanine. Similarly, these rules explain all but a few codon substitutions near to the 5′ end of the CDS (Fig. 3.5). The codon usage similarities between the protein- and density-optimised gene sequences show that the proposed framework can be applied to various types of biological data to enable forward engineering approaches. However, wet-lab experiments are required in order to determine which of the constructed predictors is better suited for codon optimisation.

3.4. DISCUSSION

We have described a generic framework for forward engineering of biological systems and demonstrated its use by optimising genes for maximum ribosome density and maximum protein levels using predictors constructed from the corresponding yeast datasets. The general agreement between the optimised gene sequences obtained by us and gene sequences optimised using an existing codon optimisation method

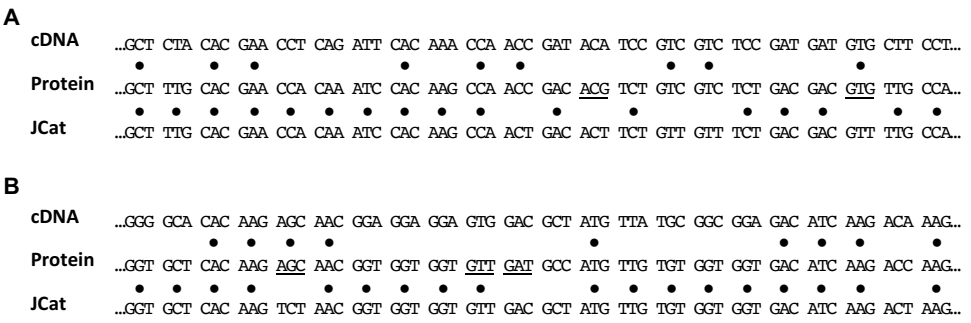

Figure 3.5: Comparison of codons 5 to 26 of sequences of the (**A**) *4CL* and (**B**) *PAL1* genes codon-optimised by JCat and DECODON for absolute protein levels.

suggests that the proposed approach can be successfully utilised for forward engineering of biological parts, whereas the differences between the sequences suggest that our codon optimisation method DECODON simultaneously considers the effects of multiple translation mechanisms to produce optimal sequences. Time complexity of DECODON is much higher than that of JCat, however, it is negligible compared to the time involved in ordering and experimenting with the synthesised DNA.

Features selected for the final ribosome density predictor and the exceptions to the "one amino acid - one codon" rule in the optimised sequences show that data-driven models can combine multiple features describing (competing) biological mechanisms in a way that best explains the available data. While the effect of combining multiple mechanisms in a single predictor is hard to observe in sequences optimised for maximum ribosome density (or protein level), we believe that it would be more pronounced in sequences optimised for intermediate ribosome density, in which no one single mechanism would have a dominating influence.

Using black-box models for combining multiple (potential) mechanisms in a single predictor is particularly useful in areas where precise workings of a system are not known, but hypotheses on its important aspects can be generated and described by features. Note that a danger associated with the interpretation of the results is that the constructed model will select features that correlate with the property it is trained to predict, rather than the features describing the actual underlying mechanisms. For example, Qian *et al.* [14] suggest that strong CUB in highly expressed genes is not related to translation rate of those genes, but is rather a consequence of random mutations and the evolutionary pressure to keep codon usage and tRNA availability of an organism balanced. Nevertheless our models exhibit the "one amino acid - one codon" behaviour when genes are optimised for maximum density/protein levels. It is, therefore, crucial to validate predictive models by testing their predictions in the wet-lab prior to their application.

For the constructed predictors (especially in the case of the protein level predictor) we observed that a single codon substitution often leads to changes in many features. These changes are often difficult to interpret and to link to the effect a substitution has

on the prediction. Nevertheless, we believe that by trading interpretability for general applicability, our framework will enable forward engineering of various parts essential for synthetic biology such as promoters, coding sequences and UTRs.

3.A. *Addendum*: EXPERIMENTAL VALIDATION

INTRODUCTION

Above we proposed DECODON, a data-driven codon optimisation method that aims at simultaneously considering the effects of multiple translation mechanisms during the optimisation process. It achieves this by optimising for the net effect of these mechanisms on the resulting expression, which is estimated by a predictor of ribosome density trained on gene sequence features that capture various mechanisms employed in the process of translation. The learned predictor showed a good fit to the ribosome density data (CV R^2 = 0.66), and sequences of genes codon-optimised by this predictor demonstrated considerable agreement with the previously established CAI-based method JCat [5], while yielding higher predicted expression. Together, this suggests that DECODON-optimised genes can potentially achieve higher expression than genes optimised solely using the CAI or similar metrics. However, because superior *in silico* performance does not guarantee successful applications to gene re-design, we sought to further validate the developed method by measuring expression of genes synthetically designed using our method. We chose to re-design the *PAL1* (*phenylalanine ammonia lyase*) gene from the flavonoid biosynthetic pathway due to its importance to the ongoing project of yeast flavonoid production [6], and availability of enzymatic activity assays [23]. Here, we report on the comparison of enzymatic activity between the different versions of the *PAL1* obtained using our method and JCat.

RESULTS

To assess the applicability of DECODON for optimising genes for high protein expression, we used it to re-design the *PAL1* cDNA obtained from *Arabidopsis thaliana* for maximum predicted ribosome density as in Chapter 3 (*dcPAL1*; Table 3.2 and Figure 3.4B), and compared it to the JCat-optimised version (*jcPAL1*) and the original *A. thaliana* cDNA (*atPAL1*) [6]. All versions of the enzyme were expressed using the TDH3 promoter on a centromeric plasmid (see Tables 3.3 and 3.4) in *Saccharomyces cerevisiae* yeast strain CEN.PK 113-5D, and had the hemagglutinin (HA) epitope attached at the C-terminus. To confirm that the HA tag does not have a significant impact on expression, we also measured expression of the *jcPAL1* version of the enzyme without the tag (*jcPAL1* no tag).

As measuring protein levels directly is not trivial, specific enzymatic activity of the *PAL1* enzyme was used as a proxy for its expression (see Materials and Methods, below). The use of enzymatic activity as a measure of protein expression additionally ensures that gene re-designs do not render the protein non-functional, and thus presents a more stringent measure of the effect of codon optimisation than protein levels. Enzymatic activity measurements (Figure 3.6) confirmed that the HA tag does not have a significant impact on *PAL1* activity; and showed that genes codon-optimised by either method achieve higher activity than the *A. thaliana* wild type sequence, suggesting

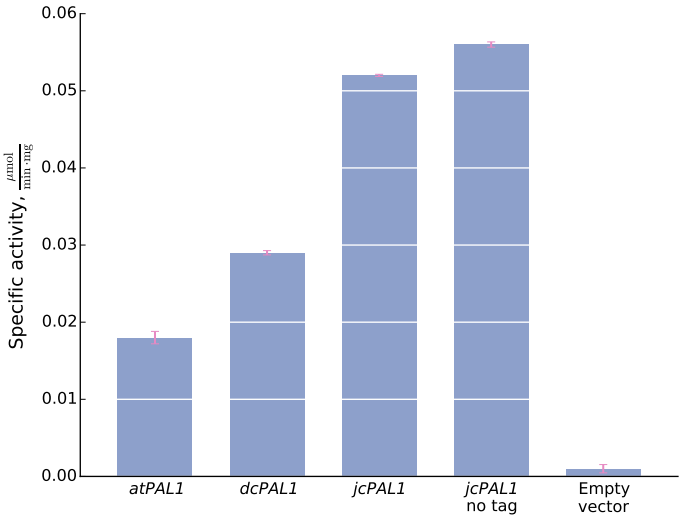

Figure 3.6: Specific activity of several versions of the *PAL1* enzyme. Standard deviations are calculated from two technical replicate measurements.

that both methods increase protein expression, presumably by increasing its translation rate. However, the results also revealed that the JCat-optimised gene achieves higher enzymatic activity than its DECODON counterpart, suggesting that translation rates achieved by our method are lower than that of the CAI-based JCat. We propose potential explanations of this result in the following section.

DISCUSSION

Comparison of the enzymatic activity of different versions of the *PAL1* gene suggests that both, our method and the CAI-based JCat, improve heterologous protein expression of the *A. thaliana* cDNA sequence in *Saccharomyces cerevisiae*. This result is in line with the goal of DECODON to optimise gene sequences for efficient translation, and thus serves as an initial validation of our method. However, we also observed that *jcPAL1* showed consistently higher activity than *dcPAL1*, thus suggesting that gene re-design using JCat yielded a higher translation rate than our method. This result is surprising in the context of our approach, which aims at optimising genes for the net effect of the various mechanisms on translation, and should have lead to gene re-design with expression comparable or better to that of JCat. We identified several possible reasons for this discrepancy.

In our experimental validation we use enzymatic activity of the *PAL1* gene as a measure of its expression, thus assuming that the introduced enzyme sequence changes do not adversely affect its mRNA levels or function. While it is unlikely that mRNA levels are affected by sequence changes aimed at increasing translation rates [24], it is

possible that changes introduced by DECODON affected the enzyme folding in a way that reduced its activity [3]. To gain more insight into the entire process of protein expression of the re-designed genes, and to rule out the possibility that *PAL1* activity was affected by changes in mRNA levels or enzyme functionality, the experiments performed here should be complemented by measurements of mRNA and protein levels (e.g., measured by qPCR and ELISA).

However, a more likely explanation for the higher expression of *jcPAL1* compared to *dcPAL1* is related to the data used for developing the DECODON method. As described in Chapter 3, DECODON uses regression to learn the relationship between gene sequence features and the resulting net effect of translation efficiency (TE) changes. The net effect of TE changes is challenging to measure, and the total ribosome density of a gene as measured by [11] was used as a proxy for it. This approach has several limitations:

1. The predictor learned by DECODON was trained on the entire set of genes with measured ribosome densities [11]. The expression of these genes spans multiple orders of magnitude, but their distribution is skewed with relatively few genes having very high expression levels (see Chapter 3, Fig. 3.2). This limits the number of examples of highly expressed genes available for training, and thus also the certainty of the corresponding predictions. The *dcPAL1* gene was optimised for highest expression, and is thus potentially affected by the prediction uncertainty. To test relationship between prediction accuracy and the predicted expression level, several DECODON-based re-designs of the *PAL1* gene should be produced and validated.

2. The predictor was trained on the first publicly available ribosome profiling dataset [11]. At the time, experimental biases of the ribosome profiling technique were poorly understood. Later studies demonstrated that cycloheximide, a chemical used in the majority of ribosome profiling protocols, is responsible for the accumulation of ribosome reads at the beginning and upstream of coding sequences [25], and for altering A-site occupancies of many codons [26]. These biases affect overall gene ribosome profiles, and thus limit the ability of the regressor used by DECODON to learn the relationship between sequence and TE. This limitation could be overcome by applying the DECODON method to the more mature ribosome profiling datasets [27], and comparing the resulting predictors.

3. Ribosome density does not allow distinguishing between the situations (i) of high ribosome density due to efficient translation, and (ii) high ribosome density due to ribosome stalling. Optimising genes for high ribosome density as a proxy for high TE may thus actually optimise for ribosome stalling and result impaired translation. While it is unlikely that under normal growth conditions many native *S. cerevisiae* genes (the data used to train the DECODON predictor) undergo ribosome stalling, inability to distinguish between the two situations of high ribosome density may still be responsible for the lower expression of *dcPAL1* compared to *jcPAL1*.

4. Although the aim of DECODON is to optimise gene sequences at the level of translation, to capture the overall effect of translation efficiency changes on

Table 3.3: *Saccharomyces cerevisiae* strains used in this study.

Strain	Genotype	Source
CEN.PK 113-5D	*MATa MAL2-8^C SUC2 ura3-53*	P. Kötter
IMC082	CEN.PK 113-5D pUDC136	This study
IMC083	CEN.PK 113-5D pUDC137	This study
IMC084	CEN.PK 113-5D pUDC138	This study
IMC085	CEN.PK 113-5D pUDC139	This study
IMC086	CEN.PK 113-5D pAG416GPD	This study

Table 3.4: Plasmids used in this study.

Strain	Description	Source
pAG416GPD-*ccdB*	Centromeric plasmid, Amp, *URA3*, P_{TDH3}-*ccdb*-T_{CYC1}, Addgene plasmid 14148	[31]
pUDC136	Centromeric plasmid, Amp, *URA3*, P_{TDH3}-*atPAL1*-HA-T_{CYC1}	This study
pUDC137	Centromeric plasmid, Amp, *URA3*, P_{TDH3}-*jcPAL1*-HA-T_{CYC1}	This study
pUDC138	Centromeric plasmid, Amp, *URA3*, P_{TDH3}-*jcPAL1*-T_{CYC1}	This study
pUDC139	Centromeric plasmid, Amp, *URA3*, P_{TDH3}-*dcPAL1*-HA-T_{CYC1}	This study
pAG416GPD	Centromeric plasmid, Amp, *URA3*, P_{TDH3}-T_{CYC1}	This study

expression, its predictor was trained on the genes' total ribosome density without normalising by the mRNA levels. Variability of cellular mRNA levels across genes could mask the differences between their translation efficiencies, which may explain the substantial agreement between the *jcPAL1* and *dcPAL1* sequences. It would be interesting to also consider a version of DECODON that is trained to predict mRNA-normalised ribosome density, and thus separates the changes in TE between genes from their potential effect on mRNA levels.

Overall, experimental validation of DECODON showed that the method is suitable for re-designing gene sequences for higher expression, thus confirming the methods potential. However, it also highlighted several limitations of the method, which we addressed in the approach developed for modelling protein synthesis in Chapter 4.

MATERIALS AND METHODS

STRAINS AND MAINTENANCE

All strains used in this study were derived from the *S. cerevisiae* CEN.PK strain family [28, 29]. Yeast cultures were grown at 30°C in 500 ml shake flasks containing 100 ml synthetic medium (SM) [30] with 20 $g \cdot L^{-1}$ glucose and growth factors to supplement auxotrophic requirements of the strains. After overnight growth, glycerol was added to achieve final concentration of 20%, and 1 ml aliquots were stored at −80°C.

STRAIN AND PLASMID CONSTRUCTION

Plasmids were constructed using the standard restriction-ligation cloning with the *SpeI* and *XhoI* restriction sites, and the pAG416GPD-*ccdB* as a backbone vector. Genes *jcPAL1*

Table 3.5: Oligonucleotide primers used in this study.

Gene	Name	Sequence
atPAL1	FK1, Fw	GCGACTAGTATGGAGATTAACGGGGC
jcPAL1	AG2, Fw	GCGACTAGTATGGAAATCAACGGTGCTC
dcPAL1	AG6, Fw	GCGACTAGTATGGAAATTAACGGTGCTCAC
jcPAL1 no tag	AG3, Rev	GCGCTCGAGTTAACAGATTGGGATTGGAGC
atPAL1	AG8, Rev	GCGCTCGAGTTAAGCGTAATCTGGAACGTCATATGGATAACATATTGGA ATGGGAGCTCC
PAL1	AG10, Rev	GCGCTCGAGTTAAGCGTAATCTGGAACGTCATATGGATAACAGATTGGG ATTGGAGC

and *dcPAL1* were respectively optimised using JCat and our method, and commercially synthesised; *atPAL1* was obtained from [6]. All genes were amplified using high fidelity PCR and primers from Table 3.5; the *SpeI* and *XhoI* restriction sites and, when appropriate, the HA-tag were added during this step.

Ligation products were transformed into *E. coli* cells using electroporation, and plated on LB-Amp plates. Isolated single colonies were transferred into liquid LB-Amp medium, and their structures were verified using restriction analysis. Yeast transformations were performed through heat shock using cells grown to OD = 0.6 after re-inoculating overnight cultures. Transformed cells were plated on SM-URA plates and stocks were prepared from single colonies verified using PCR.

MEDIA AND CULTIVATION

Yeast shake flask cultures were grown at 30°C in 500 ml flasks containing 100 ml synthetic medium with 20 $g \cdot l^{-1}$ glucose and growth factors to supplement auxotrophic requirements of the strains. Optical density at 660 nm was measured with a Libra S11 spectrophotometer (Biochrom, Cambridge, UK).

PHENYLALANINE AMMONIA LYASE (PAL) ACTIVITY

Enzyme extraction was performed according to [32]. Briefly, 25 ml of shake flask cultures were sampled at OD = 5, and their cell extract was prepared by sonication using a protocol optimised for *Saccharomyces cerevisae*.

Cell extracts were used to measure *PAL1* activity according to [23]. Specifically, 100 μl of the extract were suspended in 850 Triethanolamine-HCl (pH 8.5, at 1 M) buffer, and the enzymatic reaction was started by the addition of 50 μl of L-phenylalanine (at 0.01 M) to start the reaction. Enzymatic activity at 30°C was measured as *trans*-cinnamate absorbance at 290 nm (molar attenuation coefficient ϵ = 2.1 $mM^{-1}cm^{-1}$). To obtain specific activity, the total protein concentration was determined using the Lowry assay [33]. All measurements were performed in duplicate.

REFERENCES

[1] A. A. Gritsenko, M. J. Reinders, and D. de Ridder, *Using predictive models to engineer biology: a case study in codon optimization*, in *Pattern Recognition in Bioinformatics* (Springer, Nice, France, 2013) pp. 159–171.

[2] B. Mohammadi and O. Pironneau, *Shape optimization in fluid mechanics*, Annual Review Fluid Mechanics **36**, 255 (2004).

[3] E. Angov, *Codon usage: nature's roadmap to expression and folding of proteins*, Biotechnology Journal **6**, 650

(2011).

[4] G. Cannarozzi and A. Schneider, *Codon evolution: mechanisms and models* (OUP Oxford, 2012).

[5] A. Grote, K. Hiller, M. Scheer, R. Münch, B. Nörtemann, D. C. Hempel, and D. Jahn, *JCat: a novel tool to adapt codon usage of a target gene to its potential expression host,* Nucleic Acids Research **33**, W526 (2005).

[6] F. Koopman, J. Beekwilder, B. Crimi, A. van Houwelingen, R. D. Hall, D. Bosch, A. van Maris, J. T. Pronk, and J.-M. Daran, *De novo production of the flavonoid naringenin in engineered Saccharomyces cerevisiae,* Microbial Cell Factories **11**, 155 (2012).

[7] G. Cannarozzi, N. N. Schraudolph, M. Faty, P. von Rohr, M. T. Friberg, A. C. Roth, P. Gonnet, G. Gonnet, and Y. Barral, *A role for codon order in translation dynamics,* Cell **141**, 355 (2010).

[8] T. Tuller, I. Veksler-Lublinsky, N. Gazit, M. Kupiec, E. Ruppin, and M. Ziv-Ukelson, *Composite effects of gene determinants on the translation speed and density of ribosomes,* Genome Biology **12**, R110 (2011).

[9] B. Maertens, A. Spriestersbach, U. von Groll, U. Roth, J. Kubicek, M. Gerrits, M. Graf, M. Liss, D. Daubert, R. Wagner, *et al., Gene optimization mechanisms: A multi-gene study reveals a high success rate of full-length human proteins expressed in Escherichia coli,* Protein Science **19**, 1312 (2010).

[10] H. Drucker, C. Burges, L. Kaufman, A. Smola, and V. Vapnik, *Support vector regression machines,* Advances in Neural Information Processing Systems , 155 (1997).

[11] N. T. Ingolia, S. Ghaemmaghami, J. R. Newman, and J. S. Weissman, *Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling,* Science **324**, 218 (2009).

[12] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder, *The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing,* Science **320**, 1344 (2008).

[13] M. Yassour, T. Kaplan, H. Fraser, J. Levin, J. Pfiffner, X. Adiconis, G. Schroth, S. Luo, I. Khrebtukova, A. Gnirke, C. Nusbaum, D.-A. Thompson, N. Friedman, and A. Regev, *Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing,* Proceedings of the National Academy of Sciences **106**, 3264 (2009).

[14] W. Qian, J. Yang, N. Pearson, C. Maclean, and J. Zhang, *Balanced codon usage optimizes eukaryotic translational efficiency,* PLoS Genetics **8**, e1002603 (2012).

[15] J. Coleman, D. Papamichail, S. Skiena, B. Futcher, E. Wimmer, and S. Mueller, *Virus Attenuation by Genome-Scale Changes in Codon Pair Bias,* Science **320**, 1784 (2008).

[16] I. Hofacker, W. Fontana, P. Stadler, L. Bonhoeffer, M. Tacker, and P. Schuster, *Fast folding and comparison of rna secondary structures,* Monatshefte für Chemie/Chemical Monthly **125**, 167 (1994).

[17] C.-C. Chang and C.-J. Lin, *LIBSVM: A library for support vector machines,* ACM Transactions on Intelligent Systems and Technology **2**, 27:1 (2011).

[18] L. Wessels, M. Reinders, A. Hart, C. Veenman, H. Dai, Y. He, and L. Van't Veer, *A protocol for building and evaluating predictors of disease state based on microarray data,* Bioinformatics **21**, 3755 (2005).

[19] R. Kohavi and G. H. John, *Wrappers for feature subset selection,* Artificial Intelligence **97**, 273 (1997).

[20] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, *A fast and elitist multiobjective genetic algorithm: NSGA-II,* Evolutionary Computation, IEEE Transactions on **6**, 182 (2002).

[21] K. Fredrick and M. Ibba, *How the sequence of a gene can tune its translation,* Cell **141**, 227 (2010).

[22] P. Lu, C. Vogel, R. Wang, X. Yao, and E. Marcotte, *Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation,* Nature Biotechnology **25**, 117 (2006).

[23] Z. Xue, M. McCluskey, K. Cantera, F. S. Sariaslani, and L. Huang, *Identification, characterization and functional expression of a tyrosine ammonia-lyase and its mutants from the photosynthetic bacterium Rhodobacter sphaeroides,* Journal of Industrial Microbiology & Biotechnology **34**, 599 (2007).

[24] S. Edri and T. Tuller, *Quantifying the effect of ribosomal density on mRNA stability,* PloS One **9**, e102308 (2014).

[25] M. V. Gerashchenko and V. N. Gladyshev, *Translation inhibitors cause abnormalities in ribosome profiling experiments,* Nucleic Acids Research **42**, e134 (2014).

[26] J. A. Hussmann, S. Patchett, A. Johnson, S. Sawyer, and W. H. Press, *Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast,* PLoS Genetics **11**, e1005732 (2015).

[27] D. E. Weinberg, P. Shah, S. W. Eichhorn, J. A. Hussmann, J. B. Plotkin, and D. P. Bartel, *Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation,* Cell Reports **14**, 1787 (2016).

[28] J. Van Dijken, J. Bauer, L. Brambilla, P. Duboc, J. Francois, C. Gancedo, M. Giuseppin, J. Heijnen, M. Hoare, H. Lange, *et al., An interlaboratory comparison of physiological and genetic properties of four Saccharomyces cerevisiae strains,* Enzyme and Microbial Technology **26**, 706 (2000).

[29] J. F. Nijkamp, M. van den Broek, E. Datema, S. de Kok, L. Bosman, M. A. Luttik, P. Daran-Lapujade, W. Vongsangnak, J. Nielsen, W. H. Heijne, *et al., De novo sequencing, assembly and analysis of the genome of the laboratory strain Saccharomyces cerevisiae CEN.PK113-7D, a model for modern industrial biotechnology,* Microbial Cell Factories **11**, 1 (2012).

[30] C. Verduyn, E. Postma, W. A. Scheffers, and J. P. van Dijken, *Physiology of Saccharomyces cerevisiae in anaerobic glucose-limited chemostat cultures,* Microbiology **136**, 395 (1990).

[31] S. Alberti, A. D. Gitler, and S. Lindquist, *A suite of Gateway® cloning vectors for high-throughput genetic analysis in Saccharomyces cerevisiae,* Yeast **24**, 913 (2007).

[32] E. Postma, C. Verduyn, W. A. Scheffers, and J. P. Van Dijken, *Enzymic analysis of the Crabtree effect in glucose-limited chemostat cultures of Saccharomyces cerevisiae,* Applied and Environmental Microbiology **55**, 468 (1989).

[33] O. H. Lowry, N. J. Rosebrough, A. L. Farr, R. J. Randall, *et al., Protein measurement with the Folin phenol reagent,* Journal of Biological Chemistry **193**, 265 (1951).

4

USING RIBOSOME PROFILING DATA TO MODEL PROTEIN SYNTHESIS

Alexey A. GRITSENKO, Marc HULSMAN, Marcel J.T. REINDERS and Dick DE RIDDER

This chapter has been published in *PLoS Computational Biology* **11**, e1004336 (2015) [1].

ABSTRACT

Translation of RNA to protein is a core process for any living organism. While for some steps of this process the effect on protein production is understood, a holistic understanding of translation still remains elusive. *In silico* modelling is a promising approach for elucidating the process of protein synthesis. Although a number of computational models of the process have been proposed, their application is limited by the assumptions they make. Ribosome profiling (RP), a relatively new sequencing-based technique capable of recording snapshots of the locations of actively translating ribosomes, is a promising source of information for deriving unbiased data-driven translation models. However, quantitative analysis of RP data is challenging due to high measurement variance and the inability to discriminate between the number of ribosomes measured on a gene and their speed of translation.

We propose a solution in the form of a novel multi-scale interpretation of RP data that allows for deriving models with translation dynamics extracted from the snapshots. We demonstrate the usefulness of this approach by simultaneously determining for the first time per-codon translation elongation and per-gene translation initiation rates of *Saccharomyces cerevisiae* from RP data for two versions of the Totally Asymmetric Exclusion Process (TASEP) model of translation. We do this in an unbiased fashion, by fitting the models using only RP data with a novel optimisation scheme based on Monte Carlo simulation to keep the problem tractable. The fitted models match the data significantly better than existing models and their predictions show better agreement with several independent protein abundance datasets than existing models. Results additionally indicate that the tRNA pool adaptation hypothesis is incomplete, with evidence suggesting that tRNA post-transcriptional modifications and codon context may play a role in determining codon elongation rates.

4.1. INTRODUCTION

The process of protein synthesis is central to all living organisms. It has been actively researched for over five decades, and by now the individual steps of this process are known in great detail at the molecular and mechanistic levels [2]. Gene adaptation to the tRNA pool, mRNA secondary structure strength, codon order and local amino acid charge were independently implicated in shaping rates of protein production [3–5]. However, many disciplines would benefit from a holistic view of how these factors collectively influence translation. In particular, in biotechnology this knowledge would allow for tuning protein expression as desired with ramifications for cost-effective production of medicines and biofuels using microbes [6]. However, owing to the biological complexity of the process and the difficulty of measuring kinetic rates of the individual steps of protein synthesis, the development of computational models that would enable such applications lagged behind.

Only recently, the accumulated knowledge was integrated into several state-of-the-art models of increasing complexity. Zhang and Ignatova [7] proposed a "static" model for predicting the local speed of translation within a gene from codon-specific elongation rates derived from tRNA concentrations; their approach was extended by Reuveni *et al.* [8], who suggested using a "dynamic" model in which ribosomes initiate translation at the first codon and block each other while moving towards the end of the mRNA transcript. Siwiak and Zielenkiewicz [9] and Shah *et al.* [10] independently proposed static and dynamic full-cell models that additionally integrated the intracellular concentrations of ribosomes, mRNA and tRNA molecules, and their diffusion inside the cell in a single model. While predictions made by these models are usually in accordance with the current understanding of translation, most of their core assumptions (e.g., codon translation rates) have not been subjected to comparison against measured data.

Ribosome profiling (RP) [11, 12], a relatively new technique based on high-throughput sequencing of ribosome-protected RNA fragments (footprints), is nowadays often employed for studying translation [13–16]. It provides noisy snapshots of the locations of actively translating ribosomes attached to mRNA transcripts, thereby convolving the number of ribosomes and their speed of translation (a few stalled ribosomes can generate similar sets of footprints as many ribosomes involved in rapid translation). While in principle these data allow for simultaneously reasoning about ribosome counts and their local speed, such analysis is hampered by the limited understanding of the error model and biases of RP data [17]. To date RP measurements have been analysed either at the level of full genes [9, 10] or at single codon resolution [5, 18]. While only the latter allows for analysing the dynamics of translation, it is not clear whether codon-resolution measurements are sufficiently reliable for such quantitative analysis (see Suppl. Text, page 100). To overcome the measurement reliability issue several studies [19–21] performed "meta-codon" analysis by pooling observations from different occurrences of a particular codon together to produce an estimate of the codon elongation time. It is unclear, however, to what extent such estimates are affected by ribosomal interference.

We propose a set of methods for deriving full translation kinetics of an organism from RP data (see Fig.. 4.1). Our approach is conceptually similar to Ciandrini *et al.*

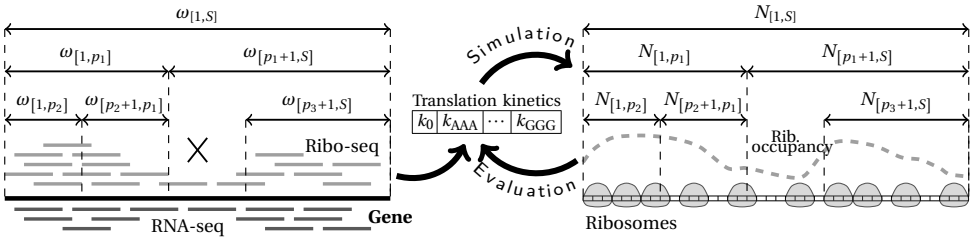

Figure 4.1: Schematic overview of the proposed approach for inferring translation kinetics from RP data. To obtain a segment tree representation of the RP data (**left**) mapped ribo-seq (light grey) and RNA-seq (dark grey) reads are assigned to nested segments of decreasing lengths (starting from segments $[1, S]$ equivalent to the full-length CDSes) while there is sufficient data. Ribosome densities ω for each segment are computed for the available replicates and are used to parameterize the log-normal distributions describing measurement error of these segments. To determine per-gene translation initiation rates k_0 and per-codon elongation rates $k_{AAA}, \ldots, k_{GGG}$ many candidate sets of translation rates are tested. For every candidate set the TASEP model of translation is simulated with the proposed rates for all genes in the *model simulation step* (**right**). Ribosome occupancy, i.e., the relative amount of time ribosomes spend at a particular location on the mRNA, obtained from the simulation (dashed grey) is then aggregated per segment to compute the average occupancies N, which are compared the log-normal distributions of the corresponding segments from the segment tree representation in the *model evaluation step*. Evaluation results are used by a genetic algorithm to propose new candidate sets of rates and repeat the simulation-evaluation cycle until the search for translation rates converges. To simplify notation, the gene index g is dropped for all gene-specific variables in the figure.

[22], who inferred translation initiation rates of yeast genes from polysome profiling data, except that we use RP for deriving these rates and additionally determine the translation elongation rates. The method is based on a novel "segment tree" multi-scale interpretation of the RP data that captures ribosome translation dynamics along mRNAs without sacrificing reliability due to measurement noise. We use this interpretation to simultaneously extract, for the first time, per-gene translation initiation rates and per-codon translation elongation rates for the bakers yeast *Saccharomyces cerevisiae* by fitting two version of the TASEP (Totally Asymmetric Exclusion Process), a simple dynamic model of translation [23], on the segment tree estimates. To make fitting tractable, we devised a highly efficient initiation rate approximation scheme and combined it with a novel Monte Carlo simulation strategy inside an evolutionary optimisation algorithm.

Fitted TASEP models match the RP data significantly better than the state-of-the-art models, and their predicted protein production rates are confirmed by several independent protein abundance (PA) datasets. In particular our models show significantly better agreement with PA than existing models when the measurements are corrected for mRNA levels, i.e., when only the effect of translation on protein levels is considered. Interestingly, the fitted codon elongation rates deviate significantly from the tRNA pool adaptation hypothesis.

4.2. MATERIALS AND METHODS

RIBOSOME PROFILING DATA

RP data for yeast *Saccharomyces cerevisiae* strain S288C [24] containing ribosome footprint reads (ribo-seq) and fragmented mRNA reads (RNA-seq) measured in duplicate were obtained from the NCBI Short Read Archive (accession SRP028552). Reads were trimmed and mapped to the latest *S. cerevisiae* strain S288C reference genome taken from the Saccharomyces Genome Database (SGD, Cherry *et al.* [25]) in two stages, and assigned to gene coding sequences (CDSes) obtained from SGD. Aligned ribosome footprint and mRNA reads were assigned to single positions within the CDSes based on respectively their inferred A-sites or the centre position of the read (see Suppl. Text, page 89 for details).

MEASUREMENT RESOLUTION

To obtain a high-resolution map of mRNA and ribosome density without sacrificing measurement accuracy, for each gene we construct a segment tree of density measurements from nested parts of the CDSes (Fig. 4.1, left). By pooling reads from all segment positions, average densities per segment can be calculated more reliably than would be possible at single codon resolution (see also Suppl. Text, page 89), while recording these densities for nested segments of decreasing lengths allows for indirectly capturing the change in density along a transcript.

Starting from an initial segment $[l,r]$ equivalent to the complete CDS we count the number of ribo-seq reads $R_{[l,r]}$ and RNA-seq reads $M_{[l,r]}$ assigned to this segment. These counts are normalised by the total number of ribo- and RNA-seq reads aligned to all CDSes (N_R and N_M respectively) and the segment length $L_{[l,r]} = r - l + 1$ to obtain ribosome and mRNA densities $d^{\text{Ribo}}_{[l,r]} = \frac{R_{[l,r]}}{L_{[l,r]}N_R}$ and $d^{\text{mRNA}}_{[l,r]} = \frac{M_{[l,r]}}{L_{[l,r]}N_M}$ for the current segment. To obtain the sought *per transcript* ribosome density (later referred to as density ratio) the ratio of the two measurements $\omega_{[l,r]} = \frac{d^{\text{Ribo}}_{[l,r]}}{d^{\text{mRNA}}_{[l,r]}}$ is calculated. The average segment ribosome density given by this ratio is normalised for transcript abundance and allows for directly comparing segments from different genes to each other. A cut point p is then chosen and the process is repeated recursively for segments $[l,p]$ and $[p+1,r]$ (see Fig. 4.1, left). The aim behind calculating $d^{\text{mRNA}}_{[l,r]}$ for each segment independently instead of estimating a single gene-specific value is to remove any local sequencing bias (presumed to be identical between RNA- and ribo-seq since very similar protocols are used for library preparation [24]) from the ratio estimates. Density measurements are computed for each replicate individually, but the same segment cut points are used in order to merge replicates later. Cut points are chosen such that the combined number of RNA- and ribo-seq reads across replicates is divided equally between the left and the right segments (see Suppl. Text, page 95 for details).

The recursive tree construction continues while there are sufficient reads for making reliable density estimates (at least 128 reads in the two replicates summed together for RNA-seq and ribo-seq, separately; see Suppl. Text, page 95 for details on choosing these thresholds) and segment length is large enough, $L_{[l,r]} \geq 20$ codons. The segment length cutoff aims at keeping the segments long enough to average out any measurement

error due to incorrect read assignment or sequence bias. Prior to interpreting the measurements, we additionally remove a systematic density-dependent bias present in the density and ratio measurements using the available replicate information (see Suppl. Text, page 97).

This procedure was used to construct segment trees for 4,892 genes with a total of 60,466 nested density estimates left after removing genes classified as dubious or located on the mitochondrial chromosome.

STATISTICAL TREATMENT OF THE MEASUREMENTS

In order to accurately capture variance of RP data, we assume that the measured segment densities follow a log-normal distribution around the density values. A similar assumption is often made for transcriptome measurements and is justified by the observation that inter-replicate errors (i.r.e.), i.e., the ratios of replicated mRNA and ribosome density measurements, follow a log-normal distribution (Suppl. Fig. 4.10 and Ingolia *et al.* [11]). It then holds that density ratios $\omega_{[l_j,r_j]}$ ($j \in J^g$, where J^g is the set of all segments of gene g) from different replicates also follow a log-normal distribution $\ln\mathcal{N}(\mu_j,\sigma_j)$ as ratios of log-normally distributed random variables - the mRNA and ribosome segment densities. Here μ_j and σ_j are used as shorthands for $\mu_{[l_j,r_j]}$ and $\sigma_{[l_j,r_j]}$ respectively.

To determine the parameters of this distribution we estimate μ_j for the j-th segment from the available replicated measurements as the log of their geometrical mean. Ideally, a separate shape parameter σ_j should also be estimated per segment, but, given the number of replicates, doing so would not yield reliable estimates. Instead it was chosen to group segments from all genes based on their length, and estimate shape parameters σ_k^{group} for group k from the i.r.e. of measurements from that group (see Suppl. Text, page 98 and Suppl. Fig. 4.12).

The proposed measurement distribution $\ln\mathcal{N}(\mu_j,\sigma_{k_j}^{\text{group}})$, where k_j denotes the length group of the j-th segment, is used throughout this paper as an error model for fitting TASEP models of translation on RP data and for comparing different models with the data.

DATA INTERPRETATION AND MODEL EVALUATION

Computational models of translation typically provide the ability to extract steady-state codon occupancy probabilities obtained from model simulations, i.e., estimates of the chance that a particular position of an mRNA is occupied by an actively translating ribosome. Similar to the ribosome profiling measurements these occupancy profiles are determined by the local speed of translation and the number of ribosomes translating an mRNA. This allows for evaluating how well a given model matches the RP data by comparing the average segment occupancies and the segment tree ratio estimates (see Fig. 4.1, right).

Quantitative measurements obtained via high-throughput sequencing such as the mRNA and ribosome densities (and hence their ratios) are measured in arbitrary units. Without explicit assumptions on the physiological characteristics of the analysed organism, such as the full size of its transcriptome [9] or the number of ribosomes

per cell [10], and on the efficiency of individual experimental steps, it is impossible to estimate sequencing depth of the RP measurements (i.e. the average number of reads per ribosome or the average number of reads per kilobase of transcript) and therefore impossible to express the measured values in physiologically meaningful units (e.g. number of ribosomes per transcript). Additionally, this unit mismatch complicates the comparison of modelled ribosome occupancies to the measured densities. To derive a model evaluation criterion, we first assume that an unknown scaling factor C that transforms model output into measurement data units is given, and propose a method for calculating it later.

Let n_i^g be the model-predicted ribosome occupancy at position i of gene g and $T^g = \left\{\left(\mu_j^g, \sigma_j^g\right) \middle| j \in J^g\right\}$ be the set of ratio distribution parameters for segments $\left[l_j^g, r_j^g\right]$. Here the upper index g denotes the gene, and for a more succinct notation we use the lower index j in place of $\left[l_j^g, r_j^g\right]$. For segment j on gene g the probability of the predicted occupancies given the segment ratio estimates can be expressed as

$$p\left(C, N_j^g \middle| \mu_j^g, \sigma_j^g\right) \propto f_C(N_j^g; \mu_j^g, \sigma_j^g), \tag{4.1}$$

where $N_j^g \equiv \sum_{i=l_j^g}^{r_j^g} n_i^g \Big/ \left(r_j^g - l_j^g + 1\right)$ is the predicted average occupancy on segment j of gene g, and $f_C(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x + \ln C - \mu)^2}{2\sigma^2}}$ is the log-normal probability density function describing the density ratio measurement error scaled by factor $\frac{1}{C}$. This formulation is used for comparing the predicted occupancies to the estimated values in a probabilistic fashion. Assuming independence between ratio estimates of the same gene and between genes, the probability of observing all estimates, denoted by n, can be expressed as

$$p(C, n|T) \propto \prod_g \prod_{j \in J^g} f_C(N_j^g; \mu_j^g, \sigma_j^g), \tag{4.2}$$

In practice these calculations are more easily performed in log space and the constant factors are dropped:

$$\begin{aligned} \psi(C, n|T) &= \sum_g \sum_{j \in J^g} \ln f_C(N_j^g; \mu_j^g, \sigma_j^g) \sim \\ &\sim \sum_g \sum_{j \in J^g} \left[-\frac{1}{2\left(\sigma_j^g\right)^2} \left(\ln N_j^g - \mu_j^g + \ln C\right)^2 - \ln N_j^g \right] \end{aligned} \tag{4.3}$$

We use $\psi(C, n|T)$ as the objective function for quantifying how well model-predicted ribosome occupancies match measured data.

To choose the scaling factor C, we note that it is the only free parameter of $\psi(C, n|T)$ if model output n and segment tree estimates T are given. In that case, the value of C

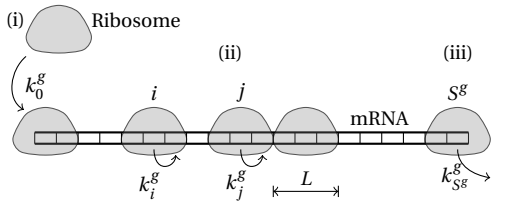

Figure 4.2: In TASEP mRNAs are modelled as one-dimensional lattices of S^g sites (codons) and ribosomes - as particles occupying L sites ($L = 3$ in the figure). During translation (**i**) ribosomes attach to the first codon with rate k_0^g if the beginning of the mRNA is not occupied by other ribosomes (initiation); (**ii**) ribosomes move from position i to $i+1$ with a site-specific rate k_i^g if not blocked by another ribosome (elongation); and, finally, (**iii**) after reaching the last codon, they detach with rate k_{Sg}^g (termination).

maximising ψ can be determined analytically:

$$\ln C = \left(\sum_{g,j \in Jg} \frac{1}{\left(\sigma_j^g\right)^2} \left(\mu_j^g - \ln N_j^g\right) \right) \Bigg/ \left(\sum_{g,j \in Jg} \frac{1}{\left(\sigma_j^g\right)^2} \right) \tag{4.4}$$

Throughout this paper, different models are evaluated at a scaling factor C maximising their fit to the data (i.e., maximising ψ). While the unknown true scaling factor is determined by the physiological properties of the cell, the efficiency of the experimental protocols and characteristics of the high-throughput sequencing measurements (see section "Initiation rate approximation" and Suppl. Text, page 101), evaluating models at the best possible scale allows for a more fair evaluation as it does not penalise models in cases when the model and the true scales mismatch.

THE TASEP MODEL OF TRANSLATION

TASEP (Totally Asymmetric Exclusion Process) models mRNAs g as one-dimensional lattices of length S^g and ribosomes as abstract "particles" occupying L sites corresponding to codons (Fig. 4.2). These particles hop on (translation initiation) and off (translation termination) the lattice at the first and last sites with rates k_0^g and k_{Sg}^g respectively. They only move towards the end of the lattice (hence the totally asymmetric) by hopping one site at a time with site-specific elongation rate k_i^g. Ribosomes interact with each other by "excluding" a volume of L sites that they cover on the lattice, meaning that a ribosome cannot continue to the next codon if it is already covered by another ribosome. The exact location of the active site among the L covered codons does not change the rules governing ribosome motion [23], but the choice of L may influence simulation dynamics in cases of high ribosome queueing. Typically, values $9 \leq L \leq 11$ are used [9, 10, 17, 22]; $L = 10$ was chosen for our simulations as it best matches the RP footprint size distribution [11].

TASEP captures the high-level physical interaction between ribosomes and transcripts by describing the ribosomes as travelling on the mRNAs. While in practice a number of varying translation scenarios are possible (e.g., RER-bound translation with

ribosomes glued to the endoplasmic reticulum and moving very slowly while the mRNA is instead pulled though the ribosomes [26]), the rich set of behaviours attainable by TASEP makes it a suitable framework for modelling translation. It requires specification of a large number of parameters, namely the gene- and site-specific elongation rates k_i^g (with the stop codon elongation rate functioning as the termination rate) and the gene-specific initiation rates k_0^g. To reduce the number of parameters we assume that the site-specific elongation rates are codon-specific and do not differ between genes. This commonly made assumption [8, 17, 22, 27] is necessary for determining model parameters from RP data as it makes the model fitting problem tractable. Depending on the experiment, either elongation rates consistent with the tRNA pool adaptation hypothesis were fixed to allow fitting the initiation rates only, or all model parameters were fit on the available data.

MONTE CARLO SIMULATIONS

Evaluation and fitting of the TASEP model requires an efficient way of obtaining steady-state ribosome occupancies. TASEP models allow limited analytical tractability and, to our knowledge, no analytical results for the steady-state codon occupancy probabilities are available for the general case. Additionally, existing TASEP mean-field approaches poorly approximate codon occupancies [28], a quantity of particular importance to this study, leaving stochastic simulations as the only suitable approach.

TASEP steady-state codon occupancies were obtained by simulating the model using a Monte Carlo algorithm, i.e., by randomly selecting an event (translation initiation, elongation or termination) in every simulation step and, if no other ribosomes interfere with the event, executing it with a probability proportional to its rate. To speed up simulation we developed a continuous time simulation method similar to the Gillespie algorithm [29], but based on the use of the Erlang distribution to only sample times between *state-changing events*, i.e., events that change the configuration of ribosomes attached to an mRNA.

Formally, the times between consecutive initiation or elongation events at position i are assumed to be exponentially distributed with rates k_0^g and k_i^g respectively (i.e., the corresponding model rate parameters, Fig. 4.2). Let $o_i, i = 1, \ldots, S^g$ be the current state of the simulated molecule:

$$o_i = \begin{cases} 1, & \text{codon } i \text{ is occupied by a ribosome (is at its A-site)} \\ 0, & \text{otherwise} \end{cases}. \tag{4.5}$$

Then the time between *any* two consecutive events is also exponentially distributed with rate $k = k_0^g + \sum_{i=1}^{S^g} o_i k_i^g$ as the minimum of independent exponentially distributed random variables. Once an event occurred, the probability that it was event j is given by $p_j = o_j k_j^g \big/ k$ (it is assumed that ribosomes are always available to initiation translation, i.e., $o_0 = 1$). Some of the events cannot be executed due to ribosomes blocking each other and do not lead to a state change. If k_+ is the sum of rates of events leading to a state change, then the number of events between consecutive state changes, denoted as e, follows a geometric distribution with parameter $p_+ = k_+ \big/ k$ and the time Δt between state changing events follows the Erlang distribution with shape e and rate k as the

sum of iid exponential random variables. The simulation proceeds by repeated random sampling of the number of events, the time between events and the event type s from the appropriate probability distributions; and updating ribosome locations in accordance to the sampled event:

$$s \sim \text{Categorial}\left(p_0, p_1, \ldots, p_{Sg}\right), \quad e \sim \text{Geometric}\left(p_+\right), \quad \Delta t \sim \text{Erlang}\left(e, k\right). \tag{4.6}$$

Simulating only state-changing events allows the simulation to progress faster, especially in cases of high ribosome queueing. The total time T_i^g spent by ribosomes at position i and the total simulation time T^g are recorded to estimate the per-transcript ribosome occupancy at this position as $n_i^g = T_i^g / T^g$, which is then used for comparing the model to RP data. Similarly the total number of translation terminations F^g is used to estimate the protein production rate $J^g = F^g / T^g$.

To reach steady-state distribution of ribosomes on mRNA irrespective of the CDS length, each mRNA was simulated until 1000 translation termination events occurred. After that the model was further simulated for up to 10^7 additional steps or until the average ribosome occupancy in the segments of interest was estimated with high precision (absolute error $\epsilon < 10^{-3}$). The latter stopping criterion is based on the observation that average ribosome occupancy over a fixed segment of the mRNA can be reliably estimated before per-position occupancies can. Segment densities were first estimated after 5×10^5 simulation steps and then every 10^6 steps. Simulation was stopped if the absolute error between consecutive estimates was smaller than ϵ.

INITIATION RATE APPROXIMATION

In addition to the elongation rates, large TASEP models require specification of hundreds gene translation initiation rates prior to simulation. Direct measurements of the initiation rates rates are unavailable and instead their values are often inferred from other sources such as ribosome profiling [9, 10] or polysome size measurements [22] data. Initiation rates estimated in such a way depend on the rates of translation elongation used in the analysis, and hence need to be optimised together with the elongation rates of the TASEP model. This leads to an explosion of the number of parameters that need to be determined, stressing the need for highly efficient initiation rate approximation strategies if the initiation and elongation rates are to be determined from the RP data simultaneously.

The problem of determining initiation rates was previously tackled by approximations neglecting ribosome queueing [9, 10], and by near-exhaustive computational search [22]. We propose a method that is a compromise between the two approaches - it allows approximating gene initiation rates for the TASEP model from RP data at a fraction of the computational cost of an exhaustive search. Briefly, we add an additional parameter $\tilde{C}$, the "proposed" scaling factor, to the list of model parameters that need to be estimated. This parameter is identical to the scaling factor C from eq. (4.4), but is used within the model to obtain biologically meaningful initiation rates. We calculated the value of $\tilde{C}$ from the number actively translating ribosomes [30] and the number of mRNA molecules [31] per cell using a procedure proposed by Siwiak and Zielenkiewicz [9]. Given some estimate of the elongation rates and $\tilde{C}$ we then find optimal initiation rates using a novel numerical approximation of ribosome density for

TASEP models that is based on the observations of Cinandrini *et al.* [22]. This approach allows us to decouple initiation rates from elongation rates and greatly reduces the number of model parameters that need to be fitted explicitly (next section). We used this method to efficiently (re-)approximate initiation rates of genes for each new set of elongation rates k_i^g. A full description of the approach is available in the Suppl. Text, page 104.

MODEL FITTING

When fitting the TASEP models, translation rates that maximise $\psi(C, n|T)$ are sought. Lacking a closed-form solution, we employed the Covariance Matrix Adaptation Evolutionary Strategy (CMA-ES [32]) to find these rates.

We considered two different TASEP models: TASEP$^{\text{init}}$ and TASEP$^{\text{elong}}$. In TASEP$^{\text{init}}$ the elongation rates are fixed at values consistent with the tRNA pool adaptation hypothesis and initiation rates are approximated as described earlier. In the TASEP$^{\text{elong}}$ model none of the parameters are fixed: also the codon-specific elongation rates are optimised with the CMA.

Since TASEP simulation output is invariant to scaling of translation rates, many equally good solutions exist. To constrain the search the elongation rate of codon GAA was fixed at its initial tRNA pool adaptation hypothesis value. The codon was chosen as it is present in many genes and segments (Suppl. Fig. 4.14). Further details regarding the use of CMA can be found in the Suppl. Text, page 107.

Despite the efficient Monte Carlo simulation and translation rate search strategies, model fitting remains a very CPU-intensive task. To speed up computations in practice, the models were fitted using hundreds of CPUs in parallel as individual genes can be simulated independently.

Because TASEP simulations of different genes are independent of each other, it may be unclear how to interpret the fitted elongation and initiation rates, as they must depend on such global biophysical quantities as the number of tRNAs or ribosomes within the cell. Nevertheless, the final simulation results are compared to whole-genome RP measurements. We can therefore expect that if our TASEP simulations agree well with RP data, the fitted translation rates used within the simulations account for the necessary biophysical parameters. Thus they should be regarded as the *effective* initiation and elongation rates that account for the relevant biophysical characteristics of the cell and growth conditions. We note that translation rates determined in such a way are condition-specific, and would likely change if fitted on a dataset obtained under different growth conditions.

COMPARISON TO OTHER MODELS

To obtain a baseline for evaluating the performance of fitted TASEP models we also evaluated several existing state-of-the-art static and dynamic models of translation and compared them to each other based on their agreement with the RP data as given by eq. (4.3). SMoPT [10] and Zhang's model [7] were chosen for evaluation on the segment tree data as other state-of-the-art models, namely the Ribosome Flow Model [8] and the model from Siwiak and Zielenkiewicz [9], do not provide ribosome occupancy profiles compatible with the segment tree interpretation. The latter model was however

compared to the fitted TASEP models based on several independent PA datasets.

When comparing models' predictions using independent protein abundance datasets, the "initiation frequency" P, "total amount of protein molecules produced from transcripts of particular type" B and the "total time of translation of one protein molecule from a given transcript" T from Siwiak and Zielenkiewicz [9] were respectively treated as translation initiation rate, the product of J and mRNA levels, and the inverse of J; the average gene total elongation time from SMoPT [10] was treated as the inverse of J; $\mathscr{P}$ from Ciandrini *et al.* [22] was treated as J.

EXPERIMENTAL SETUP

Since the sets of genes included in SMoPT and the segment trees differ, to facilitate comparison, all models were evaluated on a set of 3,617 genes (49,894 segments) that were in common between all models after removing very long genes to speed up TASEP simulations (31 genes longer than 2,000 codons). This set of genes was used to fit TASEP models inside a 5-fold stratified cross-validation (CV) loop over genes, in which the CV folds were chosen to balance the number of genes and segments between folds. In every step of the CV 1 fold was used for fitting (training set) and 4 folds were used for model evaluation (test set). Smaller training sets were used to reduce model fitting time. To evaluate predictions of the proposed TASEP models we also fitted them on all segment tree estimates. And to further reduce fitting time on this large dataset, codon elongation rates of the $\text{TASEP}^{\text{elong}}$ model were set to the geometric mean of elongation rates from CV folds, and only the initiation rates were estimated from the data.

To simplify comparison of different models, we computed CV objectives for all evaluated models, including the models that did not require any parameter fitting (i.e., SMoPT and Zhang's model). While the static Zhang model does not explicitly model the translation initiation step, SMoPT and TASEP models require initiation rates to be defined for every gene in the test sets in order to calculate the CV objective. We used the original initiation rates inferred from the RP data [10, 11] for SMoPT, and approximated TASEP initiation rates using the test set segment tree measurements.

THE tRNA POOL ADAPTATION HYPOTHESIS

Some of the experiments required the translation elongation rates to be defined. For those experiments we used translation elongation rates $k_{\text{AAA}}, \ldots, k_{\text{GGG}}$ consistent with the tRNA pool adaptation hypothesis, which could be seen as a statement that codons recognised by more abundant tRNAs are translated faster. The exact values for the elongation rates were defined based on the tRNA Adaptation Index (tAI [33]), which quantifies the decoding efficiency of a codon by simultaneously considering abundances of all tRNA species recognising it and the strength of wobble base pairing between the codon and the anticodons of the isoacceptor tRNAs. The elongation rates $k_{\text{AAA}}, \ldots, k_{\text{GGG}}$ were calculated as the inverse of the codon translation times taken from the Ribosome Flow Model [34]; and translation termination rates (i.e., $k_{\text{TAG}}, k_{\text{TAA}}, k_{\text{TGA}}$) were set to 1.

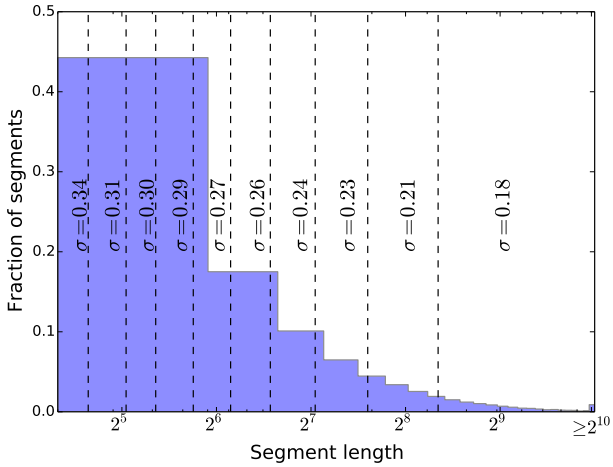

Figure 4.3: Segment length histogram overlaid with the shape parameters of the density ratio distributions for segment length groups (separated by dashed lines) shows that shorter segments tend to have more variable measurements. Segments were separated based on their length into 10 equal-content groups (group edges adjusted to allow for unique segment assignment), and the shape parameters σ were calculated from the inter-replicate errors of the measurements falling within each group (Suppl. Table 4.5).

COMPARISON TO TAI AND CAI

The tAI and CAI (Codon Adaptation Index [35]) are the most commonly used codon indices. They quantify respectively the extent to which a particular sequence consists of codons recognised by abundant tRNAs, and the extent to which a particular sequence consists of codons present in highly expressed (e.g., ribosomal and glycolytic) genes. These indices are often used as a proxy for translational efficiency of a gene and are employed to optimise its sequence for expression in a different host organism. Having determined elongation rates for the TASEP$^{\text{elong}}$ model, we sought to understand whether these rates suggest a different optimisation scheme than the one given by tAI or CAI.

For each codon the tAI (CAI) assigns a number - the absolute adaptiveness of that codon to the tRNA pool (codons used in highly expressed genes). To facilitate comparison between the different indices, following the definition of the CAI, we define the relative adaptiveness of a codon as its absolute adaptiveness normalised by the maximum adaptiveness among synonymous codons. We then use the relative adaptiveness for CAI, tAI and an index based on the TASEP$^{\text{elong}}$ elongation rates (described below), when comparing optimisation schemes.

We note that from the definitions of tAI [33] and elongation rates consistent with the tRNA pool hypothesis (previous section and [8]) it follows that the tAI absolute codon adaptiveness and the elongation rates are proportional to each other, and use this observation to define a codon index based on the fitted TASEP$^{\text{elong}}$ elongation rates. We define the relative adaptiveness of a codon according to TASEP$^{\text{elong}}$ as its elongation rate normalised as described above.

OTHER DATASETS

Protein abundance measurements were taken from Newman *et al.* [36] (YEPD and SD media) and Ghaemmaghami *et al.* [37]. 5′- and 3′ UTR lengths were determined based on Nagalakshmi *et al.* [38] and Yassour *et al.* [39] as the mean length obtained from the two studies.

4.3. RESULTS

SEGMENT TREES RELIABLY CAPTURE DENSITY CHANGES ALONG TRANSCRIPTS

Segment density ratios are estimates of the average number of ribosomes engaged in translation of a given segment (measured in arbitrary units), and are expected to become more reliable if the segment length is increased. Fig. 4.3 shows that estimates obtained for longer segments are indeed more reliable (smaller σ values) with the longest segments (rightmost group) being nearly as reliable as the full-CDS estimates from all genes (Suppl. Fig. 4.12). We note that although group widths increase almost exponentially, potentially collecting segments with different i.r.e. in the top group, the constructed groups map very well to individual levels of the segment trees because lengths of segments with each new level are halved on average. This mapping thus provides important additional information to the segment trees about the increasing reliability of measurements that are located higher within the tree.

In this way, segment trees establish a tradeoff between measurement reliability and measurement resolution by combining the use of trustworthy estimates high in the tree (corresponding to longer segments, describing high-level gene behaviour) with the use of many less reliable estimates located lower in the tree that describe the local density variation. As can be seen from the visualisation of the raw data for gene `YLR449W` and its segment tree reconstruction in Fig. 4.4, our multi-scale approach, that combines measurements from different scales (segment lengths), allows for implicitly capturing changes in ribosome density along transcripts, while at the same time keeping the average ribosome density across larger segments close to the observed levels. This representation also encodes uncertainty about the density ratio at a particular region of the gene, even if that region is not directly represented by a segment from the tree. For example, region (85,104) (highlighted in the figure) is covered by 6 segments (i.e., has depth 5 within the tree) and has one of the tightest confidence intervals (CIs) in the reconstruction. At the same time region (105,120) was not measured at the two lowest scales (has a depth 3) and its average density has to be derived from the density values of other segments and our uncertainty about them, leading to a wider CI. This example demonstrates how segment trees capture changes in ribosome density along the transcript, which are crucial for fitting translation rates and evaluating competing models.

KNOWLEDGE-BASED MODELS DO NOT FIT RP DATA

Small standard deviations of the scaling factors and objective scores (determined using CV) of the evaluated models shown in Table 4.1 suggest that the (fitted) models perform consistently across different folds. The objective scores also show that

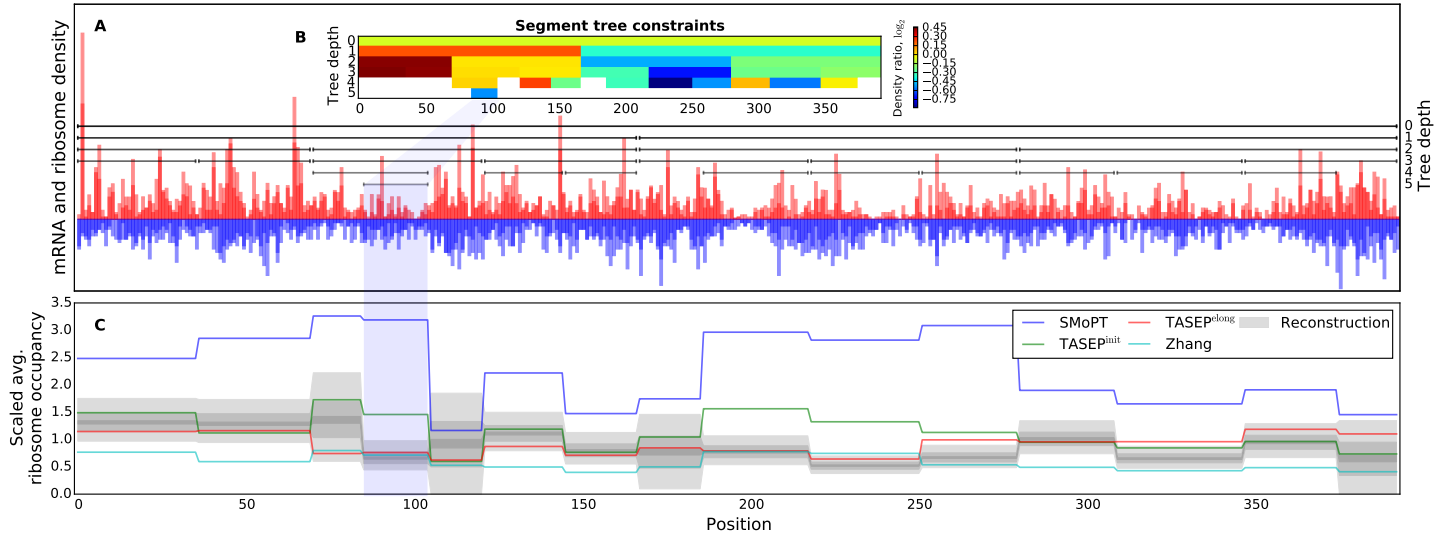

Figure 4.4: Ribosome profiling data, segment tree and simulated ribosome occupancy for gene YLR449W. (**A**) Ribo-seq (red) and mRNA-seq (blue) read density shown at single-codon resolution. Densities from the available replicates are overlaid with each other. Horizontal lines show the beginning and end of segments from the segment tree constructed for these densities. (**B**) Heatmap of the average density ratios recorded for each of the segments shows how density changes along the transcript for each of the scales (i.e. depths) within the segment tree. (**C**) Reconstruction of the per-transcript ribosome density from the segment tree (grey) shown as 90%, 50% and 10% confidence intervals (shades of grey). The reconstruction was obtained by sampling from the joint probability distribution derived from the segment tree (see Suppl. Text, page 116). Simulated ribosome occupancy for several considered models (blue, green, red and cyan solid lines) was averaged within segments and scaled to match the data.

Table 4.1: Objective ψ and scaling factor C for the evaluated models computed on the test folds inside a 5-fold CV loop.

Model	Fitted	$\ln C$	Objective ψ
Zhang	No	-4.55 ± 0.00	$-600\,286 \pm 4449$
SMoPT	No†	5.04 ± 0.01	$-244\,834 \pm 2962$
TASEP$^{\text{init}}$	Init.	5.40 ± 0.00	$99\,144 \pm 2137$
TASEP$^{\text{elong}}$	Yes	5.41 ± 0.02	$114\,865 \pm 4335$

† - RP data Ingolia *et al.* [11] was used in the original publication to set initiation rates.

knowledge-based models (i.e., the SMoPT and Zhang models) based on a manual choice of numerous translation-related parameters explain the ribosome density measurements significantly worse than the two models fitted on RP data. This can also be concluded from a visual inspection of the predictions made by these models for one of the genes in Fig. 4.4C, which shows that their ribosome occupancies tend to "miss" the measured density ratios. For the Zhang model this could be explained by the absence of gene-specific initiation rates in the model, whereas SMoPT often overshoots the measured density ratios, presumably because it over-estimates initiation rates by neglecting ribosome queueing.

The TASEP$^{\text{init}}$ model simulated with tAI-based elongation rates and fitted initiation rates achieves a significantly higher objective scores than the two state-of-the-art models. It is further improved by the TASEP$^{\text{elong}}$ model, for which the elongation rates are additionally fit on the segment tree measurements. Fig. 4.5 shows that superior objective function values of the fitted models translate to better predictions of the measured ribosome density (Pearson correlation coefficient $r = 0.77$ vs. 0.45, $p < 10^{-293}$). Although the predictions are generally better for longer segments, improvements can be observed at all scales (see Suppl. Fig. 4.13). While due to its relative simplicity only a weak positive correlation was expected for the Zhang model, for reasons unclear, a highly significant ($p < 10^{-293}$) *negative* correlation is observed (Fig. 4.5, left). This demonstrates that current knowledge-based models are not supported by RP measurements and highlights the importance of a critical evaluation of existing translation models against independent measurements.

TASEP PREDICTIONS ARE SUPPORTED BY INDEPENDENT DATASETS

Although TASEP$^{\text{init}}$ and TASEP$^{\text{elong}}$ outperformed existing models in the CV experiments, care has to be taken when interpreting these results as only the TASEP models were fitted directly on the segment tree measurements. We sought to obtain additional confirmation of the models' performance and to determine if they make biologically meaningful predictions. To this end we compared the protein production and translation initiation rates given by TASEP models fitted on all segment tree estimates to several independent large-scale PA datasets (see Materials and Methods).

Most importantly, we found that for both models the predicted protein production rates (PPRs) J positively correlate with the PA measurements (Table 4.2). As expected, because J describes PPR per transcript, these correlations improve when the product of

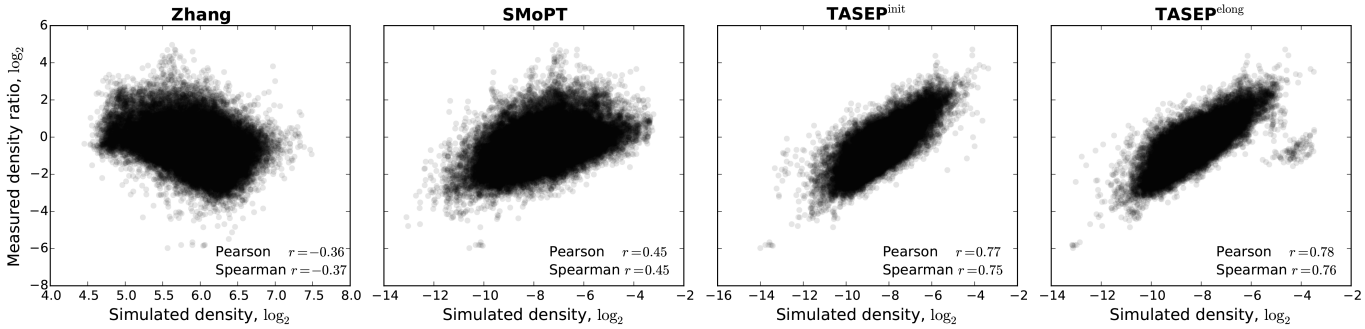

Figure 4.5: Measured segment density ratios $\mu_{[l_j,r_j]}$ plotted against the segment-averaged predicted ribosome occupancies for several existing and proposed models. Ribosome occupancy predictions made by the fitted models show significantly better agreement with the RP data. Reported correlations are highly significant ($p < 10^{-293}$).

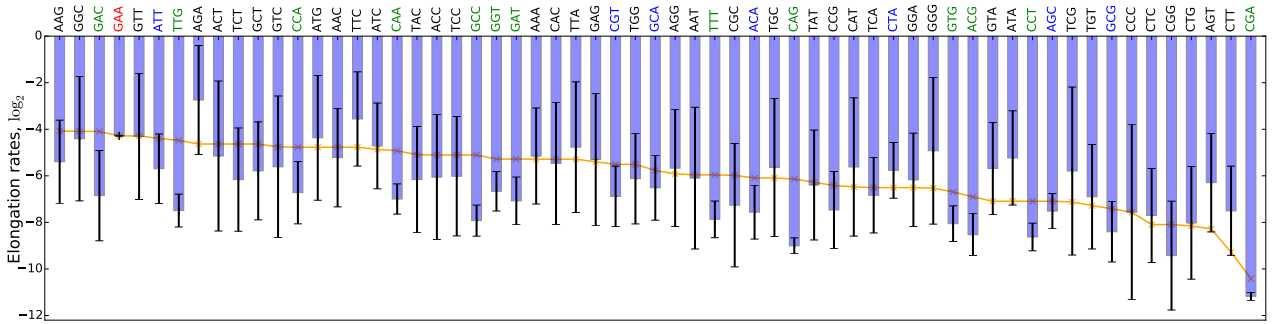

Figure 4.6: Mean and SD of the codon elongation rates fitted in different CV folds, compared to the tAI-based rates. For many codons elongation rates (depicted as mean and SD, blue bars) are reproducible across CV folds. This becomes evident for codons with smaller SDs (blue labels, $\sigma < 1.5$), and codons whose elongation rates are significantly different from the tAI-based rates (green labels; t-test, $p < 0.05$). tAI-based rates (orange line) are plotted as a reference. The rate of codon GAA (red label) was not optimised. Stop codons were excluded from the figure as their fitted termination rates remained very close to the original values of 1.

Table 4.2: Correlations of TASEP predictions with independent PA datasets. Spearman rank correlation coefficients r for are reported; J' is the partial correlation between J and PA given mRNA.

	$\text{TASEP}^{\text{init}}$		
	Newman YEPD	Newman SD	Ghaemmaghami
Init. rate	$r = 0.56^{***}$	$r = 0.55^{***}$	$r = 0.49^{***}$
J	$r = 0.57^{***}$	$r = 0.56^{***}$	$r = 0.50^{***}$
$J \times$ mRNA	$r = 0.72^{***}$	$r = 0.70^{***}$	$r = 0.63^{***}$
J'	$r = 0.52^{***}$	$r = 0.49^{***}$	$r = 0.39^{***}$
	$\text{TASEP}^{\text{elong}}$		
	Newman YEPD	Newman SD	Ghaemmaghami
Init. rate	$r = 0.54^{***}$	$r = 0.53^{***}$	$r = 0.49^{***}$
J	$r = 0.56^{***}$	$r = 0.53^{***}$	$r = 0.49^{***}$
$J \times$ mRNA	$r = 0.72^{***}$	$r = 0.70^{***}$	$r = 0.63^{***}$
J'	$r = 0.52^{***}$	$r = 0.48^{***}$	$r = 0.39^{***}$

* - $p-$value $< 10^{-5}$ ** - $p-$value $< 10^{-20}$ *** - $p-$value $< 10^{-100}$

J and mRNA levels ($J \times$ mRNA; mRNA levels taken from the RP data) is considered. Even when both J and PAs are corrected for mRNA levels (thereby removing transcriptional regulatory influences in order to study translational regulation in isolation), the remaining (partial) correlation between J' and PA$'$ is still significant, indicating that our TASEP models adequately capture the effects of protein translation on protein levels. These correlations are superior compared to correlations observed for state-of-the-art models (Table 4.3), especially when the partial correlations are considered. While strong positive partial correlations would be expected, we find these only for the fitted TASEP models. Unexpectedly low and negative partial correlations between PA$'$ and J' for other models, together with strong correlations between PPR and mRNA levels (Table 4.4) suggest that existing models are overfit on transcript levels and may not accurately model translation. These findings provide an independent confirmation that the TASEP models with fitted translation rates accurately capture the dynamics of the *S. cerevisiae* translation machinery.

Looking more in detail (Table 4.4), we find that for both models the fitted initiation rates agree well with the rates inferred by the existing full-cell models of Shah *et al.* (SMoPT), and of Siwiak and Zielenkiewicz. However, we did not find the previously reported strong negative correlation between initiation rates and CDS length [10, 22]. We note that this correlation is also not supported by the model of Siwiak and Zielenkiewicz. The initiation rates also exhibit a weak correlation with the 3$'$ UTR lengths (similar correlations also found for several other models), supporting the hypothesis of more efficient translation re-initiation in genes with longer 3$'$ UTRs.

Interestingly, we did not find the tendency for genes with short 5$'$ UTRs to exhibit high initiation rates suggested by Shah *et al.* and supported by Ciandrini *et al.* [22] in our models or the model of Siwiak and Zielenkiewicz. We also note that no relationship or a negative relationship can be observed between initiation rates and 5$'$ UTR lengths corrected for CDS lengths can be found in most considered models. This suggests that the previously observed inverse relationship between 5$'$ UTR lengths and initiation rates may not be indicative of a 5$'$ UTR-mediated initiation rate regulation mechanism, but

Table 4.3: Correlations of predictions made by existing models with independent PA datasets. Spearman rank correlation coefficients r are reported.

	Siwiak and Zielenkiewicz		
	Newman YEPD	Newman SD	Ghaemmaghami
Init. rate	$r = 0.45^{**}$	$r = 0.48^{***}$	$r = 0.40^{***}$
J	$r = 0.33^{**}$	$r = 0.36^{**}$	$r = 0.37^{***}$
$J \times$ mRNA	$r = 0.58^{***}$	$r = 0.54^{***}$	$r = 0.50^{***}$
J'	$r = -0.12^{*}$	$r = -0.07$	$r = -0.01$
	SMoPT		
	Newman YEPD	Newman SD	Ghaemmaghami
Init. rate	$r = 0.45^{**}$	$r = 0.49^{***}$	$r = 0.44^{***}$
J	$r = 0.21^{**}$	$r = 0.23^{**}$	$r = 0.26^{**}$
$J \times$ mRNA	$r = 0.45^{**}$	$r = 0.46^{**}$	$r = 0.46^{***}$
J'	$r = -0.26^{**}$	$r = -0.21^{*}$	$r = -0.13^{*}$
	Ciandrini *et al.* [22]		
	Newman YEPD	Newman SD	Ghaemmaghami
Init. rate	$r = 0.44^{***}$	$r = 0.43^{***}$	$r = 0.43^{***}$
J	$r = 0.45^{***}$	$r = 0.44^{***}$	$r = 0.44^{***}$
$J \times$ mRNA	$r = 0.57^{***}$	$r = 0.56^{***}$	$r = 0.55^{***}$
J'	$r = 0.10^{*}$	$r = 0.10^{*}$	$r = 0.14^{*}$

* - $p-$value $< 10^{-5}$ ** - $p-$value $< 10^{-20}$ *** - $p-$value $< 10^{-100}$

could be merely a consequence of a positive correlation between 5′ UTR lengths and CDS lengths.

While correlations observed for the fitted models do not change between TASEP$^{\text{init}}$ and TASEP$^{\text{elong}}$ (Table 4.4), the latter model makes considerably better ribosome occupancy predictions. It can be seen from the example in Fig. 4.4C that fitting the elongation rates allows the segment-averaged ribosome occupancy of TASEP$^{\text{elong}}$ to follow the reconstructed density considerably better than any of other model.

FITTED ELONGATION RATES ARE NOT EXPLAINED BY ADAPTATION TO TRNA LEVELS ALONE

Since the TASEP$^{\text{elong}}$ model achieves a significantly better fit to the RP data compared to TASEP$^{\text{init}}$ with tAI-based rates (Table 4.1), having fitted its elongation rates on different CV folds, we sought to interpret the obtained values and their variance. We first, however, confirmed that elongation rates determined from different RP datasets agree qualitatively with each other by fitting a new TASEP$^{\text{elong}}$ model on the dataset of Ingolia *et al.* [11] and comparing its translation rates to the original model (see Suppl. Text, page 108).

It can be seen from Fig. 4.6 that despite the generally large SDs, for many codons the elongation rates fitted in different folds of the CV are spread compactly around codon-specific values. This is clearly visible for codons with smaller SDs (green and blue), for which similar rates were found in different folds. Nonetheless the rate SDs differ considerably between codons. While the majority of the fitted elongation rates are consistently different from tAI-based rates, only for 13 codons this difference is

Table 4.4: Comparison of TASEP predictions to existing models. Spearman rank correlation coefficients r are reported. When "corrected for" column is non-empty, partial correlations are reported.

Variable 1	Variable 2	Corrected for	Correlation coeff.	p-value
$\text{TASEP}^{\text{init}}$ init. rates	SMoPT init. rates		$r = 0.67$	$p < 10^{-298}$
	Siwiak and Zielenkiewicz init. rates		$r = 0.74$	$p < 10^{-298}$
	Ciandrini *et al.* init. rates		$r = 0.47$	$p < 10^{-197}$
$\text{TASEP}^{\text{elong}}$ init. rates	$\text{TASEP}^{\text{init}}$ init. rates		$r = 0.94$	$p < 10^{-298}$
	SMoPT init. rates		$r = 0.65$	$p < 10^{-298}$
	Siwiak and Zielenkiewicz init. rates		$r = 0.73$	$p < 10^{-298}$
	Ciandrini *et al.* init. rates		$r = 0.46$	$p < 10^{-182}$
CDS lengths	$\text{TASEP}^{\text{init}}$ init. rates		$r = -0.07$	$p < 10^{-4}$
	$\text{TASEP}^{\text{elong}}$ init. rates		$r = -0.05$	$p < 10^{-2}$
	SMoPT init. rates		$r = -0.52$	$p < 10^{-257}$
	Siwiak and Zielenkiewicz init. rates		$r = -0.02$	$p > 10^{-1}$
	Ciandrini *et al.* init. rates		$r = -0.65$	$p < 10^{-298}$
5′ UTR lengths	$\text{TASEP}^{\text{init}}$ init. rates		$r = -0.01$	$p > 10^{-1}$
	$\text{TASEP}^{\text{elong}}$ init. rates		$r = -0.02$	$p > 10^{-1}$
	SMoPT init. rates		$r = -0.06$	$p < 10^{-3}$
	Siwiak and Zielenkiewicz init. rates		$r = 0.00$	$p > 10^{-1}$
	Ciandrini *et al.* init. rates		$r = -0.09$	$p < 10^{-10}$
	$\text{TASEP}^{\text{init}}$ init. rates	CDS lengths	$r = 0.00$	$p > 10^{-1}$
	$\text{TASEP}^{\text{elong}}$ init. rates	CDS lengths	$r = -0.01$	$p > 10^{-1}$
	SMoPT init. rates	CDS lengths	$r = 0.03$	$p > 10^{-1}$
	Siwiak and Zielenkiewicz init. rates	CDS lengths	$r = 0.03$	$p < 10^{-1}$
	Ciandrini *et al.* init. rates	CDS lengths	$r = -0.06$	$p < 10^{-3}$
3′ UTR lengths	$\text{TASEP}^{\text{init}}$ init. rates		$r = 0.04$	$p < 10^{-2}$
	$\text{TASEP}^{\text{elong}}$ init. rates		$r = 0.04$	$p < 10^{-1}$
	SMoPT init. rates		$r = 0.06$	$p < 10^{-3}$
	Siwiak and Zielenkiewicz init. rates		$r = 0.07$	$p < 10^{-5}$
	Ciandrini *et al.* init. rates		$r = 0.03$	$p < 10^{-1}$
	$\text{TASEP}^{\text{init}}$ init. rates	CDS lengths	$r = 0.04$	$p < 10^{-1}$
	$\text{TASEP}^{\text{elong}}$ init. rates	CDS lengths	$r = 0.04$	$p < 10^{-1}$
	SMoPT init. rates	CDS lengths	$r = 0.07$	$p < 10^{-4}$
	Siwiak and Zielenkiewicz init. rates	CDS lengths	$r = 0.08$	$p < 10^{-6}$
	Ciandrini *et al.* init. rates	CDS lengths	$r = 0.02$	$p > 10^{-1}$
mRNA levels	$\text{TASEP}^{\text{init}}$ init. rates		$r = 0.36$	$p < 10^{-115}$
	$\text{TASEP}^{\text{elong}}$ init. rates		$r = 0.33$	$p < 10^{-93}$
	SMoPT init. rates		$r = 0.58$	$p < 10^{-298}$
	Siwiak and Zielenkiewicz init. rates		$r = 0.33$	$p < 10^{-117}$
	Ciandrini *et al.* init. rates		$r = 0.62$	$p < 10^{-298}$
	$\text{TASEP}^{\text{init}}$ J		$r = 0.34$	$p < 10^{-97}$
	$\text{TASEP}^{\text{elong}}$ J		$r = 0.37$	$p < 10^{-115}$
	SMoPT J		$r = 0.65$	$p < 10^{-298}$
	Siwiak and Zielenkiewicz J		$r = 0.69$	$p < 10^{-298}$
	Ciandrini *et al.* J		$r = 0.63$	$p < 10^{-298}$
mRNA levels	Newman YEPD PA		$r = 0.58$	$p < 10^{-209}$
	Newman SD PA		$r = 0.57$	$p < 10^{-194}$
	Ghaemmaghami PA		$r = 0.54$	$p < 10^{-273}$

Continued on next page.

Table 4.4 continued.

CDS lengths	Newman YEPD PA		$r = -0.13$	$p < 10^{-10}$
	Newman SD PA		$r = -0.14$	$p < 10^{-12}$
	Ghaemmaghami PA		$r = -0.16$	$p < 10^{-22}$
	Newman YEPD PA	mRNA	$r = 0.32$	$p < 10^{-60}$
	Newman SD PA	mRNA	$r = 0.28$	$p < 10^{-42}$
	Ghaemmaghami PA	mRNA	$r = 0.21$	$p < 10^{-36}$
	mRNA		$r = -0.53$	$p < 10^{-298}$
	5′ UTR lengths		$r = 0.14$	$p < 10^{-20}$
	3′ UTR lengths		$r = -0.03$	$p < 10^{-1}$

statistically significant (single sample t-test for population mean difference, $p < 0.05$; Fig. 4.6, Supplementary Data): GAC, TTG, CCA, CAA, GCC, GGT, GAT, TTT, CAG, GTG, ACG, CCT and CGA. Although these differences between the tAI-based and fitted elongation rates are challenging to explain, their presence suggests that additional unknown factors are shaping these rates.

Having identified differences in elongation rates between the TASEP$^{\text{init}}$ and TASEP$^{\text{elong}}$ models, we sought to understand their effect on models' predictions. As could be expected from the similar correlations in Table 4.4 and Fig. 4.5, the two models make very similar PPR and ribosome density predictions (Suppl. Fig. 4.11). However, ribosome density predicted by the TASEP$^{\text{elong}}$ model with fitted elongation rates agrees better with RP measurements. To understand the cause of this improvement we looked for genes whose fit to the RP data improved when fitted elongation rates were used. These genes can be classified into two groups: (i) genes that have a very similar initiation rate in both models (Fig. 4.7, left); and (ii) genes that have a considerably lower initiation rate in the TASEP$^{\text{elong}}$ model (Fig. 4.7, right). Because all 13 codons with significantly different elongation rates were predicted to be slower, their presence in CDSes generally leads to higher predicted ribosome occupancy, especially if the genes initiation rate remains unchanged. For genes from the first group, such as YOR202W shown on the left panel of Fig. 4.7, this already results in a more accurate ribosome occupancy prediction. For most other genes, the second group, this increase in codon elongation times yields ribosome occupancy that is too high under the current initiation rate. For these genes (e.g., YGR284C on the right panel of Fig. 4.7) a smaller fitted initiation rate is required to reduce ribosome occupancy that would otherwise be too high due to the effects of slow codons and high ribosomal flux (due to high initiation rate). Together these effects allow the model to better match the ribosome density changes along the transcript.

SIGNIFICANCE OF THE FITTED ELONGATION RATES FOR CODON OPTIMISATION

Codon optimisation, the process of substituting codons with synonymous alternatives that are elongated faster, thus contributing to the overall protein production rate, is routinely used to improve protein expression [40, 41]. Nonetheless, it remains a controversial tool because the same optimisation techniques can lead to contradicting results when applied to different proteins [42]. Here we compare our fitted elongation rates to codon optimality estimated by the commonly used tAI [33] and CAI [35] indices.

We considered the relative adaptiveness of a codon (see Materials and Methods)

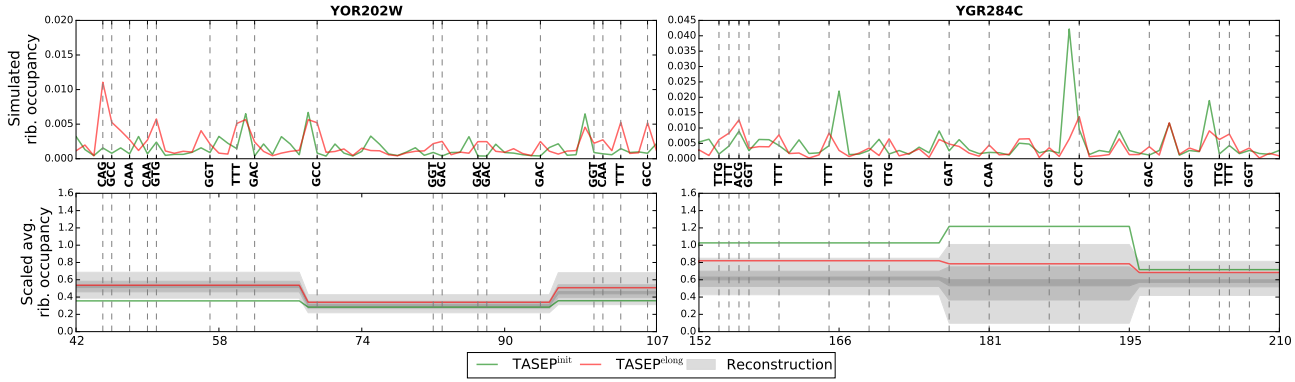

Figure 4.7: Ribosome density reconstruction (grey, bottom panels) and simulated ribosome occupancy (top) for selected regions of genes YOR202W (left panels) and YGR284C (right panels) plotted for the TASEP$^{\text{init}}$ (green) and TASEP$^{\text{elong}}$ (red) models. Presence of codons with significantly different elongation rates (vertical dashed lines) increases simulated ribosome occupancy. Higher increase can be observed for segments containing more such codons. This is clearly seen for gene YOR202W (**left**) with similar initiation rates in the TASEP$^{\text{init}}$ and TASEP$^{\text{elong}}$ models (0.24×10^{-4} and 0.22×10^{-4} respectively), for which the predicted occupancy only increases when fitted elongation rates are used. For most genes, such as YGR284C (**right**) this increase in density is compensated by reducing the initiation rate (from 0.72×10^{-4} to 0.36×10^{-4}), which leads to an overall better agreement between simulated ribosome occupancy and the segment tree measurements (bottom right). To keep the visualisation comprehensible, only selected regions of genes YOR202W and YGR284C were used. However, the described trends also hold for the remainder of these genes and for other genes.

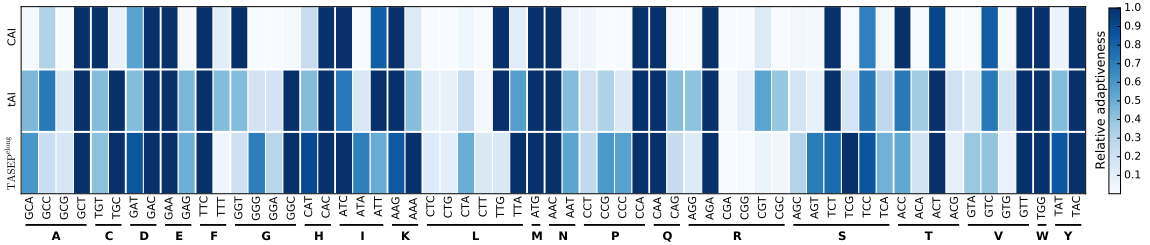

Figure 4.8: tAI and CAI compared to a measure of codon adaptation derived from the fitted TASEP$^{\text{elong}}$ elongation rates. Relative adaptiveness of codons grouped by their corresponding amino acids (columns) plotted for the three measures of codon adaptation (rows) shows that the considered measures often agree on the optimal codon. In particular, the tAI and TASEP$^{\text{elong}}$ measure agree on the optimal codons for all but 4 amino acids (I, K, L and S).

given by the CAI, the tAI and the fitted elongation rates of the TASEP^{elong} model. Fig. 4.8 shows that the three measures of codon adaptation often agree on the optimal codon for a particular amino acid (relative adaptiveness of 1.0, dark blue), which further demonstrates that our findings are in line with the earlier work. In particular, despite significant differences between the fitted elongation rates and elongation rates given by the tRNA adaptation hypothesis, the two sets agree on optimal codons for all but four amino acids. Only for isoleucine, leucine, lysine and serine the TASEP^{elong} model suggests codons `ATC`, `AAA`, `TTA` and `TCG` instead of `ATT`, `AAG`, `TTG` and `TCT` respectively. An interesting observation is that the bottom row in Fig. 4.8 is much more blue than the top ones, suggesting codon optimisation is less black-and-white than suggested by tAI and in particular CAI, meaning that many more codons are "reasonably good", i.e., there may be less to gain by codon optimisation than thought before. This observation is also corroborated by Leavitt and Alper [43], who noted that the level of control achievable in yeast through codon optimisation is considerably smaller than what can be achieved through transcriptional regulation.

TRANSLATION INITIATION LIMITS PROTEIN PRODUCTION

It is still unclear whether translation of endogenous yeast genes is limited by initiation or elongation [44, 45]. To test whether translation is limited by the initiation rates or by the elongation rates we artificially increased the initiation rate of each gene from the TASEP^{elong} model by 10%. To obtain robust results the experiment was repeated 5 times with different random initialisations and the average increase in PPRs was calculated for every gene.

Fig. 4.9 shows the relative differences in PPRs for all genes. In almost all cases (except 7 genes) the PPR increased substantially (relative difference > 0.02) when increasing the initiation rate, supporting the hypothesis that under exponential growth in the rich medium translation in *S. cerevisiae* operates in an initiation-limited regime. This also explains why fitting the codon elongation rates in TASEP^{elong} did not improve the PA correlations compared to the TASEP^{init} model. Elongation-limited production for these genes can be explained by the very high initiation rates predicted for them, which shift the rate-limiting step from translation initiation to translation elongation. Interestingly, groups of genes that had a low, medium and high PPR increase are enriched for several biological functions (FDR < 0.05, Fig. 4.9). Notably, genes in the high increase group are involved in negative regulation of various biosynthetic and metabolic processes. This suggests that yeast cells may have evolved to rapidly "switch on" negative regulation by keeping a buffer of the required mRNA transcripts that are efficiently translated only once there is demand.

4.4. DISCUSSION

For the first time, we described an approach that derives complete translation kinetics of an organism from ribosome profiling data and used it to simultaneously infer the translation elongation, translation initiation and protein production rates all together without neglecting the effects of ribosomal interference. We applied our methodology to the ribosome and RNA sequencing data of the baker's yeast *Saccharomyces cerevisiae*.

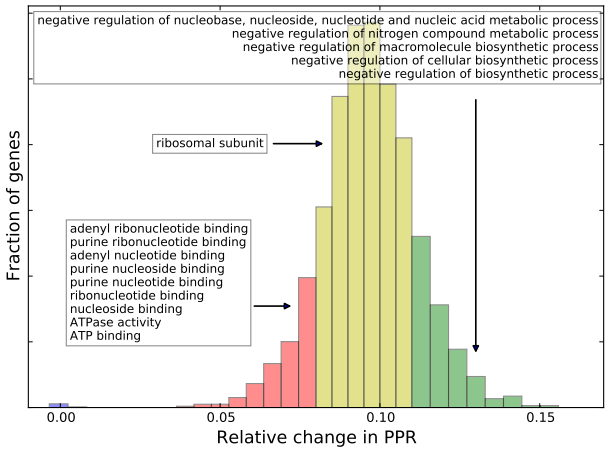

Figure 4.9: Distribution of the relative changes in PPR after a 10% increase in initiation rates shows that translation initiation is the rate limiting step for the protein production for most *S. cerevisiae* genes from the considered RP dataset. Groups of genes with low (≥ 0.02 and < 0.08, red), medium (≥ 0.08 and < 0.11, yellow) and high (≥ 0.11, green) increase in PPRs are enriched for several biological functions (white boxes in the figure, FDR < 0.05).

The fitted yeast translation models agree considerably better with independent protein abundance datasets than existing models. In particular, our TASEP models are the only ones that maintain strong correlations with protein abundance after removing the effect of transcriptional regulation.

While translation initiation rates provided by the models are similar to rates from other studies, we did not find the previously reported negative correlation between initiation rates and CDS lengths. The observed negative correlations between PA and CDS length, which one would expect to see as a result of this correlation, can alternatively be explained by transcriptional regulation, i.e., the strong negative correlation between mRNA levels and CDS lengths (Table 4.4). An alternative explanation can be offered by a mechanism driven by amino acid chain elongation rather than translation initiation. For example, abortive translation or the degradation of misfolded proteins [46], since the chance of producing a misfolded protein is expected to increase with protein length.

We also found that translation elongation rates deviate considerably from the widely accepted tRNA pool adaptation hypothesis, for 13 codons significantly so. Differences in elongation rates of these codons between the tRNA pool adaptation hypothesis and $\text{TASEP}^{\text{elong}}$ may be partially explained by nucleotide modifications of their respective tRNAs, which are known to modify the specificity and efficiency of messenger decoding [47]. As such, some of these 13 codons were shown to be affected by post-transcriptional nucleotide modifications of tRNAs in different organisms [48]. We speculate that for these codons the concentration of (un)modified tRNAs, rather than the total tRNA concentration, plays a non-negligible role in determining their elongation rates [19].

An additional factor that possibly contributes to the observed deviation from the tRNA pool adaptation hypothesis is its implicit assumption that different tRNA genes from the same family contribute equally to determining the rate of translation. This assumption should be revisited in light of the recent finding of Bloom-Ackermann *et al.* [49] that the contributions of different gene copies from the same tRNA family to the tRNA pool and cellular fitness are far from equal.

In our experiments we found that SDs of elongation rates from different CV folds differ markedly between codons. In order for the elongation rates to be specified with high precision by the RP data, small changes in the rates must lead to detectable differences in ribosome density. However, in light of our finding that yeast translation is initiation-limited and the observation of Bloom-Ackermann *et al.* [49] that *S. cerevisiae* is robust to deletions of tRNA genes, especially in rich medium used to produce the ribosome profiling measurements analysed here, it is unlikely that in the considered physiological conditions the elongation rates exert a strong enough effect on ribosome density to allow the RP data to specify elongation rates with high precision. We speculate that found SDs reflect the robustness of the yeast translation system w.r.t. the codon translation rates, with the system being more sensitive to changes in rates of those codons that have smaller SDs. In this case, yeast translation appears to be robust to fold changes in codon translation rates and, consequently, to the aminoacyl-tRNA availability that these rates are thought to be determined by [45].

Alternatively, the SDs may reflect the extent to which codon translation rates change between CV folds due to codon context, i.e., the local sequence around a codon which may alter its elongation rate (see Suppl. Text, page 108). It is unlikely that the TASEP model captures the full complexity of the translation process by assuming that codon elongation rates are determined solely by the codon identity, and not also by the sequence surrounding the codons as was previously suggested [3, 4]. Such a constraint limits the models ability to capture the underlying translation dynamics and may bias it towards fitting different rates on different sets of genes (e.g., CV folds) with varying codon contexts, thereby inflating the SDs. The observed variation in fitted elongation rates puts forward codon context as a factor that may significantly modulate the baseline elongation rates.

Using our models we found that under exponential growth in rich medium translation initiation appears to be the main limiting factor of protein production of endogenous genes in *Saccharomyces cerevisiae*, with protein production being limited by initiation rates for all but 7 genes with very high initiation rates. These findings suggest that rational design of $5'$ UTRs involved in translation initiation [50, 51] may be a more promising avenue for achieving protein over-expression than the routinely used codon optimisation techniques. It is likely, however, that further over-expression could be achieved using codon optimisation. Because once the gene is put under the translational control of an efficient $5'$-UTR, which is usually the case in heterologous gene expression, translation elongation is expected to become a rate-limiting factor. In such cases we recommend performing codon optimisation using the fitted $\text{TASEP}^{\text{elong}}$ elongation rates, which, while mostly agreeing with existing techniques, also demonstrate several differences.

Although we found that translation initiation appears to be the main factor limiting

protein production in yeast under exponential growth in rich medium, it is possible that different mechanisms are dominant in other organisms. For example, Li *et al.* [52] and Guimaraes *et al.* [53] discuss greater contribution of protein elongation respectively by anti-Shine-Dalgarno sequences and codon usage in *E. coli*. Our method could be applied to ribosome profiling data of other organisms to delineate the relative contribution of initiation and elongation.

All translation models proposed to date, including TASEP$^{\text{init}}$ and TASEP$^{\text{elong}}$, assume that translation elongation rates are not influenced by *codon context*, i.e., the sequence around a particular codon, although various factors affecting the speed of elongation have been suggested [3–5]. Variation in fitted elongation rates and the highly varying codon translation times recently observed by Dana and Tuller [54] suggest that codon context may play a more compelling role in determining translation rates than previously thought. Fortunately investigations of codon context are becoming feasible thanks to the growing adoption of ribosome profiling as a standard technique for studying translation. With the increasing amount of ribosome profiling measurements, data-driven approaches, such as the one described here, will become instrumental for delineating the effects of multiple competing translation mechanisms, for generating new hypothesis, and for constructing predictive models for use in other fields. These goals can be achieved by incorporating alternative translation mechanisms as sequence- and position-specific effects altering the codon elongation rates.

4.A. SUPPLEMENTARY INFORMATION

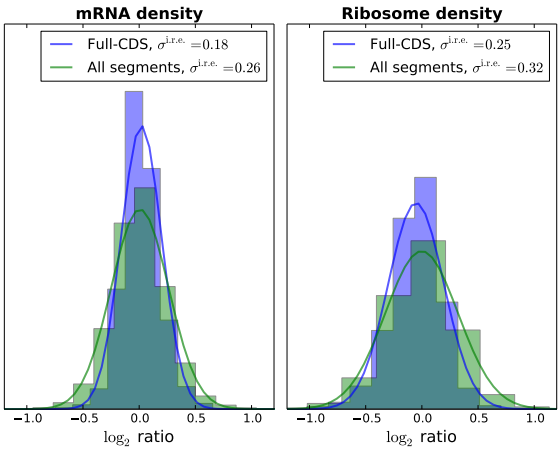

Figure 4.10: Histograms of the $\log_2$ inter-replicate errors (ratios of replicated measurements) of reliable ribosome and mRNA density measurements show that the full-CDS and segment tree density estimates follow comparable log-normal distributions. Distributions fitted into data (solid lines) are centered around zero, but their SDs differ.

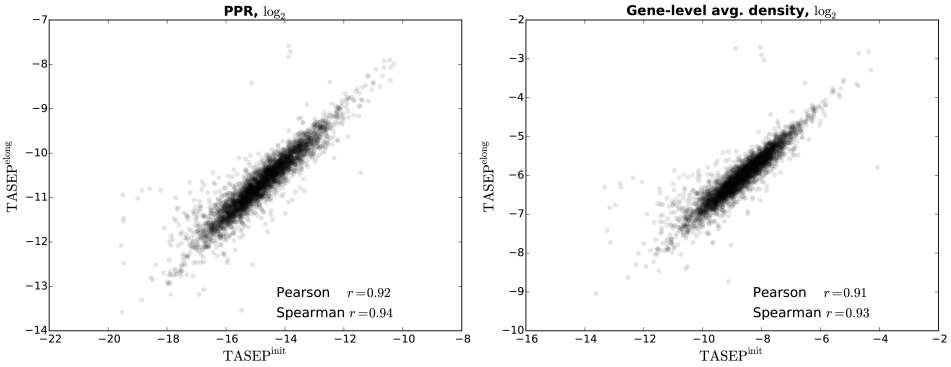

Figure 4.11: Agreement between the PPR (**left**) and gene-level average ribosome occupancy (**right**) predictions made by $\text{TASEP}^{\text{init}}$ and $\text{TASEP}^{\text{elong}}$ models.

RIBOSOME PROFILING READ PROCESSING

With the exception of elongation rate reproducibility analysis, the RP data for yeast *Saccharomyces cerevisiae* strain S288C [24] were used for all analyses. These data are available as raw sequencing reads that needed to be trimmed and aligned to the genome prior to any analysis. The read mapping procedure from [13] was used

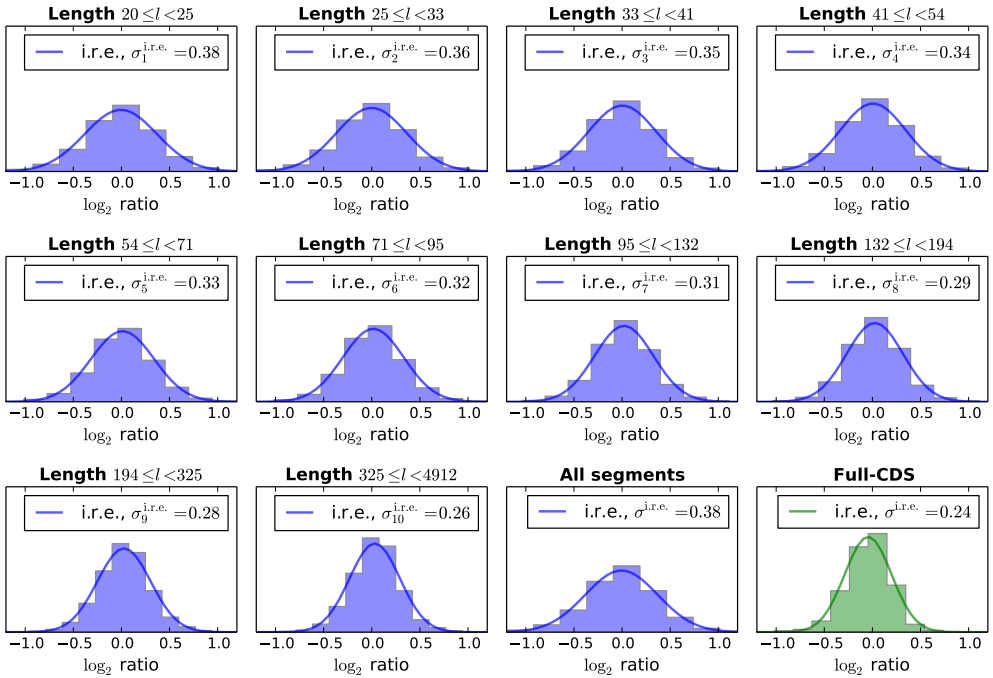

Figure 4.12: Histograms of the $\log_2$ inter-replicate errors of reliable density ratio measurements show similar error profiles in full-CDS and segment tree estimates. The group shape parameters of the i.r.e. and the density ratio distributions are related as $\sigma_k^{\text{group}} = \frac{1}{\sqrt{2}}\sigma_k^{\text{i.r.e.}}$.

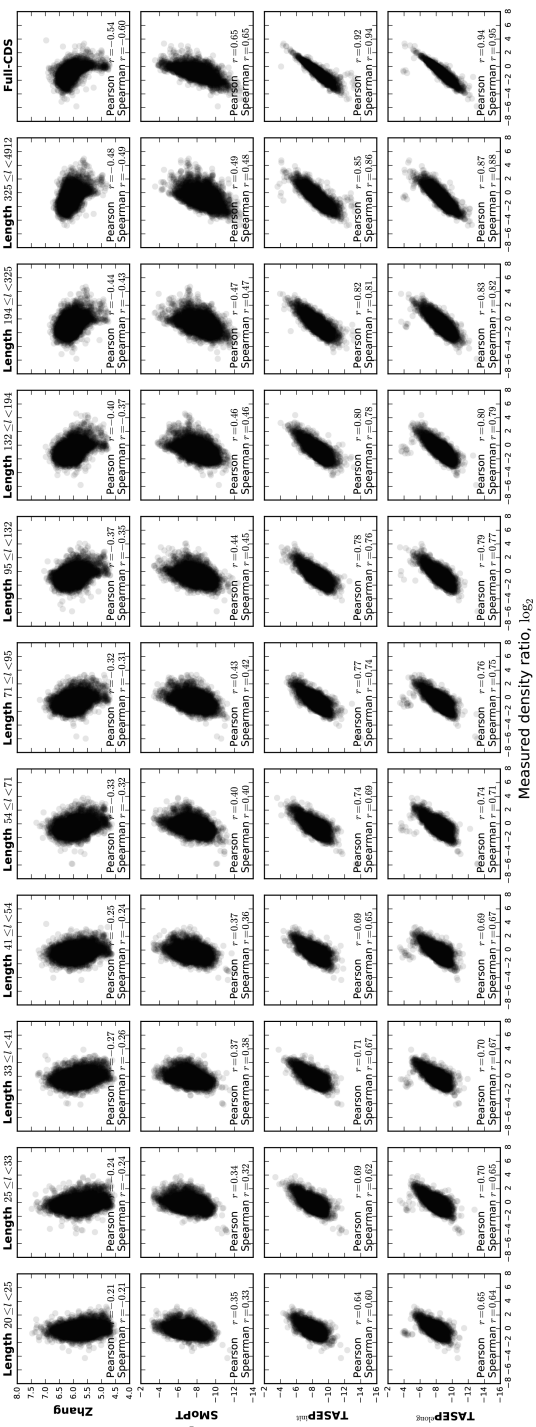

Figure 4.13: Measured segment density ratios $\mu_{\left[l_j, r_j\right]}$ plotted against the segment-averaged predicted ribosome occupancies for segments of varying size and for several existing and proposed models. $\mathrm{TASEP}^{\mathrm{init}}$ and $\mathrm{TASEP}^{\mathrm{elong}}$ significantly improve over existing models for all segment length groups.

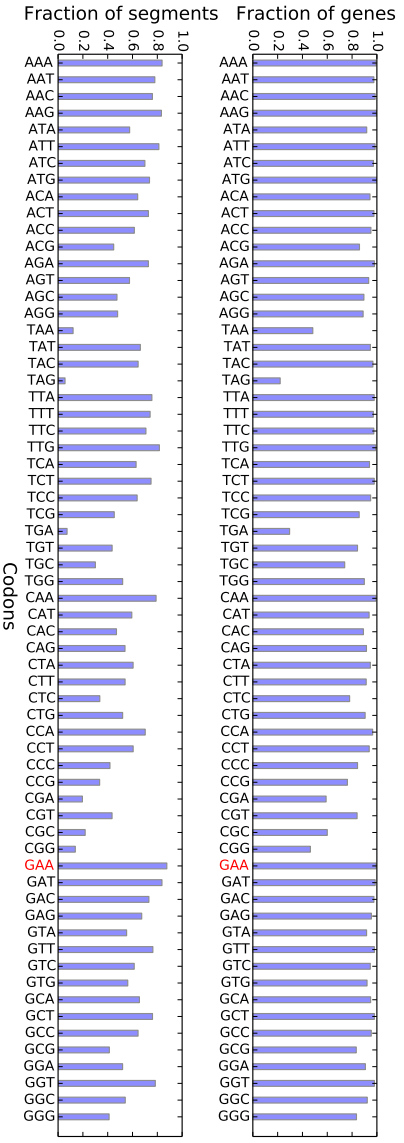

Figure 4.14: Presence of codons in gene and segment sequences from the segment tree. Translation rate of codon GAA (red) was fixed in elongation rate fitting experiments as it is present in many genes and segments.

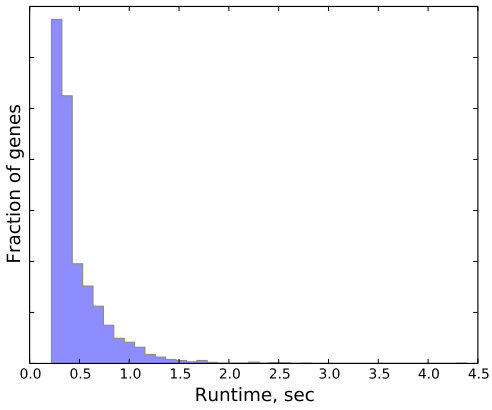

Figure 4.15: Histogram of the running times (average over 3 replicates) of the TASEP model simulations for genes in the evaluation set. tAI-based elongation rates and initiation rates of 1.0 were used in the simulations.

Table 4.5: Shape parameters of the density ratio distributions for segments grouped by length. Left (inclusive) and right (exclusive) edges give the range of segment lengths of a given group.

#	Left	Right	Group size	Shape parameter σ, $\log_2$
1	20	25	5284	0.235565455789103
2	25	33	6804	0.216285904079930
3	33	41	5591	0.207713921552678
4	41	54	6361	0.199316542112745
5	54	71	6163	0.183918710525198
6	71	95	5989	0.177805660887211
7	95	132	6097	0.164353210454788
8	132	194	6063	0.159682537763375
9	194	325	6057	0.142796300946171
10	325	4912	6057	0.128137654321086

to align RP reads to the yeast genome. After trimming the reads to a length of 21 nt to remove any linker-adapter sequences, the trimmed reads were aligned to the S288C reference genome sequences (release R64, accessed on January 14, 2014) using Bowtie [55]. First, the reads were mapped to the annotated CDSes and UTRs (taken from SGD [25]) of S288C extended by 100 nt on each side, and then unaligned reads were mapped to the entire reference genome sequence. These alignments were then extended up to the original read length to minimise the number mismatches between the untrimmed read, the reference and the linker sequences. Sequences `CTGTAGGCACCATCAAT` and `AGATCGGAAGAGCACACGTCTGA` were used for the RP linker and Illumina sequencing adapter during extension. Alignments with up to 2 mismatches were accepted, and multiple alignments were allowed for a single read, but alignments with fewer mismatches were preferred. Following McManus *et al.* [24] only RNA- and ribo-seq alignments of lengths $27 \leq l \leq 40$ and $27 \leq l \leq 33$ respectively were kept for

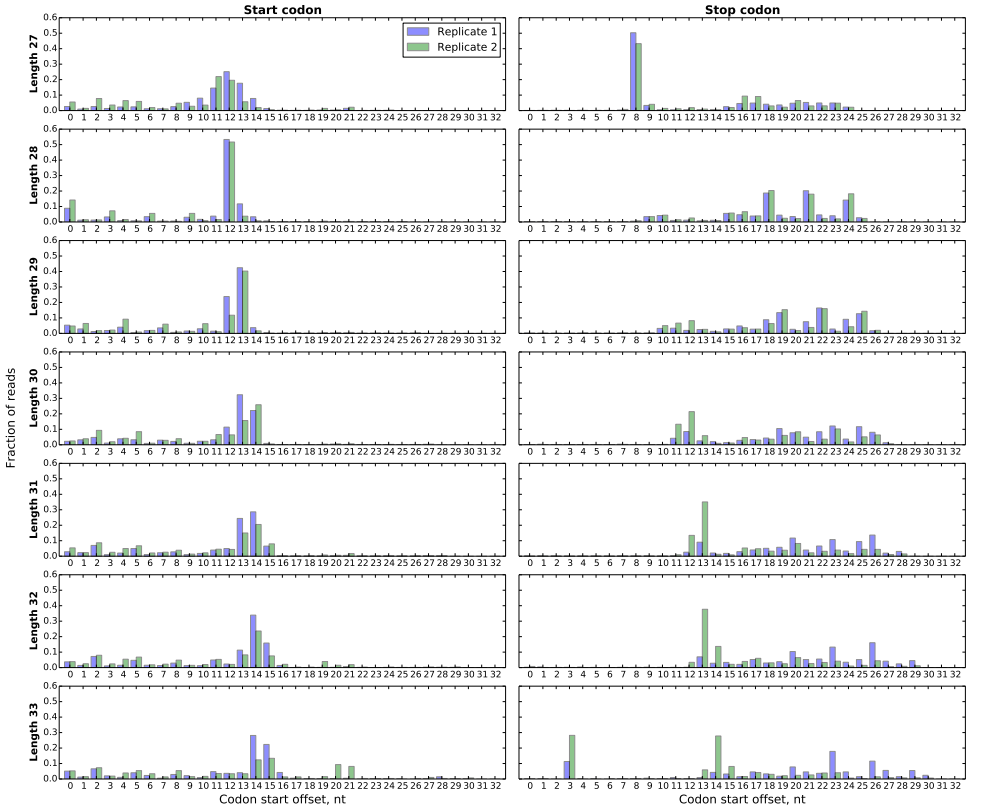

Figure 4.16: Histograms of the locations of the start and stop codons within ribosomal footprints of various lengths. Offsets give distances from the beginning of the read to the first nucleotides of the start or stop codons. High peaks around positions 11-14 on the start codon offset histograms directly give position of the P-site in ribosomal footprints, whereas the leftmost periodic peaks in stop codon offset histograms are located 6 nt upstream of the P-site. Offsets fixed for every footprint length are given in Table 4.6.

analysis.

We then sought to assign reads to the (parts of) CDSes that they originate from. Ribo-seq reads should be assigned to CDSes based on the position of the A-site codon in the read, which may differ with read length. Metagene analysis [13] was used to calibrate the position of the A-site codon for various footprint lengths. Reads with alignments containing start or stop codons of annotated CDSes were considered and the positions of these codons were recorded. Histograms of the positions of the start and stop codons in Fig. 4.16 were then used to determine the location of the P-site for each footprint length. The footprints were then assigned to CDSes based on the alignment coordinate of the overlap of the second nucleotide of the A-site codon (i.e., P-site offset +4) with annotated CDSes. RNA-seq reads were similarly assigned to CDSes based on the coordinate of their central nucleotide. For reads that map to multiple locations (ambiguous reads) an equal fraction of the read count was counted towards each location; and reads

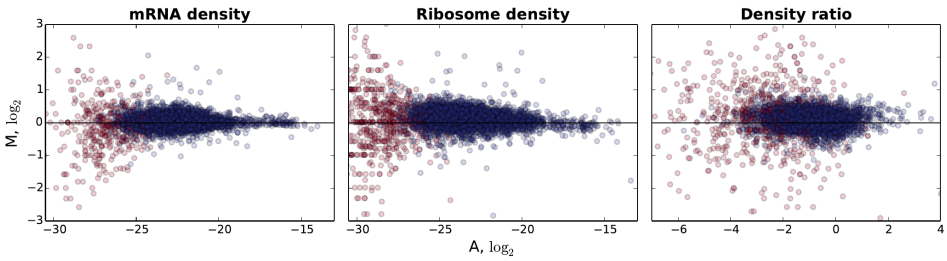

Figure 4.17: MA-plot of the full-CDS density estimates. Measurement variance for ribosome and mRNA density estimates is higher for genes with low density. Unreliable density estimates (< 128 total RNA-seq reads or < 128 ribo-seq reads; coloured red) show increased measurement variance.

assigned to multiple CDSes ("overlapping" reads) contributed their full read count to each assignment region. Read assignment was performed separately for each of the two mRNA- and ribo-seq biological replicates (see Table 4.7 for statistics).

SETTING THE READ COUNT THRESHOLDS

MA-plots typically applied to the analysis and normalisation of 2-channel microarray data [56] were used to visualise density measurement differences between replicates by plotting the log-ratio of the measurements $M = \log_2 a - \log_2 b$ against the log of their geometrical mean $A = 0.5 \cdot (\log_2 a + \log_2 b)$. Here a and b are density and density ratio measurements for the same gene (segment) from two different replicates. These plots were made for full-length CDSes and for segment trees.

MA plots for the full length genes (Fig. 4.17) were used to manually set the total read count thresholds for reliable density estimates. The chosen thresholds of 128 reads are identical to the ones used in Ingolia *et al.* [11] for defining reliably measured genes.

SEGMENT TREE CONSTRUCTION

When constructing segment trees, cut points p are chosen such that the combined number of RNA- and ribo-seq reads across replicates is divided equally between the left and the right segments. This is achieved by simultaneously minimising for the available replicates the sum of absolute per-replicate differences in the combined numbers of RNA- and ribo-seq reads between the left and the right segments.

When recursively splitting segments, cuts where both segments pass the minimum length criterion are preferred to cuts that minimise the read count imbalance. If multiple cut points with the same imbalance are available, the leftmost one is chosen. Measurements from segments, in which one or more density estimates are based on read counts containing ≥ 20% ambiguous or overlapping reads with other CDSes, are discarded, but tree construction is allowed to continue.

Table 4.6: P-site offsets for various footprint lengths determined individually for each replicate based on Fig. 4.16.

Length	Replicate 1		Replicate 2		Final offset
	Start	Stop	Start	Stop	
27	12	11	11	11	11
28	12	12	12	12	12
29	13	13	13	13	13
30	13	13	14	14	14
31	14	14	14	14	14
32	14	14	14	14	14
33	14	14	14	14	14

Table 4.7: Read alignment statistics for the *S. cerevisiae* genome release R64, and varying alignment strictness. Counts for reads aligned to any position in the reference genome and counts for reads assigned to coding sequences are reported separately; CDS counts were rounded to the nearest integer (in cases when reads had multiple alignments, only the fraction of alignments assigned to CDSes was counted).

Type	Name	Max. subst.	Min. length	Max. length	Unique	Replicate 1		Replicate 2	
						mRNA	Ribosome	mRNA	Ribosome
Ref.	Filtered	2	27	33/40	No	13,504,666	19,990,853	13,567,103	22,766,209
	Strict	1	27	33/40	No	12,827,218	18,782,657	12,870,000	21,360,880
	Unique	1	27	33/40	Yes	11,468,200	15,962,773	11,534,794	17,796,356
CDS	Filtered	2	27	33/40	No	10,859,192	15,983,167	10,911,163	16,173,551
	Strict	1	27	33/40	No	10,314,950	15,072,767	10,350,534	15,300,543
	Unique	1	27	33/40	Yes	9,037,207	12,936,748	9,081,107	12,945,505

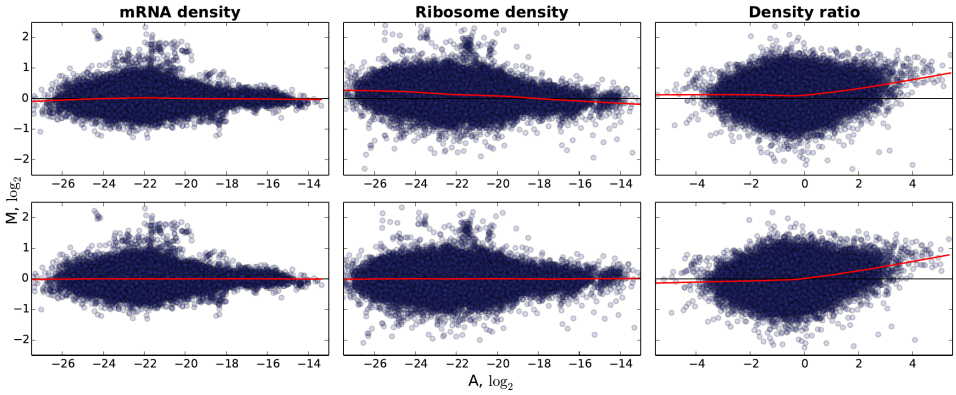

Figure 4.18: MA-plots of the segment tree density and ratio measurements (**top**) before bias correction show a density-dependent systematic bias, which (**bottom**) after removal of mRNA and ribosome density bias is no longer present in density estimates, but is amplified for the ratio estimates. An identical procedure is applied to correct this amplified bias (see Fig. 4.19). The locally estimated mean $\bar{M}$ (red line) was obtained using LOWESS [57] using 33% of the data.

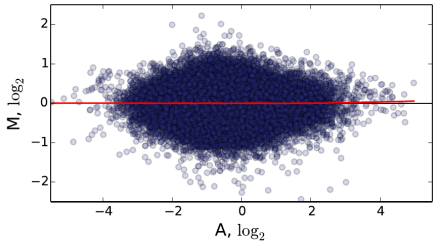

Figure 4.19: MA-plot of the segment tree density ratio estimates after normalisation shows that no significant bias is present at the extreme density ratio values.

DENSITY-DEPENDENT BIAS CORRECTION

MA-plots in Fig. 4.18 (left) suggest presence of a systematic density-dependent bias in density and ratio measurements. To remove the bias, density measurements were normalised by first estimating the local bias using LOWESS regression (red line in MA-plots) and then (i) subtracting it from M and (ii) subtracting a half of it from A. Although this bias is negligible for mRNA and density measurements, it may get amplified when the ratio of the two biased measurements is computed. This effect can already be seen from the density ratio MA-plots, where the bias only becomes more pronounced after density normalisation. The remaining bias in the density ratio estimates is removed be applying the normalisation procedure to the ratio estimates (Fig. 4.19).

We note that although this bias correction procedure does not allow for removing bias from the original measurements, it provides bias-corrected versions of quantities

M and A, which are sufficient to determine parameters of the log-normal distributions describing the segment tree measurements. Quantity A directly gives the scale parameter μ of the distribution for the corresponding segment, and M is essentially the i.r.e. which is used to determine segment length group shape parameters.

ESTIMATING SHAPE PARAMETERS FROM I.R.E.

To estimate shape parameters σ_j of the log-normal distributions $\ln\mathcal{N}(\mu_j,\sigma_j)$ describing density ratio measurements, segments were divided into 10 groups based on their length. Because each segment in a group follows a different distribution, with its own parameter μ_j, the per-group shape parameter cannot be estimated directly from the density ratio measurements. Instead, for each group k we estimate it via the shape parameter $\sigma_k^{\text{i.r.e.}}$ of the i.r.e. for measurements in this group. If $X, Y \sim \ln\mathcal{N}(\mu_j,\sigma_j)$ are random variables representing two independent replicated measurements of the density ratios, then the shape parameters of the i.r.e. and the density ratio distributions are related as

$$\left(\sigma_k^{\text{i.r.e.}}\right)^2 = \text{Var}(X) + \text{Var}(Y) = 2\cdot\left(\sigma_k^{\text{group}}\right)^2 \tag{4.7}$$

Using this equation the group shape parameter is calculated as $\sigma_k^{\text{group}} = \frac{1}{\sqrt{2}}\sigma_k^{\text{i.r.e.}}$ and used in place of σ_j for all segments in the group.

VARIANCE STRUCTURE IN SEGMENTS WITH HIGH AND LOW INITIATION RATES

It is possible that segments originating from genes with high initiation rates have a different variance structure (e.g., are more reliable) than genes with low initiation rates. If present, this kind of relationship would be missed by the proposed segment grouping strategy and render it problematic. To confirm that gene initiation rates do not significantly alter variance structure of their corresponding segments we plotted inferred gene initiation rates from several existing datasets [9, 10, 22] against the segments inter-replicate errors (i.e., M from the MA-plots) used for estimating the log-normal distribution shape parameters σ_k^{group} as described before. Figs 4.20 to 4.22 show that no strong relationship between the initiation rates and inter-replicate errors is present. Only for initiation rates obtained from Shah *et al.* [10] (Fig. 4.22) there appears to be a weak tendency of increasing i.r.e. for lower initiation rates. We believe that absence of strong dependencies between initiation rates and i.r.e. justifies our segment grouping approach and use it to derive log-normal distribution shape parameters as described above.

RIBOSOME OCCUPANCY PROFILES

The mRNA and ribosome occupancy profiles were obtained by assigning reads to coding sequences as in the case of segment tree construction. The nucleotide occupancy counts were then normalised by dividing them by N_R or N_M (the total number of ribo- and RNA-seq reads mapped to CDSes) depending on the profile; and the normalised counts were coarse-grained into codon-resolution count profiles by summing counts over nucleotide positions of the corresponding codons.

To obtain per-transcript ribosome occupancy profiles, the ribosome count profiles were further normalised in two different ways: either by dividing the per-position

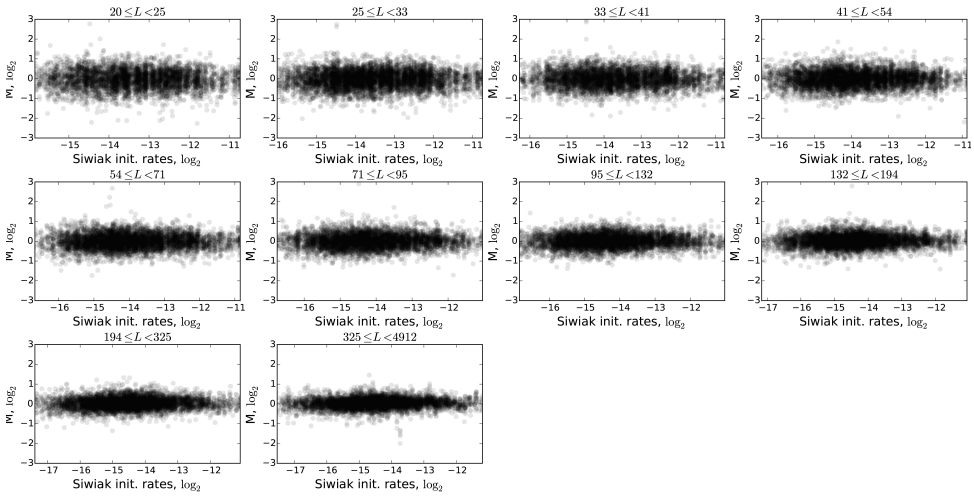

Figure 4.20: Density ratio i.r.e. (M from the MA-plot in Fig. 4.19) plotted against initiation rates obtained from Siwiak and Zielenkiewicz [9] for each of the 10 segment length groups used in the main text. No relationship between i.r.e. and initiation rates can be observed.

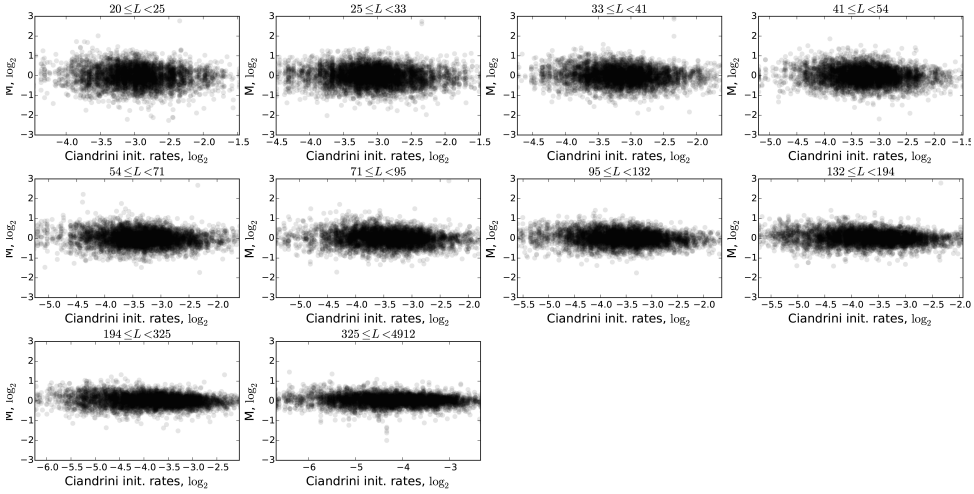

Figure 4.21: Density ratio i.r.e. M plotted against initiation rates obtained from Ciandrini *et al.* [22] for each of the 10 segment length groups used in the main text. No relationship between i.r.e. and initiation rates can be observed.

counts of the ribosome profiles by the average count of the corresponding mRNA profile (referred to as mean normalisation); or by dividing per-position counts of the ribosome profiles by the respective (same CDS and same position in the CDS) counts of the mRNA profiles (referred to as profile normalisation). The latter normalisation method is conceptually similar to the way in which density ratios in segment trees are calculated,

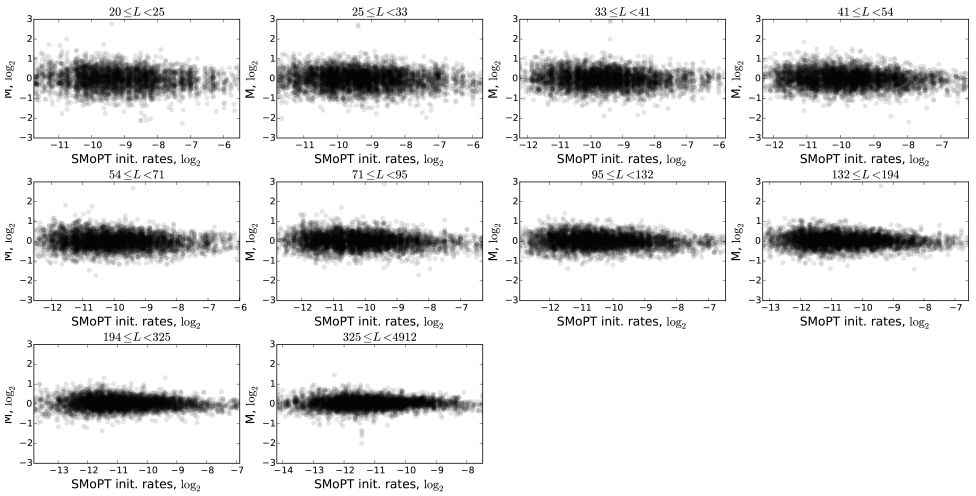

Figure 4.22: Density ratio i.r.e. M plotted against initiation rates obtained from Shah *et al.* [10] for each of the 10 segment length groups used in the main text. Only a weak relationship (the scatterplot has a droplet-like shape if viewed horizontally; clearly seen for group $194 \leq L < 325$) between i.r.e. and initiation rates can be observed.

but is applied at single codon resolution.

HIGH VARIANCE OF RP MEASUREMENTS AT SINGLE CODON RESOLUTION COMPLICATES INFERENCE OF TRANSLATION KINETICS

Prior to choosing for a "multi-scale" segment tree approach to interpreting the RP data, we evaluated its *quantitative* reproducibility at single-codon resolution by comparing ribosome occupancy profiles between replicates. To this end we obtained occupancy profiles using either profile or mean normalisation (PN and MN respectively). Separate ribosome occupancy profiles were obtained for the available biological replicates. For each reliably measured gene (as defined in Section 4.A) Pearson correlation between profiles obtained from the two replicates were calculated. Profile positions, for which it was impossible to obtain a profile (e.g., due to zero mRNA profile counts) in at least one of the replicates, were left out of the analysis.

We found that profile correlations demonstrate limited agreement of the ribosome occupancy profiles (median correlation coefficient $\tilde{r} = 0.55$; Fig. 4.23, left). This conclusion does not change even when more stringent read filtering is used (Fig. 4.24). Since PN profiles can be viewed as an extreme case of a segment tree, where segments do not overlap and are one codon in length, they were computed as a reference for the segment tree interpretation. We found that PN scheme performed worse than MN (median correlation coefficient of $\tilde{r} = 0.32$; Fig. 4.23, right), presumably because it introduced additional noise into the profiles through division by poorly estimated counts.

In the original publication Ingolia *et al.* [11] showed that RP data has good reproducibility when analysed at whole-gene scale. Given the low Pearson correlation

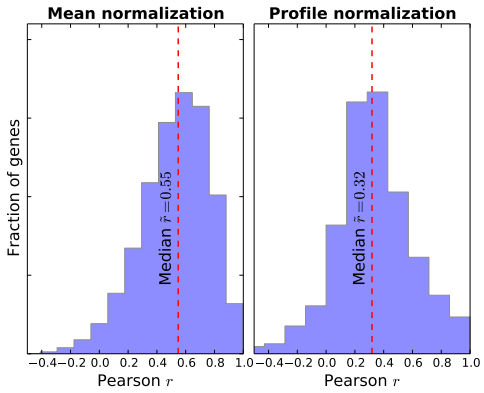

Figure 4.23: Histograms of Pearson correlation coefficients for ribosome occupancy profiles obtained from two biological replicates demonstrate limited reproducibility of the profiles in a majority of the reliably measured genes irrespective of the used normalisation method.

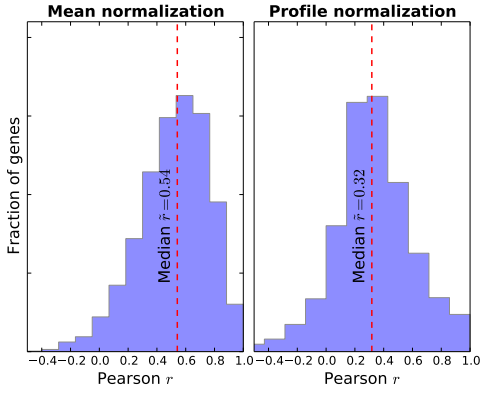

Figure 4.24: Histograms of Pearson correlation coefficients for ribosome occupancy profiles obtained from two biological replicates. Only uniquely mapping reads with at most 1 mismatch were used to construct the profiles (see Table 4.7 for statistics on read mapping). Nonetheless, the resulting correlation coefficients did not improve compared to the case of using less stringent read filtering (Fig. 4.23).

coefficients obtained for occupancy profiles evaluated at single codon resolution, we concluded that the RP data interpreted at single codon resolution would not allow for quantitative inference of translation kinetics; and devised a segmentation approach that estimates (local) average ribosome occupancy of a gene at multiple scales.

OBJECTIVE FUNCTION DERIVATION

In order to derive the objective function we assume that the density ratios obtained from the RP data follow the log-normal distribution. Further we assume that measurements for different genes and segments are independent from each other. We then seek to

quantify how likely it is that a certain *simulated* segment density matches the measured one by plugging it into the log-normal probability density function (PDF) of that segment. The objective function can then be written as a product of PDFs of individual segments:

$$F(n|T) = \prod_g \prod_{j \in J^g} f_x\left(N_j^g; \mu_j^g, \sigma_j^g\right),$$

where $f_x\left(x;\mu,\sigma\right) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$ is the log-normal PDF and other variables hold the following meaning:

- g denotes the gene.
- J^g denotes the set of segments of gene g.
- j denotes the segment within this gene.
- μ_j^g and σ_j^g are the scale and shape parameters of the log-normal distribution describing density ratios of segment j from gene g.
- T denotes the set of μ_j^g and σ_j^g for all genes and segments.
- n denotes the simulated ribosome occupancy for all genes at single codon resolution.
- N_j^g denotes the average simulated ribosome occupancy for segment j of gene g, which is matched against the estimated density ratio of the same segment.

The comparison of simulated average ribosome occupancy and the estimated density ratios is complicated by the fact that these are measured on different scales. We therefore need to rescale the simulated occupancy and the measured data to the same scale. We do this by scaling the ratio density distributions by a factor $\frac{1}{C}$, which for the moment we assume to be known. This is equivalent to transforming the PDF of the measured data into the PDF of the simulated occupancy and using the latter to evaluate simulation results. To derive the transformed PDF $f_y(y;\mu,\sigma)$ we apply Theorem 5.11 from [58].

Let x be the random variable with PDF $f_x(x;\mu,\sigma)$ representing the data distribution and let $y = \frac{1}{C}x$ be the rescaled version of this random variable that is on the simulation scale. The goal is then to determine $f_y(y;\mu,\sigma)$, i.e., the PDF of the rescaled variable. According to Theorem 5.11 f_y can be written as

$$f_y(y;\mu,\sigma) = \frac{1}{\frac{1}{C}} f_x(\frac{1}{\frac{1}{C}} y) = C f_x(Cy) = \frac{C}{Cy\sigma\sqrt{2\pi}} e^{-\frac{(\ln(Cy)-\mu)^2}{2\sigma^2}} = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{(\ln y+\ln C-\mu)^2}{2\sigma^2}} =$$
$$= \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{(\ln y-(\mu-\ln C))^2}{2\sigma^2}} = f_x(y;\mu-\ln C,\sigma)$$

We can now go back to the objective function $F(n|T)$ and replace f_x with f_y (denoted as f_C in the main text):

$$F(C,n|T)=\prod_g\prod_{j\in J^g} f_x\left(N_j^g;\mu_j^g-\ln C,\sigma_j^g\right)=\prod_g\prod_{j\in J^g}\frac{1}{N_j^g\sigma_j^g\sqrt{2\pi}}e^{-\frac{\left(\ln N_j^g-\mu_j^g+\ln C\right)^2}{2\left(\sigma_j^g\right)^2}}$$

To avoid working with multiplications, we will instead consider the logarithm of F:

$$\begin{aligned}\ln F(C,n|T)&=\sum_g\sum_{j\in J^g}\left[-\frac{\left(\ln N_j^g-\mu_j^g+\ln C\right)^2}{2\left(\sigma_j^g\right)^2}+\ln\left(\frac{1}{N_j^g\sigma_j^g\sqrt{2\pi}}\right)\right]=\\&=\sum_g\sum_{j\in J^g}\left[-\frac{1}{2\left(\sigma_j^g\right)^2}\left(\ln N_j^g-\mu_j^g+\ln C\right)^2-\ln\left(N_j^g\sigma_j^g\sqrt{2\pi}\right)\right]=\\&=\sum_g\sum_{j\in J^g}\left[-\frac{1}{2\left(\sigma_j^g\right)^2}\left(\ln N_j^g-\mu_j^g+\ln C\right)^2-\ln N_j^g-\ln\left(\sigma_j^g\sqrt{2\pi}\right)\right]\end{aligned}$$

If we now drop the constants from $\ln F(C,n|T)$, we get the final objective function:

$$\psi(C,n|T)=\sum_g\sum_{j\in J^g}\left[-\frac{1}{2\left(\sigma_j^g\right)^2}\left(\ln N_j^g-\mu_j^g+\ln C\right)^2-\ln N_j^g\right].$$

We now address an earlier assumption that the scaling factor C is known. To this end we find a C that maximises the objective ψ when all other variables are given (i.e., when the simulated occupancy n and the parameters of the log-normal distributions are available). To this end we take the derivative of ψ with respect to $\ln C$ and equate it to zero:

$$\frac{\partial\psi}{\partial\ln C}=\sum_g\sum_{j\in J^g}\left[-\frac{2}{2\left(\sigma_j^g\right)^2}\left(\ln N_j^g-\mu_j^g+\ln C\right)\right]=\sum_g\sum_{j\in J^g}\left[-\frac{1}{\left(\sigma_j^g\right)^2}\left(\ln N_j^g-\mu_j^g+\ln C\right)\right]=0$$

$$\sum_g \sum_{j \in J^g} \left[-\frac{1}{\left(\sigma_j^g\right)^2} \left(\ln N_j^g - \mu_j^g + \ln C \right) \right] = 0$$

$$\sum_g \sum_{j \in J^g} \frac{1}{\left(\sigma_j^g\right)^2} \left(\ln N_j^g - \mu_j^g \right) + \sum_g \sum_{j \in J^g} \frac{1}{\left(\sigma_j^g\right)^2} \ln C = 0$$

$$\sum_g \sum_{j \in J^g} \frac{1}{\left(\sigma_j^g\right)^2} \left(\ln N_j^g - \mu_j^g \right) = -\sum_g \sum_{j \in J^g} \frac{1}{\left(\sigma_j^g\right)^2} \ln C$$

$$-\sum_g \sum_{j \in J^g} \frac{1}{\left(\sigma_j^g\right)^2} \left(\ln N_j^g - \mu_j^g \right) = \sum_g \sum_{j \in J^g} \frac{1}{\left(\sigma_j^g\right)^2} \ln C$$

$$\sum_g \sum_{j \in J^g} \frac{1}{\left(\sigma_j^g\right)^2} \left(\mu_j^g - \ln N_j^g \right) = \sum_g \sum_{j \in J^g} \frac{1}{\left(\sigma_j^g\right)^2} \ln C$$

$$\frac{\sum_g \sum_{j \in J^g} \frac{1}{\left(\sigma_j^g\right)^2} \left(\mu_j^g - \ln N_j^g \right)}{\sum_g \sum_{j \in J^g} \frac{1}{\left(\sigma_j^g\right)^2}} = \ln C$$

The end result

$$\ln C = \frac{\sum_g \sum_{j \in J^g} \frac{1}{\left(\sigma_j^g\right)^2} \left(\mu_j^g - \ln N_j^g \right)}{\sum_g \sum_{j \in J^g} \frac{1}{\left(\sigma_j^g\right)^2}}$$

also allows an appealing interpretation as the weighted sum of scale differences and matches the intuition about the scaling factor.

INITIATION RATE APPROXIMATION

We propose a method for approximating initiation rates that is based on the observation of Ciandrini *et al.* [22] that the average number of ribosomes bound to a segment (i.e., the segments average ribosome occupancy) as a function of the initiation rate, $A(k_0)$, reaches saturated state either smoothly or abruptly (Figs 4.25A and 4.27). This limited set of steady-state behaviours allows for efficiently approximating $A(k_0)$. For each segment j we approximate the *shape* of this function as

$$f_j(k_0) = \begin{cases} f_j^-(k_0), & k_0 \le e_j \\ f_j^+(k_0), & k_0 > e_j \end{cases}$$

$$f_j^-(k_0) \equiv a_j k_0 / (b_j + k_0) \quad ,$$

$$f_j^+(k_0) \equiv c_j k_0 + d_j$$

where $a_j, b_j, c_j \ge 0$, and d_j are parameters that need to be determined. Here $f_j^-(k_0)$ and $f_j^+(k_0)$ are used to approximate the unsaturated and saturated parts of $A_j(k_0)$

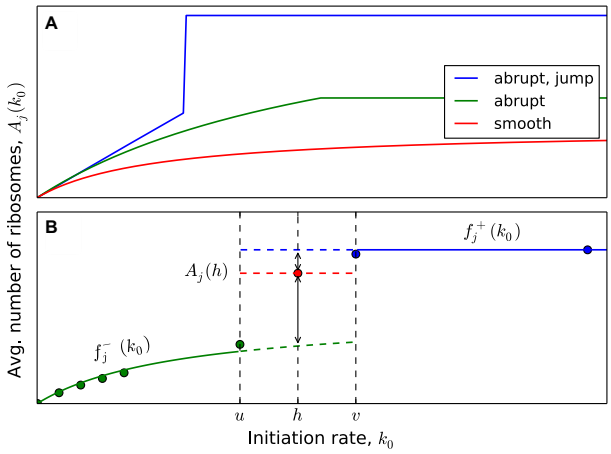

Figure 4.25: $A_j(k_0)$, the average number of ribosomes attached to segment j at steady-state increases as a function of initiation rate k_0 and reaches saturated state (**A**) smoothly, abruptly or abruptly with a "jump"; this observation can be used to efficiently approximate the shape of $A_j(k_0)$. At each step of the bracket search (**B**) the decision whether the gene is saturated at initiation rate h, the midpoint of bracket $[u,v]$, is made based on the squared distance from $A_j(h)$ (red dashed line) to $f_j^-(h)$ (green lines) and $f_j^+(h)$ (blue lines) - the approximation functions fitted into model simulation results (circles); and the approximations are refit in accordance to the decision made.

respectively. Approximation parameters are iteratively updated inside a bracket search (Fig. 4.25B). The approximation of $A_j(k_0)$ is then used to approximate gene initiation rates as discussed below.

In order to approximate initiation rates, we assume that codon elongation rates are given and a "proposed" scaling factor $\tilde{C}$, an estimate of the unknown true scaling factor C, is available. Gene initiation rates k_0 are then chosen to maximise the objective $\psi(C, n|T)$ for $\tilde{C}$ and the approximations $f_j(k_0)$ obtained earlier. In practice the objective evaluated for $f_j(k_0)$ is a unimodal function of the initiation rate (see Fig. 4.26) and ternary search is used to efficiently find the k_0 that maximises it, i.e., the sought initiation rate approximation.

To determine approximation parameter values for segment j, the model is simulated for E low and high initiation rates and f_j^- and f_j^+ are first fit onto points $\left(k_0, A_j(k_0)\right)$ recorded for low and high initiation rates respectively by minimising the summed squared error. $E = 5$ was used as it gives robust estimates in practice and the low and high initiation rates are equally spaced in $\left(0, \min_i k_i/2\right]$ and $[\max_i k_i, 1]$ respectively. Bracket search is then used to find the switch point e_j.
Starting from bracket $[u,v] = [0,1]$, a midpoint $h = (u+v)/2$ is chosen and $A_j(h)$ is found by simulation. $A_j(h)$ is compared to $f_j^-(h)$ and $f_j^+(h)$ to determine whether at

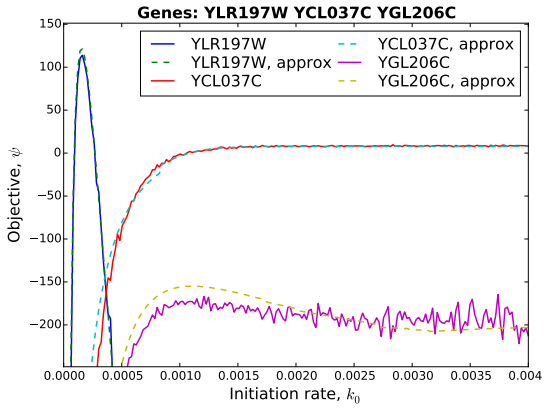

Figure 4.26: Objective function value for genes YLR197W, YCL037C and YGL206C plotted as a function of the initiation rate for proposed scale $\tilde{C} \approx 181.956$ (see next section). The objective calculated using the approximation of the average number of ribosomes (dashed lines) and the objective calculated based on the simulation output (solid lines) demonstrate near-identical behaviour. tAI-based elongation rates were used in simulations.

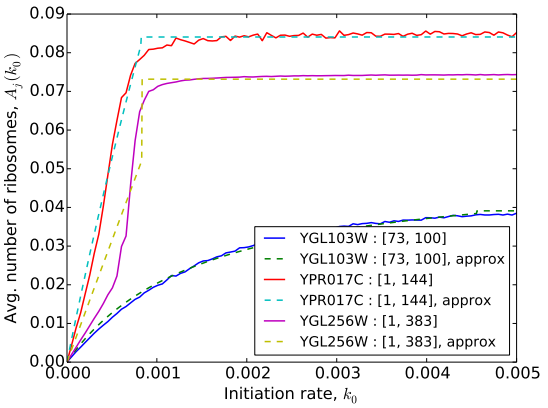

Figure 4.27: Examples of the three types of behaviour of the average number of ribosomes $A_j(k_0)$ as a function the initiation rate k_0 (solid lines) and their approximations (dashed lines): smooth - YGL103W (blue, green); abrupt - YPR017C (red, cyan); abrupt jump - YGL256W (magenta, yellow). tAI-based elongation rates were used in simulations.

initiation rate h the mRNA is already in saturated state, and the bracket is updated as

$$\begin{aligned} u &\leftarrow h, \quad \text{if } \left(f_j^-(h) - A_j(h)\right)^2 < \left(f_j^+(h) - A_j(h)\right)^2 \\ v &\leftarrow h, \quad \text{otherwise.} \end{aligned}$$

Parameters a_j, b_j or c_j, d_j are then refit with the new point $\left(h, A_j(h)\right)$ depending on whether the left or the right edge of the bracket was changed. The process continues

until the bracket length becomes smaller than 10^{-6} and the switch point is then calculated as $e_j = (u+v)/2$.
Approximations $f_j(k_0)$ are obtained for each segment j and are recomputed each time elongation rates change.

THE PROPOSED SCALING FACTOR $\tilde{C}$

The proposed scaling factor $\tilde{C}$ and the scaling factor C used in the objective function ψ are both responsible for matching scales between simulated ribosome occupancy and the measured data. However, unlike the freely changing parameter C, the proposed scale $\tilde{C}$ is fixed and defines which portion of the difference between the simulated and measured densities (N_j^g and μ_j^g respectively) can be attributed to the scale mismatch, and which portion must be explained the TASEP model. Choosing $\tilde{C}$ to be a good estimate of the unknown true scaling factor is required to ensure that the fitted translation rates have biologically meaningful values.

The true scaling factor is impossible to measure exactly, but it is possible to estimate it. It is determined by the amount of DNA available for sequencing, which in turn depends on the number of actively translating ribosomes (for ribo-seq), the total size of the coding transcriptome (for RNA-seq) and the efficiencies of individual steps of the experimental protocol. To estimate this factor we assume that individual steps of the ribosome profiling protocol are either highly efficient (i.e. only a moderate portion of the genetic material and ribosomes are lost during their execution), or that they are equally inefficient for the ribo-seq and RNA-seq measurements, and use the procedures from Siwiak and Zielenkiewicz [9] to estimate the number of actively translating ribosomes P_{active} and the size of the coding transcriptome Q.

The total number of ribosomes $P_{\text{total}} = 2\times10^5$ and the fraction of ribosomes involved in active translation $\rho = 0.85$ [30] were used to calculate $P_{\text{active}} = \rho \cdot P_{\text{total}} = 1.7\times10^5$. The size of the coding transcriptome was computed as

$$Q = 3 \cdot \sum_g S^g K_g,$$

where S^g and K_g are respectively the length in codons of CDSes and the absolute number of transcripts of gene g. Quantities K_g were calculated as the relative mRNA abundance $m_g = \frac{M_{[1,S^g]}}{N_M}$ scaled by the total number of mRNA molecules per cell $E = 36{,}139$ [31], yielding $Q = 3E \cdot \sum_g S^g m_g \approx 3.09\times10^7$.
Using these quantities the proposed scaling factor can be estimated as

$$\tilde{C} = \frac{Q}{P_{\text{active}}} \approx 181.956.$$

The described procedure was used to set the proposed scale $\tilde{C}$ for all analysed datasets individually.

CMA PARAMETER SETTINGS

CMA search space was constrained to $[-12, 12]$. This way, when sigmoid-transformed prior to TASEP simulations, the rates would occupy the interval $(0, 1)$ almost entirely.

When fitting elongation rates, rates consistent with the tRNA pool adaptation hypothesis were used as a starting point and their standard deviation (SD) was used to set CMA parameter $\sigma_0 \approx 0.94$. CMA was run with the default population size $\mu = 16$. To control the runtime of the algorithm, it was stopped if the search ran for at least 300 generations and the overall best solution did not improve over the last 50 generations.

COMPARISON TO OTHER MODELS

Zhang's model [7] was designed to predict the relative local speed of translation at a given position from codon elongation rates around it. Codon elongation times t_i (inverse of elongation rates) consistent with the tRNA pool adaptation hypothesis were used to parameterize it as in Wohlgemuth *et al.* [59].

To obtain per-gene translation time profiles from the Zhang model, individual codon translation times t_i were smoothed with a moving average window of 19 codons as in the original publication. For model evaluation these profiles were treated as codon occupancy probabilities output by other models.

SMoPT (Stochastic Model of Protein Translation [10]) is a full-cell model developed and parameterised for yeast using the RP data [11]. It describes the movement of ribosomes on mRNA transcripts with a TASEP-like process while also taking tRNA and ribosome concentration into account.

To obtain ribosome occupancy profiles, the model was simulated with default settings for the maximum allowed time (2.4×10^6 seconds; Fig. 4.28) and snapshots of the state of the model with exact locations of ribosomes on all transcripts were taken every second. These snapshots were processed into ribosome occupancy estimates by recording how often a ribosome was seen at a particular location and dividing this number by the total simulation time. Observations for different transcripts of the same gene were combined into a single occupancy profile normalised by the number of transcripts. Since SMoPT implicitly assumes that termination is instantaneous, codon occupancies for stop codons were set to zero.

TRANSLATION RATE REPRODUCIBILITY ANALYSIS

In order to determine how robust translation rates fitted on the McManus *et al.* [24] dataset are, we set out to repeat model fitting on an independent dataset. The RP data for yeast *Saccharomyces cerevisiae* strain BY4741 [11] was used for this purpose. It is available as a read mapping against the reference genome sequence of *S. cerevisiae* strain S288C taken from SGD on June 22, 2008 (release R58). These data were re-mapped to release R64 (January 14, 2014) by sequentially updating the alignment coordinates according to the sequence changes file available from SGD. The updated alignments were filtered as in in the original publication. We intentionally did not use a stringent cutoff and kept alignments with up to 2 mismatches, as we expect that a fraction of the mismatches originates from the use of the S288C reference genome for reads of a different strain. After discarding footprint lengths for which location of the A-site could not be reliably determined, reads of lengths $22 \leq l \leq 32$ and $27 \leq l \leq 32$ for mRNA- and ribo-seq reads respectively were used in the analysis. A-site determination (Fig. 4.29 and Table 4.8) and assignment of reads to CDSes (see Table 4.9 for statistics) were performed

Table 4.8: P-site offsets for various footprint lengths for the Ingolia *et al.* [11] dataset determined individually for each replicate based on Fig. 4.29.

Length	Replicate 1		Replicate 2		Final offset
	Start	Stop	Start	Stop	
27	12	12	13	12	12
28	12	12	12	12	12
29	13	12	13	12	13
30	13	13	13	13	13
31	13	13	13	13	13
32	12	12	12	NA	12

Table 4.9: Ingolia dataset [11] read alignment statistics for the *S. cerevisiae* genome release used in the original publication (R58) and the genome release used in this study (R64), and varying alignment strictness. The switch from R58 to R64 gives an increase in the number of reads with up to 2 mismatches.

Type	Release	Name	Max subst.	Min length	Max length	Unique	Replicate 1		Replicate 2	
							mRNA	Ribosome	mRNA	Ribosome
Ref.	R58	Raw	2	20	35	No	2,219,009	1,868,620	4,895,436	3,424,182
	R64	Raw	2	20	35	No	2,226,949	1,870,581	4,898,568	3,439,420
Ref.	R64	Filtered	2	22/27	32	No	1,924,200	1,175,283	4,583,654	1,699,871
		Strict	1	22/27	32	No	1,825,219	1,109,120	3,643,440	1,281,387
		Unique	1	22/27	32	Yes	1,618,952	926,029	3,216,498	1,062,049
CDS	R64	Filtered	2	22/27	32	No	1,365,482	1,128,724	3,451,402	1,649,033
		Strict	1	22/27	32	No	1,297,286	1,069,840	2,729,730	1,245,578
		Unique	1	22/27	32	Yes	1,125,445	897,963	2,369,133	1,034,200

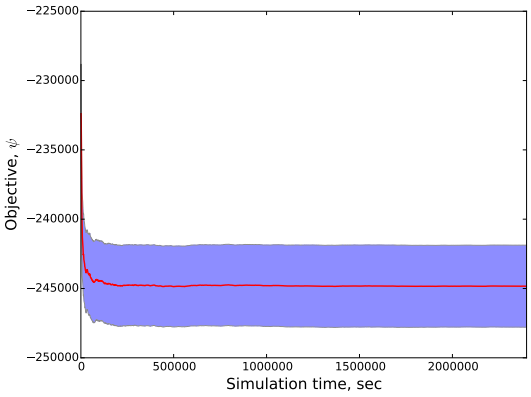

Figure 4.28: CV objective ψ for the SMoPT simulation output (red line) and its standard deviation (blue area) plotted as the function of the simulation time demonstrates that reliable density estimates were determined by the end of the simulation period.

as before.

MA-plots for the full length genes (Fig. 4.30) were similarly used to set read count thresholds for reliable density estimates to 128 and 64 mRNA- and ribo-seq reads respectively. Segment trees constructed with these thresholds were then bias-corrected (Figs 4.31 and 4.32) and ratio measurement errors (Table 4.10) were determined as before.

The set of 2,949 genes common between the McManus *et al.* [24] and Ingolia *et al.* [11] datasets with 13,443 Ingolia and 51,223 McManus segments was then used to independently fit two TASEP$^{\text{elong}}$ models on the two datasets inside a common 5-fold CV loop.

MODEL FITTING IS ROBUST AGAINST EXPERIMENTAL BIASES

Translation rates obtained from the two datasets (Fig. 4.33) and initiation rates show qualitatively similar behaviour (Fig. 4.34) despite significant differences in protocols, strains, sequencing depth and computational processing between the two datasets. Ingolia *et al.* [11] describe the first application of the ribosome profiling method. They used an RP protocol based on poly-A tailing of ribosomal footprints, which was later substituted by ligation of an adapter sequence to the 3′ end (e.g., McManus *et al.* [24]). The latter is a standard procedure in small RNA sequencing, as it simplifies the experimental protocol and subsequent short read mapping. This difference in protocols results in substantially different biases due to sequence preferences of poly-A polymerase and RNA ligase [60], which could explain the discrepancies between the elongation rates fitted on the two datasets. Relatively low sequencing depth of the Ingolia *et al.* dataset is also a likely factor contributing to higher SDs of the fitted elongation rates and moderate reproducibility. Nevertheless, the general agreement between the two sets of translation rates shows that our approach is robust against experimental biases.

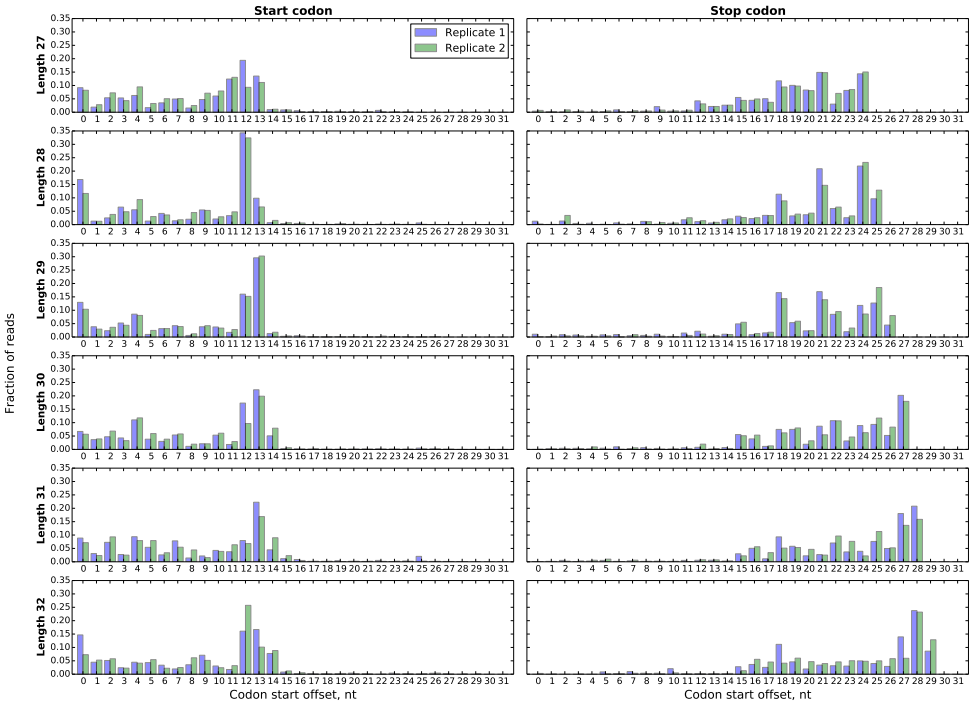

Figure 4.29: Histograms of the locations of the start and stop codons within ribosomal footprints of various lengths for the Ingolia *et al.* [11] dataset. Offsets fixed for every footprint length are given in Table 4.8.

Table 4.10: Shape parameters of the Ingolia *et al.* [11] dataset density ratio distributions for segments grouped by length. Values of σ determined for this dataset are considerably higher than the values obtained for McManus data.

#	Left	Right	Group size	Shape parameter σ, $\log_2$
1	20	31	1304	0.320130113060900
2	31	47	1400	0.292439449293707
3	47	67	1359	0.278283672264551
4	67	94	1366	0.261941203773427
5	94	130	1399	0.245270164406985
6	130	176	1359	0.239063834104316
7	176	246	1353	0.225808138327335
8	246	354	1375	0.214356415855909
9	354	560	1374	0.206160115988558
10	560	4912	1367	0.199898527308120

MODEL FITTING IS ROBUST TO CHANGES IN THE GENES USED FOR FITTING

We also sought to compare the McManus translation rates obtained in the previous section to the rates we determined on the set of genes common between the McManus dataset and the SMoPT model. Translation rates for these two sets of genes quantitatively agree with each other (Fig. 4.35). This suggests that the rate fitting procedure is robust to

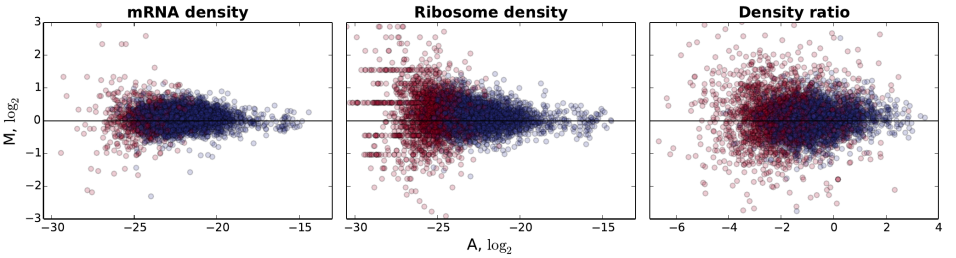

Figure 4.30: MA-plot of the full-CDS density estimates computed for the Ingolia *et al.* [11] dataset. Unreliable density estimates (< 128 total RNA-seq reads or < 64 ribo-seq reads; coloured red) show increased measurement variance.

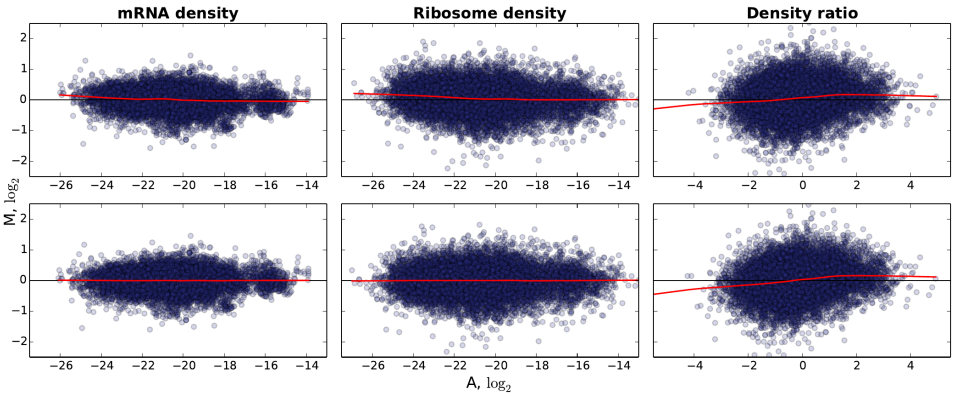

Figure 4.31: MA-plots of the segment tree density and ratio measurements for the Ingolia *et al.* [11] dataset (**top**) before bias correction show a density-dependent systematic bias, which (**bottom**) after removal of mRNA and ribosome density bias is no longer present in density estimates. An identical procedure is applied to correct density-dependent amplified in ratio measurements (see Fig. 4.32).

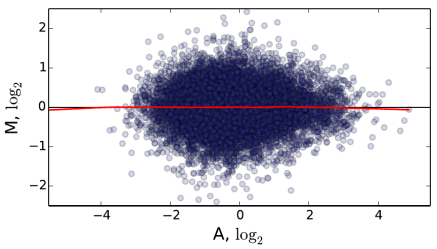

Figure 4.32: MA-plot of the segment tree density ratio estimates of the Ingolia *et al.* [11] dataset after normalisation shows that no significant bias is present at the extreme density ratio values.

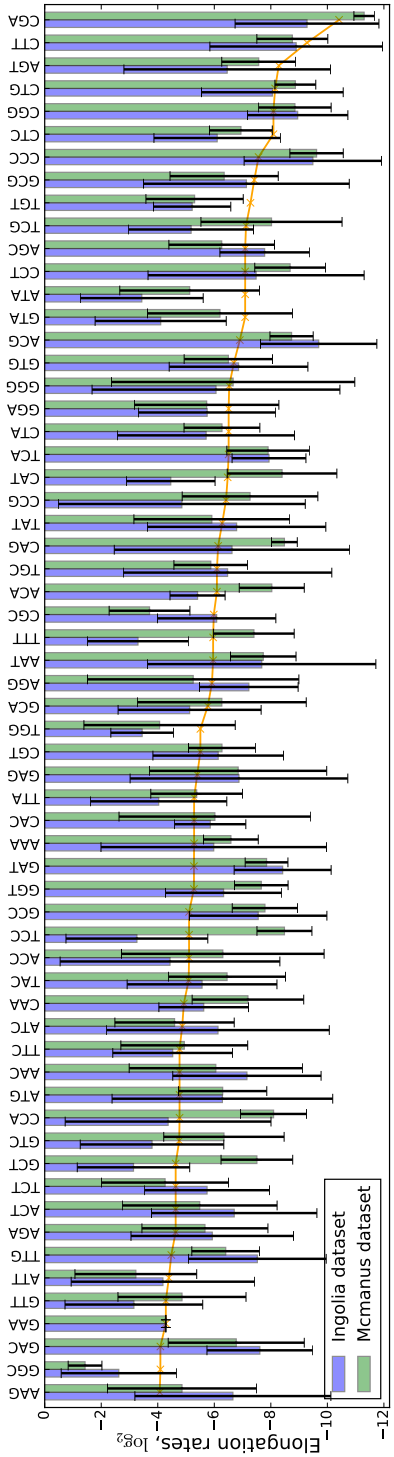

Figure 4.33: Mean and SD of the codon elongation rates fitted in different CV folds of McManus and Ingolia datasets, compared to the tAI-based rates (orange line). Many codons show agreement between the elongation rates fitted on the Ingolia dataset (blue) and the elongation rates fitted on the McManus dataset (green).

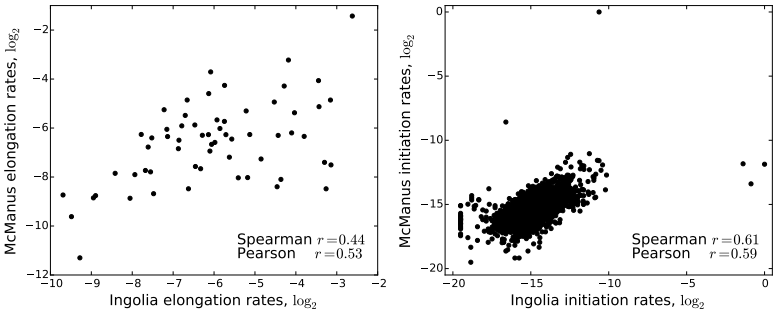

Figure 4.34: Translation elongation rates (**left**) and translation initiation rates (**right**) fitted independently on the McManus and Ingolia datasets also show qualitatively similar behaviour (Pearson $r = 0.531$, $p < 10^{-4}$ for elongation rates; and $r = 0.592$, $p < 10^{-277}$ for initiation rates).

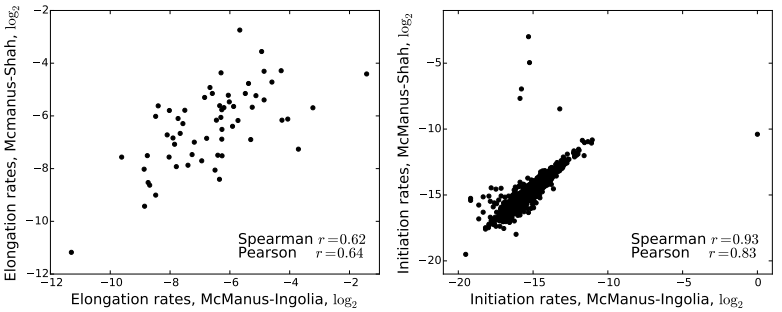

Figure 4.35: Translation elongation rates (**left**) and translation initiation rates (**right**) fitted independently on different sets of genes from McManus datasets show similar behaviour (Pearson $r = 0.644$, $p < 10^{-7}$ for elongation rates; and $r = 0.830$, $p < 10^{-293}$ for initiation rates). As expected, elongation rates fitted on different sets of genes from the McManus dataset agree well with each other.

the gene set used to obtain the rates. However, the agreement between the rates is not perfect. This could be a result of overfitting the rates in individual CV folds due to using only a single fold for training - a limitation dictated by the computational complexity of the fitting procedure.

Alternatively the non-perfect agreement may also be a consequence of the implicit assumption, that codon elongation rates are independent of codon context, being incorrect. All translation models proposed to date, including ours, assume that translation elongation rates are constant and are not influenced by the sequence around a particular codon, although various factors affecting the speed of elongation have been suggested [4]. Differences between elongation rates induced by local sequences that are over-represented in a particular gene set provide an alternative explanation for the limited agreement between translation rates fitted on different sets.

Table 4.11: Correlations of TASEP$^{\text{init}}$ predictions with independent PA datasets for the full-CDS model. Spearman rank correlation coefficients r for are reported; J' is the partial correlation between J and PA given mRNA.

	TASEP$^{\text{init}}$ full-CDS		
	Newman YEPD	Newman SD	Ghaemmaghami
Init. rate	$r = 0.56^{***}$	$r = 0.55^{***}$	$r = 0.48^{***}$
J	$r = 0.57^{***}$	$r = 0.55^{***}$	$r = 0.48^{***}$
$J \times$ mRNA	$r = 0.71^{***}$	$r = 0.69^{***}$	$r = 0.61^{***}$
J'	$r = 0.50^{***}$	$r = 0.47^{***}$	$r = 0.35^{***}$

* - $p-$value $< 10^{-5}$ ** - $p-$value $< 10^{-20}$ *** - $p-$value $< 10^{-100}$

MODEL FITTING WITHOUT SEGMENT TREES

In order to assess the effect of gene segmentation on the fitted translation rates we sought to compare our model to one fitted without the use of segment trees. To this end we restricted our segment trees to a maximal depth of 1 (i.e., they were allowed to contain only the top-level segment, corresponding to the entire CDS). The 4,768 full-CDS segments obtained in such way were bias-corrected as described in Section 4.A (results are shown in Figs 4.36 and 4.37), and a single shape parameter $\sigma \approx 0.17$ ($\log_2$ scale) was estimated for all full-CDS segments. The set of full-CDS segments was then used to fit the TASEP$^{\text{init}}$ model and compare its predictions to the independent PA datasets and to the TASEP$^{\text{init}}$ model fitted with the use of segment trees.

It can be seen from Fig. 4.38 that (i) predictions of the full-CDS model compare favourably to the predictions made by the original TASEP$^{\text{init}}$ model; and that (ii) the predicted ribosome occupancy of the full-CDS model and the measured per-transcript density are also in agreement. A cloud of outlier points that can be clearly seen in Fig. 4.38 (right) consists of genes with low fitted initiation rates. This suggests that the initiation rate approximation procedure used for segment trees may not be very suitable for the case when only the full-CDS segments are used. Table 4.11 shows that the correlation between the predictions made by the full-CDS TASEP$^{\text{init}}$ model and independent PA datasets are lower, but comparable to the correlations obtain by that model fitted on segment trees. Together these findings suggest that the use of segment trees, when compared to the traditional full-CDS approach, does not introduce any significant biases into the fitted rates, but instead makes the initiation rate approximation procedure more accurate.

We also used the set of full-CDS segments to fit elongation and initiation rates of the TASEP$^{\text{elong}}$ model as described before. The best fits (note that the CMA evolutionary strategy will at best find one of the many equally good solutions if the problem is underdetermined) from each of the folds were used to calculate the CV mean rate and its SD for each of the 61 codons (shown in Fig. 4.39). It can be seen from Fig. 4.39 that fitted elongation rates differ considerably between CV folds, suggesting that the full-CDS approach does not provide sufficient constraints for simultaneously determining translation elongation and translation initiation rates of the full-CDS TASEP$^{\text{elong}}$ model.

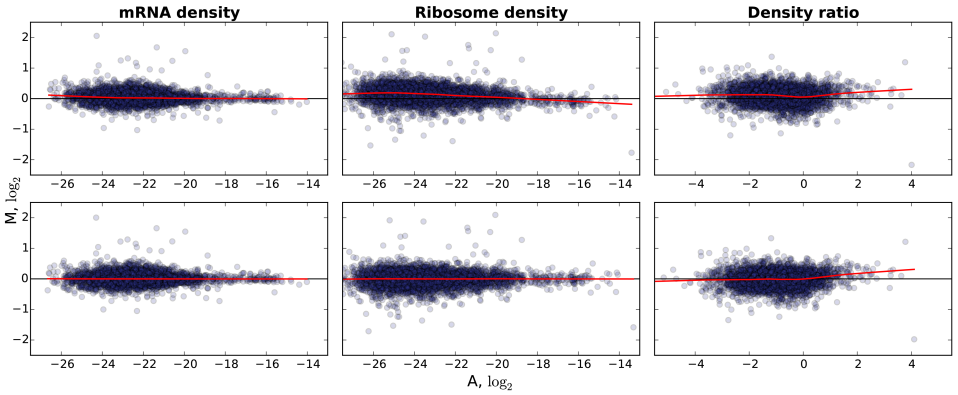

Figure 4.36: MA-plots of the full-CDS density and ratio measurements for the McManus *et al.* [24] dataset (**top**) before bias correction show a density-dependent systematic bias, which (**bottom**) after removal of mRNA and ribosome density bias is no longer present in density estimates. An identical procedure is applied to correct the density-dependent bias amplified in ratio measurements (see Fig. 4.32).

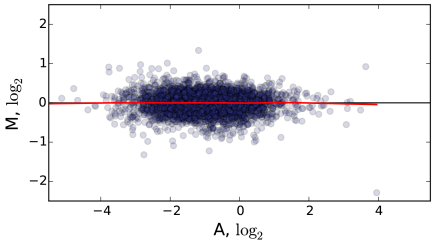

Figure 4.37: MA-plot of the full-CDS density ratio estimates of the McManus *et al.* [24] dataset after normalisation shows that no significant bias is present.

FUNCTIONAL ENRICHMENT

Gene ontology functional enrichment analysis was performed using the DAVID tool [61] with functional categories `GOTERM_BP_FAT`, `GOTERM_CC_FAT` and `GOTERM_MF_FAT`. A score cutoff of 0.1 and a size cutoff of 2 were used in the analysis. Only enrichments significant at 0.05 FDR were reported.

SEGMENT TREE RECONSTRUCTION

To visualise the change of density along transcripts and the uncertainty about it captured by the segment trees, we sought to obtain a reconstruction of the per-transcript ribosome density of the tree, which could be directly plotted. Since every node within the segment tree defines a probability distribution (PD) of the average density ratio of the corresponding segment, together these segments define a joint probability distribution of segment-averaged (i.e., piecewise constant) density of the entire gene. Samples from this joint PD can be used to reconstruct the encoded density and to obtain confidence

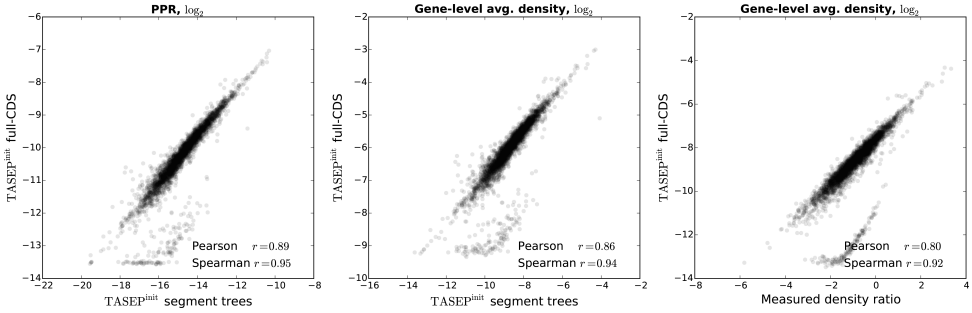

Figure 4.38: PPR (**left**) and gene-level average density (**middle**) predicted by the TASEP$^{\text{init}}$ model fitted on full-CDS segments and the TASEP$^{\text{init}}$ fitted on segment trees agree well with each other. Similarly, gene-level average density predicted by TASEP$^{\text{init}}$ fitted on full-CDS estimates agrees well with density ratios obtained from RP data (**right**).

bounds on the reconstruction. We note that the logarithm of the PDF of this distribution has the same form as the objective function ψ evaluated for the same gene.

Formally we assign a random variable $x_{[l_j,r_j]}$, describing the average density of the corresponding segment to every leaf segment $[l_j, r_j]$ in the tree, and a random variable $x_{[l_k r_k]}$ with the same meaning to every segment $[l_k, r_k]$ that needs to be added to the tree in order for each parent node to have exactly two children (see Fig. 4.40). These variables are used to compute the average density at every non-leaf segment as the weighted mean of values $x_{[l_j,r_j]}$ falling within that segment with segment lengths used as weights. We then assume a wide uniform prior for variables x and use Markov chain Monte Carlo [62] to sample them from the joint PD.

When building reconstructions for visualisation we obtained 2×10^8 samples with a burn phase of 10^8 samples and thinning of 100, yielding a total of 10^6 samples. The 10%, 50% and 90% highest posterior density (HPD) intervals calculated from this sample were then used to plot the reconstruction.

4.B. SUPPLEMENTARY DATA

Supplementary data is available from *PLoS Computational Biology* online[1].

REFERENCES

[1] A. A. Gritsenko, M. Hulsman, M. J. Reinders, and D. de Ridder, *Unbiased quantitative models of protein translation derived from ribosome profiling data,* PLoS Computational Biology **11**, e1004336 (2015).

[2] T. M. Schmeing and V. Ramakrishnan, *What recent ribosome structures have revealed about the mechanism of translation,* Nature **461**, 1234 (2009).

[3] G. Cannarozzi, N. N. Schraudolph, M. Faty, P. von Rohr, M. T. Friberg, A. C. Roth, P. Gonnet, G. Gonnet, and Y. Barral, *A role for codon order in translation dynamics,* Cell **141**, 355 (2010).

[4] T. Tuller, I. Veksler-Lublinsky, N. Gazit, M. Kupiec, E. Ruppin, and M. Ziv-Ukelson, *Composite effects of gene determinants on the translation speed and density of ribosomes,* Genome Biology **12**, R110 (2011).

[5] C. A. Charneski and L. D. Hurst, *Positively charged residues are the major determinants of ribosomal velocity,* PLoS Biology **11** (2013).

[6] J. D. Keasling, *Manufacturing molecules through metabolic engineering,* Science **330**, 1355 (2010).

[7] G. Zhang and Z. Ignatova, *Generic algorithm to predict the speed of translational elongation: implications for protein biogenesis,* PLoS One **4**, e5036 (2009).

[8] S. Reuveni, I. Meilijson, M. Kupiec, E. Ruppin,

[1] http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004336

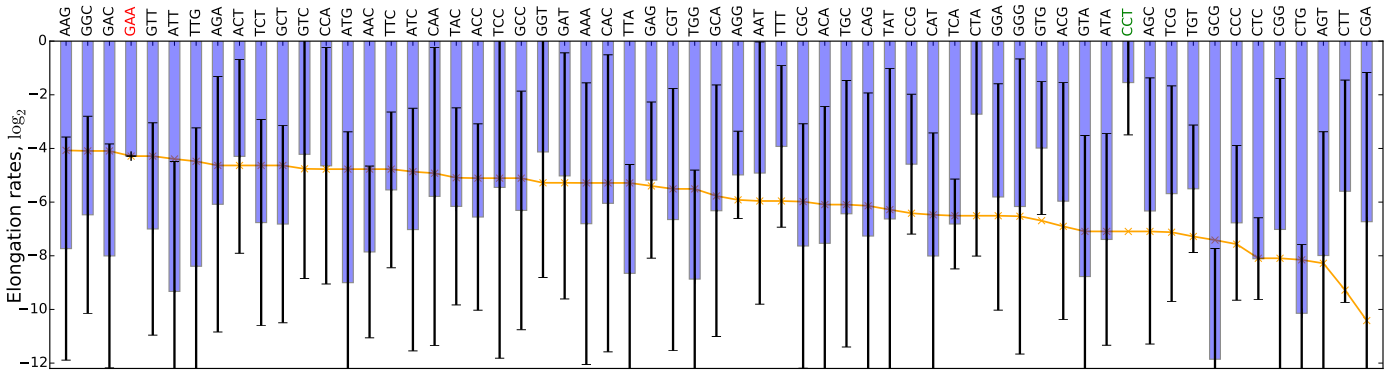

Figure 4.39: Mean and SD of the codon elongation rates fitted in different CV folds of the McManus dataset using the full-CDS approach, compared to the tAI-based rates (orange line). For most codons fitted elongation rates show large variation between CV folds.

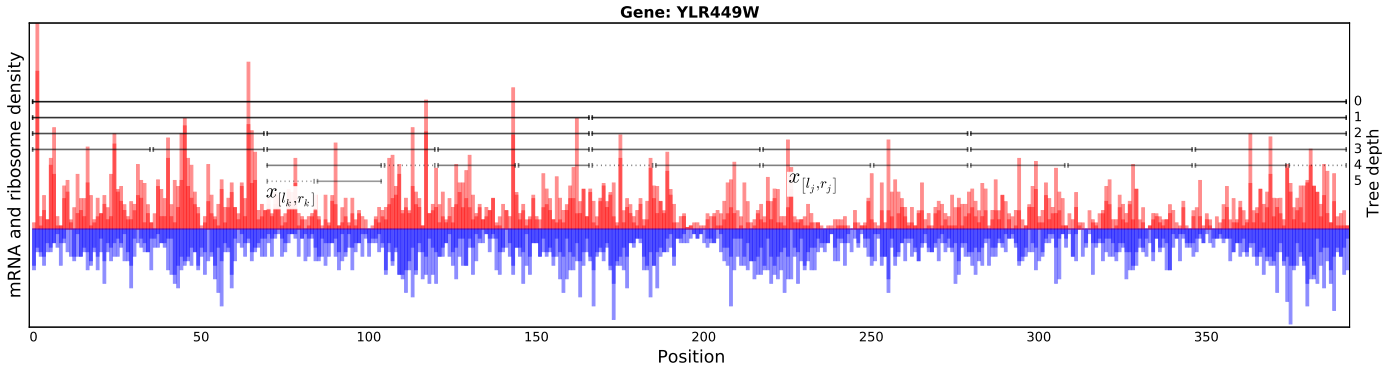

Figure 4.40: Segment tree shown for gene YLR449W visualised as in the main text. Segments from the segment tree (solid lines) are assigned variables $x_{[l_j,r_j]}$, whereas segments that need to be added to the tree (dotted lines) are assigned variables $x_{[l_k,r_k]}$.

and T. Tuller, *Genome-scale analysis of translation elongation with a ribosome flow model,* PLoS Computational Biology **7**, e1002127 (2011).

[9] M. Siwiak and P. Zielenkiewicz, *A comprehensive, quantitative, and genome-wide model of translation,* PLoS Computational Biology **6**, e1000865 (2010).

[10] P. Shah, Y. Ding, M. Niemczyk, G. Kudla, and J. B. Plotkin, *Rate-limiting steps in yeast protein translation,* Cell **153**, 1589 (2013).

[11] N. T. Ingolia, S. Ghaemmaghami, J. R. Newman, and J. S. Weissman, *Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling,* Science **324**, 218 (2009).

[12] N. T. Ingolia, *Ribosome profiling: new views of translation, from single codons to genome scale,* Nature Reviews Genetics **15**, 205 (2014).

[13] N. T. Ingolia, L. F. Lareau, and J. S. Weissman, *Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes,* Cell **147**, 789 (2011).

[14] G. A. Brar, M. Yassour, N. Friedman, A. Regev, N. T. Ingolia, and J. S. Weissman, *High-resolution view of the yeast meiotic program revealed by ribosome profiling,* Science **335**, 552 (2012).

[15] M. V. Gerashchenko, A. V. Lobanov, and V. N. Gladyshev, *Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress,* Proceedings of the National Academy of Sciences **109**, 17394 (2012).

[16] C. G. Artieri and H. B. Fraser, *Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation,* Genome Research **24**, 2011 (2014).

[17] A. Dana and T. Tuller, *Determinants of translation elongation speed and ribosomal profiling biases in mouse embryonic stem cells,* PLoS Computational Biology **8**, e1002755 (2012).

[18] W. Qian, J. Yang, N. Pearson, C. Maclean, and J. Zhang, *Balanced codon usage optimizes eukaryotic translational efficiency,* PLoS Genetics **8**, e1002603 (2012).

[19] B. Zinshteyn and W. V. Gilbert, *Loss of a conserved tRNA anticodon modification perturbs cellular signaling,* PLoS Genetics **9**, e1003675 (2013).

[20] J. Gardin, R. Yeasmin, A. Yurovsky, Y. Cai, S. Skiena, and B. Futcher, *Measurement of average decoding rates of the 61 sense codons in vivo,* eLife **3**, e03735 (2014).

[21] L. F. Lareau, D. H. Hite, G. J. Hogan, and P. O. Brown, *Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments,* eLife **3**, e01257 (2014).

[22] L. Ciandrini, I. Stansfield, and M. C. Romano, *Ribosome traffic on mRNAs maps to gene ontology: genome-wide quantification of translation initiation rates and polysome size regulation,* PLoS Computational Biology **9**, e1002866 (2013).

[23] C. T. MacDonald, J. H. Gibbs, and A. C. Pipkin, *Kinetics of biopolymerization on nucleic acid templates,* Biopolymers **6**, 1 (1968).

[24] C. J. McManus, G. E. May, P. Spealman, and A. Shteyman, *Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast,* Genome Research **24**, 422 (2014).

[25] J. M. Cherry, E. L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, E. T. Chan, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. R. Engel, *et al., Saccharomyces Genome Database: the genomics resource of budding yeast,* Nucleic Acids Research , gkr1029 (2011).

[26] Y. Shibata, G. K. Voeltz, and T. A. Rapoport, *Rough sheets and smooth tubules,* Cell **126**, 435 (2006).

[27] J. Racle, F. Picard, L. Girbal, M. Cocaign-Bousquet, and V. Hatzimanikatis, *A genome-scale integration and analysis of Lactococcus lactis translation data,* PLoS Computational Biology **9**, e1003240 (2013).

[28] L. B. Shaw, J. P. Sethna, and K. H. Lee, *Mean-field approaches to the totally asymmetric exclusion process with quenched disorder and large particles,* Physical Review E **70**, 021901 (2004).

[29] D. T. Gillespie, *Exact stochastic simulation of coupled chemical reactions,* The Journal of Physical Chemistry **81**, 2340 (1977).

[30] D. Zenklusen, D. R. Larson, and R. H. Singer, *Single-RNA counting reveals alternative modes of gene expression in yeast,* Nature Structural & Molecular Biology **15**, 1263 (2008).

[31] F. Miura, N. Kawaguchi, M. Yoshida, C. Uematsu, K. Kito, Y. Sakaki, and T. Ito, *Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs,* BMC Genomics **9**, 574 (2008).

[32] N. Hansen, S. D. Müller, and P. Koumoutsakos, *Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES),* Evolutionary Computation **11**, 1 (2003).

[33] M. dos Reis, R. Savva, and L. Wernisch, *Solving the riddle of codon usage preferences: a test for translational selection,* Nucleic Acids Research **32**, 5036 (2004).

[34] H. Zur and T. Tuller, *RFMapp: ribosome flow model application,* Bioinformatics **28**, 1663 (2012).

[35] P. M. Sharp and W.-H. Li, *The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications,* Nucleic Acids Research **15**, 1281 (1987).

[36] J. R. S. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi, and J. S. Weissman, *Single-cell proteomic analysis of* S. cerevisiae *reveals the architecture of biological noise,* Nature **441**, 840 (2006).

[37] S. Ghaemmaghami, W.-K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman, *Global analysis of protein expression in yeast,* Nature **425**, 737 (2003).

[38] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder, *The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing,* Science **320**, 1344 (2008).

[39] M. Yassour, T. Kaplan, H. Fraser, J. Levin, J. Pfiffner, X. Adiconis, G. Schroth, S. Luo, I. Khrebtukova, A. Gnirke, C. Nusbaum, D.-A. Thompson, N. Friedman, and A. Regev, *Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing,* Proceedings of the National Academy of Sciences **106**, 3264 (2009).

[40] C. Gustafsson, S. Govindarajan, and J. Minshull, *Codon bias and heterologous protein expression,* Trends in Biotechnology **22**, 346 (2004).

[41] M. Welch, S. Govindarajan, J. E. Ness, A. Villalobos, A. Gurney, J. Minshull, and C. Gustafsson, *Design parameters to control synthetic gene expression in Escherichia coli,* PloS One **4**, e7002 (2009).

[42] A. M. Lanza, K. A. Curran, L. G. Rey, and H. S. Alper, *A condition-specific codon optimization approach for improved heterologous gene expression in Saccharomyces cerevisiae,* BMC Systems Biology **8**, 33 (2014).

[43] J. M. Leavitt and H. S. Alper, *Advances and current limitations in transcript-level control of gene expression,* Current Opinion in Biotechnology **34**, 98 (2015).

[44] J. B. Plotkin and G. Kudla, *Synonymous but not the same: the causes and consequences of codon bias,* Nature Reviews Genetics **12**, 32 (2010).

[45] H. Gingold and Y. Pilpel, *Determinants of translation efficiency and accuracy,* Molecular Systems Biology **7** (2011).

[46] C. Kraft, M. Peter, and K. Hofmann, *Selective autophagy: ubiquitin-mediated recognition and beyond,* Nature Cell Biology **12**, 836 (2010).

[47] E. M. Gustilo, F. A. Vendeix, and P. F. Agris, *tRNA's modifications bring order to gene expression,* Current Opinion in Microbiology **11**, 134 (2008).

[48] P. F. Agris, *Decoding the genome: a modified view,* Nucleic Acids Research **32**, 223 (2004).

[49] Z. Bloom-Ackermann, S. Navon, H. Gingold, R. Towers, Y. Pilpel, and O. Dahan, *A comprehensive tRNA deletion library unravels the genetic architecture of the tRNA pool,* PLoS Genetics **10**, e1004084 (2014).

[50] H. M. Salis, E. A. Mirsky, and C. A. Voigt, *Automated design of synthetic ribosome binding sites to control protein expression,* Nature Biotechnology **27**, 946 (2009).

[51] S. Dvir, L. Velten, E. Sharon, D. Zeevi, L. B. Carey, A. Weinberger, and E. Segal, *Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast,* Proceedings of the National Academy of Sciences **110**, E2792 (2013).

[52] G.-W. Li, E. Oh, and J. S. Weissman, *The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria,* Nature **484**, 538 (2012).

[53] J. C. Guimaraes, M. Rocha, and A. P. Arkin, *Transcript level and sequence determinants of protein abundance and noise in Escherichia coli,* Nucleic Acids Research **42**, 4791 (2014).

[54] A. Dana and T. Tuller, *The effect of tRNA levels on decoding times of mRNA codons,* Nucleic Acids Research **42**, 9171 (2014).

[55] B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, *et al.*, *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome,* Genome Biology **10**, R25 (2009).

[56] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed, *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation,* Nucleic Acids Research **30**, e15 (2002).

[57] W. S. Cleveland, *Robust locally weighted regression and smoothing scatterplots,* Journal of the American Statistical Association **74**, 829 (1979).

[58] R. D. Yates and D. J. Goodman, *Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers*, 2nd ed. (John Wiley & Sons, 2005).

[59] S. E. Wohlgemuth, T. E. Gorochowski, and J. A. Roubos, *Translational sensitivity of the* Escherichia coli *genome to fluctuating tRNA availability,* Nucleic Acids Research **41**, 8021 (2013).

[60] C. A. Raabe, T.-H. Tang, J. Brosius, and T. S. Rozhdestvensky, *Biases in small RNA deep sequencing data,* Nucleic Acids Research **42**, 1414 (2014).

[61] D. W. Huang, B. T. Sherman, and R. A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists,* Nucleic Acids Research **37**, 1 (2009).

[62] A. Patil, D. Huard, and C. J. Fonnesbeck, *PyMC: Bayesian stochastic modelling in Python,* Journal of Statistical Software **35**, 1 (2010).

5

Deciphering sequence features of Internal Ribosome Entry Sites

Alexey A. Gritsenko*, Shira Weingarten-Gabbay*, Shani Elias-Kirma, Ronit Nir, Dick de Ridder and Eran Segal

*These authors contributed equally.

Submitted for publication.

ABSTRACT

5

Translation of mRNAs through Internal Ribosome Entry Sites (IRESs) has emerged as a prominent mechanism of cellular and viral initiation. It supports cap-independent translation of select cellular genes under normal conditions, and in conditions when cap-dependent translation is inhibited. IRES structure and sequence are believed to be involved in this process. However due to the small number of IRESs known, there have been no systematic investigations of the determinants of IRES activity With the recent discovery of thousands of novel IRESs in human and viruses [1], the next challenge is to decipher the sequence determinants of IRES activity.

We present the first in-depth computational analysis of a large body of IRESs, exploring RNA sequence features predictive of IRES activity. We identified predictive *k*-mer features resembling IRES *trans*-acting factor (ITAF) binding motifs across human and viral IRESs, and found that their effect on expression depends on their sequence, number and position. Our results also suggest that the architecture of retroviral IRESs differs from that of other viruses, presumably due to their exposure to the nuclear environment. Finally, we measured IRES activity of synthetically designed sequences to confirm our prediction of increasing activity as a function of the number of short IRES elements.

5.1. INTRODUCTION

Translation of mRNA into protein is an essential step in the process of gene expression. Eukaryotic translation begins with the formation of the pre-initiation complex after the delivery of the Met-tRNA$_i^{Met}$ initiator tRNA to the P-site of the 40S ribosomal subunit by the eukaryotic initiation factor eIF2. The pre-initiation complex is then recruited to the 5′ untranslated region (5′-UTR) of the mRNA via the interaction between the 5′ m^{7}GpppN cap structure, the poly-A tail of the mRNA, the poly-A binding protein (PABP) and additional initiation factors (eIF3 and eIF4) and begins scanning the 5′ UTR for the start `AUG`. Once the `AUG` is found in a favourable context, the 60S ribosomal subunit is assembled on the mRNA to begin protein synthesis [2, 3]. This translation initiation route accounts for more that 95% of cellular mRNAs [4], however, in a growing number cases alternative strategies are employed to initiate translation [5, 6]. One such strategy relies on the Internal Ribosome Entry Site (IRES) element, a *cis*-regulatory mRNA element that can attract the ribosome in a cap-independent manner. IRESs were first described as elements driving translation in poliovirus RNAs that do not possess the 5′ cap structure [7]. But IRESs were since discovered in other viruses, including HCV and HIV [8–10], in cellular genes such as p53 [11], XIAP [12] and Bcl-2 [13]. They were also shown to support the ongoing protein synthesis under conditions in which cap-dependent translation is inhibited, such as mitosis or cellular stress. The latter commonly occurs during viral infections, cancer and other human diseases [14–16]. Emerging evidence also suggests that in addition to this "back-up" mechanism, cellular IRESs also play important roles under conditions in which cap-dependent translation is intact: they facilitate the translation of different proteins from cellular bicistronic transcripts [17]; guide ribosomes to produce N-truncated isoforms from alternative downstream `AUG` codons [18–20]; and enable translation of transcripts with locally inhibited cap-dependent translation [21].

Despite this accumulating evidence of relevance of IRES elements to numerous diseases and cellular processes, compared to cap-dependent translation, relatively little is known about mechanisms of IRES-mediated translation. However, it is believed that a combination of primary sequence and RNA structure is functionally important for IRES activity [14, 22–24], which is achieved either via direct recruitment of the ribosome by the structured RNA, or through mediation by a combination of canonical initiation factors and additional IRES *trans*-acting factors (ITAFs; [24–26]). Precisely how ITAFs regulate IRES translation is not fully understood, but they are thought to function either as RNA chaperons, i.e. RNA-binding proteins (RBPs) that alter or stabilise RNA secondary structure in order to allow for ribosome binding, or as adaptor proteins interacting with the ribosome and other initiation factors [27]. Over a dozen proteins have been suggested to function as ITAFs [8, 25], but only few have been studied extensively. Among them, the PTB (polypyrimidine tract-binding protein) and PCBP (poly-C binding protein) RNA chaperon ITAFs were shown to remodel RNA structures of cellular IRESs [28, 29] for interactions with the 40S ribosomal subunit, and were proposed to have a similar role in viral IRESs [30, 31]. Whereas the hnRNP (heterologous nuclear nucleoproteins) C1/C2, the La autoantigen and Unr were implicated in modulating activity of multiple IRESs, but not in RNA structure remodelling [25].

Systematic methods to investigate mRNA translation have lagged behind the field

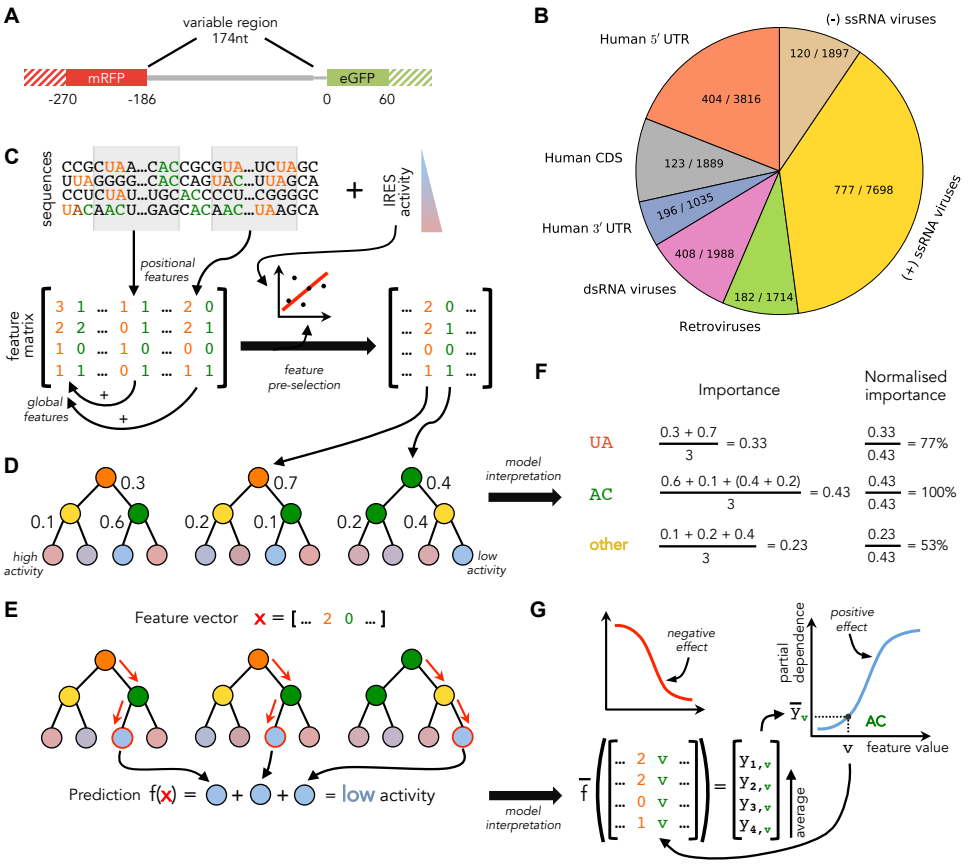

Figure 5.1: Overview of the available data and our analysis approach. (**A**) Schematic representation of the bicistronic reporter construct used in [1] with eGFP (green) expression used to measure IRES activity of variable sequences (grey), and constitutively expressed mRFP used to control for unique genomic integration. To capture context effects, in our analyses the assayed variable sequences (thick grey) were extended to include flanking regions (solid filling). (**B**) The available sequences can be divided into 7 groups based on their origin species and location within transcripts. Number of active sequences, i.e. sequences with IRES activity above background levels, and the total number of RNA sequences are shown for each class. (**C**) Sequences from each of the groups are represented as vectors of sequence k-mer features (UA - orange, AC - green), which are recorded globally and in windows (grey shading). From this large set of features, those unlikely to be predictive are removed based on their weak correlation with IRES activity. Surviving features are used to construct a reduced feature matrix. (**D**) The reduced feature matrix is used for Random Forest training. Each RF tree consists of decision nodes (coloured according to the variables selected by those nodes during training) and leaf nodes that predict IRES activity (coloured according to their prediction). RF trees are constructed by iteratively selecting for each node a variable and split that yield the highest reduction in weighted variance in the nodes children; normalised variance reduction is shown for every node as a number. (**E**) Trained RFs are used to make IRES activity predictions for feature vectors x of unseen sequences by following each tree to the leaf node corresponding to x (path and leaves marked in red), and accumulating leaf node predictions to obtain the overall RF prediction $f(x)$. (**F**) To select features that are most predictive of IRES activity, variance reduction values from (D) are accumulated per tree and averaged across trees to obtain *feature importance*. Normalised importance is also calculated for use in model interpretation. (**G**) To understand the effect of a feature (e.g. the AC k-mer), for each of its possible values v the expected prediction $\bar{y}_v$ is plotted (blue curve). The resulting curve allows for characterising v either as having a positive (increasing curve, blue), or a negative (decreasing curve, red) effect on IRES activity. Expected predictions $\bar{y}_v$ are approximated as the average of predictions made for training samples with the corresponding feature vector components substituted by value v.

of transcriptional control. Although isolated examples of IRESs with known ITAF binding sites or resolved three-dimensional structure are available [32–34], there are currently no systematic studies that aim at deciphering sequence elements governing cap-independent translation regulation. A major hindrance to progress in this direction is the relatively low number of known IRESs. The identification of novel IRES elements requires a series of labour-intensive reporter assays to confirm expression and to rule out the presence of cryptic promoter or splicing activity, so that only ≈120 IRESs were reported until recently [8]. Thus, unlike transcriptional regulation [35–37], attempts to systematically decipher determinants of cap-independent translation initiation were not feasible until now. In a recent work we developed a high-throughput IRES activity assay, and used it to identify thousands of novel IRESs in human and viral genomes [1], thereby expanding the dataset of known IRESs by 50-fold and allowing for the first time the construction and interpretation of predictive models.

Here we perform an in-depth computational analysis of data from our high-throughput IRES activity assay [1] to explore the relationship between RNA sequence and IRES activity. We find several common sequence k-mer features predictive of IRES activity that are shared between (i) sets of viral IRESs originating from viruses of the same type, and (ii) sets of cellular IRESs originating from similar locations within human transcripts, as well as features specific to retroviral IRESs. These features include the poly-U, poly-A and C/U-rich k-mers, many of which are found upstream of the start AUG in distinct "location islands", continuous stretches of positions where these sequence features have the strongest effect, suggesting that positions of ITAF binding sites relative to the AUG are important determinants of IRES activity. Finally, systematic measurements of hundreds of fully designed synthetic oligos confirmed our finding of a positive relationship between the number of short IRES elements in a sequence and its IRES activity. Together, we provide the first in-depth computational analysis of thousands of IRESs from the human genome and different types of viruses and offer novel insights into the relationship between RNA sequence and IRES activity.

5.2. MATERIALS AND METHODS

DATASET

In a recent study [1] we described a high-throughput IRES activity assay that we used to measure IRES activity for thousands of sequences. Briefly, we obtained a mixed pool of oligonucleotides, 210nt (174nt variable region plus constant primer sequences) in length, using parallel DNA synthesis technology [38–40]. We then amplified the library using constant primers, cloned it into the lentiviral bicistronic plasmid 12nt upstream of the eGFP (enhanced Green Fluorescent Protein) coding sequence (Fig. 5.1A) and infected H1299 human lung cells so that each cell integrates a single oligo. In this plasmid mRFP (monomeric Red Fluorescent Protein) is translated in a cap-dependent manner, whereas eGFP translation requires alternative mechanisms. We thus used eGFP expression as a proxy for IRES activity induced by the variable sequence. To obtain eGFP expression we sorted the resulting pool of cells into 16 bins according to eGFP fluorescence, while also filtering based on mRFP fluorescence to control for cell state, and used deep sequencing to compute a score for the expression of each

designed oligo based on the distribution of its sequence reads across expression bins. Additionally, we controlled for eGFP expression that could arise due to RNA splicing or promoter activity of the measured sequences (see Suppl. Text, page 145). Using this approach, we measured IRES activity of a library of 55,000 sequences, including 28,669 native fragments from the human and viral genomes. In the current study we use these measurements to uncover RNA sequence and structure determinants of IRES activity.

The library measured in [1] includes sequences originating from human transcripts and viral genomes. In particular, the library sequences were generated by (i) taking the sequences directly upstream of transcripts' translation start site; and (ii) by tiling transcripts and viral genomes with sequences to be measured. Because most sequences in such library are not expected to have IRES activity, ≈ 11% of the sequences showed activity above background levels (see Fig. 5.1B and Suppl. Fig. 5.8). Library sequences were taken from genomes of viruses with considerably different life cycles and replication strategies. For example, viruses from positive-sense ssRNA class replicate in the cytoplasm and their transcripts lack the 5′ cap structure, which is normally acquired in the nucleus. Thus, these viruses rely heavily on cap-independent translation mechanisms for gene expression. In contrast, retroviruses are integrated into the host genome and thus undergo transcription, RNA processing, and cap-dependent translation similarly to cellular genes. These differences in the available host gene expression machinery and subjection to distinct selection pressures due to the employed replication strategies [41, 42] may have prompted different viral classes to evolve distinct cap-independent translation strategies [43]. For this reason we separated viral sequences into (i) positive-sense (+) ssRNA viruses; (ii) negative-sense (−) ssRNA viruses; (iii) dsRNA viruses; and (iv) retroviruses based on their viral class (Fig. 5.1B). We similarly divided human sequences from the library into those originating from (i) the coding sequences (CDSes); (ii) the 5′ UTRs; and (iii) the 3′ UTRs of human transcripts, due to mechanistic differences between these regions [44, 45].

We analysed the above seven groups of sequences both together and individually. For each of the groups we learned a predictor of IRES activity from RNA features with the goal of elucidating sequence features that may determine IRES activity, and would consequently provide a prediction of the IRES activity for novel sequences.

RANDOM FOREST MODEL LEARNING

Our approach for learning sequence models of IRES activity is depicted in Fig. 5.1C-E. We chose Stochastic Gradient Boosting Random Forest regression for learning sequence models for several reasons. First, Random Forests (RFs) allow for construction of nonlinear predictors that offer established model interpretation techniques. Second, stochastic gradient boosting allows for achieving highly accurate predictions by fitting the gradient of the residual error with every new tree added to the forest, while being fairly robust to overfitting in practice [46]. The latter is especially important in our case, because for some of the considered groups of IRES sequences only a few hundred training instances are available (sequences with measured IRES activity) while thousands of features (M) are being used, leading to a situation that can easily result overfitting.

We used the scikit-learn software [47] to learn RFs from training data. We chose to

train 1000 trees per forest. To speed-up the training process, each tree only evaluated $\sqrt{M}$ features when choosing split features. The trees were allowed to have arbitrary depth, but their complexity was controlled by parameter m, defining the minimum allowed number of training samples per leaf node. This parameter was set, together with the learning rate r and subsampling fraction f, using a double-loop 10-fold cross-validation (CV) scheme on the available training data (described in detail in Suppl. Fig. 5.8). Briefly, each outer CV training set was randomly partitioned into 10 sets; every time, 9 of these sets were used as an inner training set and the remaining set was used for validation. For each of the 10 inner training sets, we learned an RF for every combination of the parameters (m, r, f) from a pre-defined grid and evaluated its performance (in terms of the R^2 statistic) on the held-out inner validation set. The parameter set with the highest average performance across the 10 validation sets was used for learning the final predictor on the outer CV training data, which was evaluated on the outer CV validation set. When randomly partitioning sequences into CV folds, we ensured that the numbers of sequences with background levels of IRES activity were balanced across sets.

k-MER FEATURE PRE-SELECTION

To explore the relationship between IRES sequence and activity, we described its primary sequence using numerical features which could be related to IRES activity by the learned RFs. We chose to represent IRES RNA sequences using k-mers and counted how many times every possible RNA subsequence of length $k \leq 5$ occurs the training sequences (see example in Fig. 5.1C). These counts were recorded for the entire sequences (global counts), as well as in moving windows of 20nt with a 10nt overlap (positional counts) to generate position-sensitive k-mer features. To assess the added predictive power of the k-mer copy numbers, we also created a k-mer occurrence feature description of the available RNA sequences, in which k-mer counts were capped at a maximum value of 1.

Because this representation of IRES sequences generates thousands of features, to facilitate model learning and interpretation we sought to reduce the number of used features by pre-selecting them prior to RF training. To this end, on the inner training set for each feature we (i) computed correlation coefficient and p-value for the Spearman rank correlation between feature values and IRES activity for k-mer counts; or, for k-mer occurrences, the Mann-Whitney U-test statistic and p-value to assess the difference between IRES activity distributions for sequences with and without the feature; and (ii) counted in how many training samples the feature value was non-zero. Only features with an association significant at a false discovery rate of 0.05 (controlled using the Benjamini-Hochberg procedure) and features present in at least 10% of the sequences were used for model learning.

RANDOM FOREST FEATURE INTERPRETATION

Unlike linear models relying on L_1 regularisation (e.g. [48, 49]), RFs cannot perform simultaneous feature selection and learning. This means that all features provided to RFs will generally be used by the learned model to make predictions. This property of RFs complicates model interpretation by increasing the number of features of the learned model that need to be examined. To efficiently sift through the features we calculate

their *feature importances* as in [50] and use them to select and prioritise interesting features (see Fig. 5.1D and F). For each tree in an RF, the feature importance of a variable captures its contribution to the resulting prediction by quantifying the total reduction in variance the variable provides each time it is selected as a split feature in this tree. The importance of a variable in an RF is then calculated as its average feature importance across all RF trees. To facilitate comparison of feature importances across models with different numbers of features, i.e. models obtained for different CV folds or sequence groups, we normalised importances of every model by dividing its feature importances by the maximum feature importance attained.

Similarly, because RFs do not provide a direct way of evaluating the direction of the effect (positive or negative) features have on the resulting prediction, we computed the *partial dependence* [50] of an RF w.r.t. its features at all possible values (see Fig. 5.1E and G). Partial dependence of a feature provides an estimate of the expected prediction (IRES activity) of a sequence with a given value for this feature. When plotted for all possible values of a selected feature, partial dependence allows for graphic inspection of the relationship between the feature and IRES activity. We observed that in practice, partial dependence often shows near-monotonic behaviour (see Suppl. Fig. 5.9 for representative examples), i.e. the expected prediction either tends to increase (or to decrease) with increasing feature values, and used this property to determine directionality of each feature based on the average derivative of its partial dependence. Features were classified as increasing IRES activity (positive) if their average derivative was positive, otherwise they were classified as negative (decreasing IRES activity). This classification can be thought of as a generalisation of the linear model variable separation into positive and negative based on their slopes (i.e. model coefficients).

To obtain robust results, partial dependences and feature importances were averaged across 10 RFs models trained on different outer CV folds.

SYNTHETIC DATA DESIGN AND ANALYSIS

We designed a total of 1024 oligos in which we planted the sequence of two short elements with experimentally validated IRES activity in 1-8 copies: (i) the TEV IRES (`UACUCCC`) and (ii) the Poliovirus type-2 IRES (`CGUCAAUUCCUUUA`) [51, 52]. Each oligo is composed of 164nt of variable sequence, 10nt of unique barcode at the 5′ end (barcodes differ by at least 3nt from each other) and constant primer sequences to amplify the oligos with PCR reaction. We chose one native and one synthetic background sequences (see Suppl. Table 5.1), which lack intrinsic IRES activity: (i) 164nt of the human beta-globin gene (HBB, NM_000518) that was used as a negative control in a previous study [53], and (ii) a concatenation of a 9-mer that was used as a spacer between multiple copies of the Gtx IRES in a previous study (Spacer1: `TTCTGACAT`; [54]). This set of 1024 sequences was measured for IRES activity as part of a 55,000 oligos library in a high-throughput bicistronic assay described before [1] and analysed here for the first time.

Synthetic construct measurements were filtered as the original data (see Suppl. Text, page 145) and analysed using ANOVA. Each of the two short elements (TEV and Poliovirus type-2) was analysed independently. To account for possible effects of background sequences on IRES activity, we averaged activity measurements across all

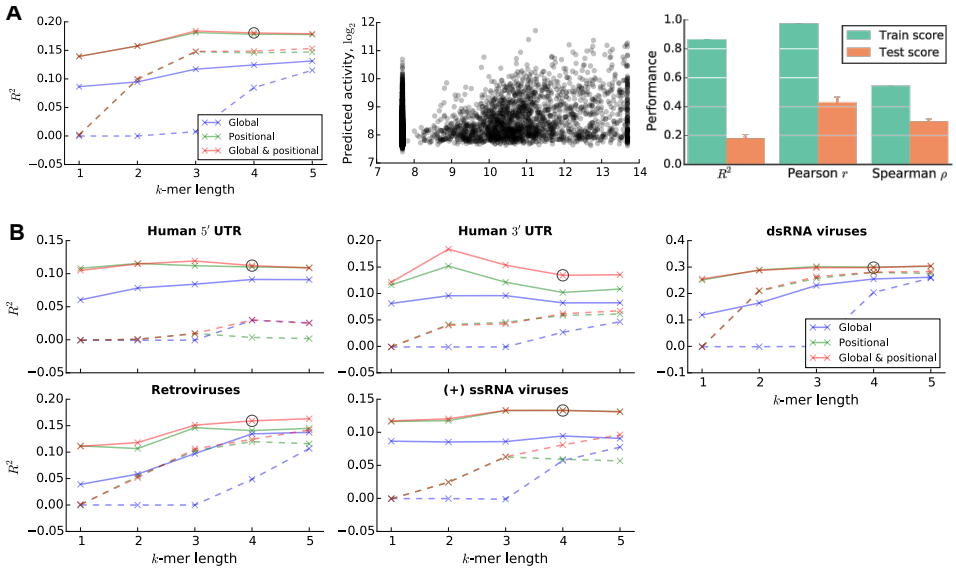

Figure 5.2: Performance of trained predictors. (**A**) Cross-validation (CV) performance of a model trained on all available native IRES sequences shown for different combinations of k-mer lengths, and k-mer count (solid lines) or presence (dashed lines) features (left), with the selected combination marked with a circle. Scatter plot of predicted and true IRES activities for the selected model (middle). And training and test performance of the selected model evaluated using several metrics. (**B**) CV performance of models trained for different groups of sequences. Only results for groups with models achieving sufficiently high performance are shown.

sequences with the same number of sites and background sequence type (HBB or Spacer1). However, to increase the number of samples, and the power of our analyses, we treated replicates as independent measurements and performed joint analyses of all samples from the two backgrounds. We analysed the samples in a two-way ANOVA model with factors "background sequence" (binary) and "number of sites" (integer):

$$\text{activity} = \alpha \cdot \text{background} + \beta \cdot \text{Nsites} + \gamma.$$

After obtaining a least squares fit of α, β and γ for $\log_2$ IRES activity on the available measurements, we tested for their significant with IRES activity using the F-test. For convenience of visualisation and continuity, we represented IRES activity values as the $\log_2$ fold change over background levels, as in [1].

5.3. RESULTS

PREDICTION OF IRES ACTIVITY FROM SEQUENCE

With the recent discovery of thousands of novel IRESs in human and viruses, providing a 50-fold increase over previously available data [1], the next big challenge is to uncover the RNA sequence features predictive of IRES activity. We sought to employ a machine learning approach for this purpose, in which we train Random Forests to predict IRES

activity from RNA sequence features, and then use the trained forests to uncover predictive sequence features. To this end we computed k-mer and structural features for all 20,872 available native IRES sequences, randomly partitioned the sequences into 10 sets of near-equal size and used them in a cross-validation scheme to train and test 10 independent RF models (see Materials and Methods). To get a comprehensive evaluation of model performance, we evaluated them using three metrics: the R^2 statistic, which quantifies the portion of variance in the data that is explained by the models, the Pearson correlation, r, and the Spearman rank correlation, ρ, calculated on test set predictions.

In a previous study we found that the effect of mutations on expression was not uniform across the IRES sequence, suggesting that in addition to the sequence of the functional elements, their position within the IRES is also important [1]. Thus, we tested the effect of both, global sequence features (counts of k-mers within the examined sequence) and positional sequence features (counts of k-mers within a specific region of the examined sequence; Fig. 5.1C). Further, we sought to check whether k-mer copy number information provides additional predictive power, compared to k-mer presence (k-mer counts capped at a maximum value of 1), and considered both feature representations in our models. We first learned combined models of IRES activity on the entire set of sequences without separation into groups based on virus type or location within transcripts. The models were learned for different combinations of k-mer length and k-mer feature types (global or positional; count or presence). The highest predictive power was achieved by a model that makes use of the global and positional 3-mer or 4-mer count features (see Fig. 5.2A, left). We selected this model with $k = 4$ for further analysis. Its test set R^2 is 0.18, indicating that RNA sequences can explain 18% of the variance in IRES activity of cellular and viral IRESs in human cells. The agreement between R^2 and the Pearson r of 0.429 (Fig. 5.2A, right) suggests that our models correctly capture the mean IRES activity in unseen test data. However, the differences between the test set Pearson and Spearman correlations ($r = 0.429$ and $\rho = 0.297$; Fig. 5.2A, right) indicate that the models are biased towards better prediction of extreme IRES activity values. This behaviour is expected from the skewed IRES activity distribution of the available sequences (see Suppl. Fig. 5.7), in which the negative skew can be explained by the relatively low abundance of IRESs in human and viral genomes [55]; and by potential underestimation of IRES activity due to its dependence on cellular conditions. Given the good agreement between the three evaluation metrics, we chose to use the R^2 statistic in all our analyses.

We hypothesised that IRESs from different virus types and locations within human transcripts may have evolved distinct initiation mechanisms, which would be easier to learn in isolation. To test this hypothesis we separated the available human data based on their location within transcripts into sequences from (i) human 5′ UTRs, (ii) human 3′ UTRs and (iii) human CDSes; and the available viral data based on their virus type into sequences from (iv) positive-sense ssRNA viruses, (v) negative-sense ssRNA viruses, (vi) dsRNA viruses and (vii) retroviruses, irrespective of their position in the viral genome of origin. We then learned RF models for each of the groups as before. As can be seen from their test R^2 in Fig. 5.2B, in line with our hypothesis, IRES activity in some groups could be predicted much better than in others. In particular,

the variation in R^2 between defined groups is significantly higher than expected by chance (see Suppl. Text, page 150). Specifically, the R^2 statistic for the group of dsRNA viruses is 0.298, a considerable improvement in predictive power over the combined model. At the same time we also found that in some groups cannot be predicted by the proposed approach (e.g. the human CDSes, $R^2 \approx 0$, or the negative-sense ssRNA viruses, $R^2 = 0.036$; see Suppl. Fig. 5.10). Translation initiation of IRESs from these groups may rely on mechanisms that are poorly captured by primary sequence features, such those involving pseudoknots and the three-dimensional structure of RNA molecules. Additionally, these groups have the lowest absolute and relative incidence of active IRESs ($\approx 6.4\%$), which makes it difficult to learn predictive models (see Suppl. Fig. 5.11).

Interestingly, models based on the k-mer count features consistently achieved higher performance than their k-mer presence counterparts across all sequence groups. While this result is unsurprising, given that the count features provide a richer description of the sequences than the capped presence features, it also suggests possibilities for a regulatory effect of k-mer copy number on IRES activity.

We have also considered several types of RNA structure features, which captured local RNA accessibility and base pairing between regions of the RNA. Individual structural features were pre-selected based on their correlations with IRES activity and used for model training in the same way as k-mer count features were (see Suppl. Text, page 147). However, despite being weakly predictive when used alone ($R^2 < 0.02$; Suppl. Text, page 147), the considered types of structural features did not allow for increasing model predictive power beyond what could be achieved using k-mer features alone.

GLOBAL SEQUENCE FEATURES RESEMBLE ITAF BINDING MOTIFS

Having obtained several predictive models, we sought to use them to elucidate individual sequence features that are strong determinants of IRES activity. Given the superior performance of models trained on the combination of global and positional count features (Fig. 5.2), we chose to interpret them, as it would provide a more faithful view of IRES features. Additionally, we chose to interpret models with $k = 4$ for all sequence groups irrespective of whether the highest predictive power is achieved at this k-mer length. This choice facilitates feature comparison at the cost of a negligible drop in performance for some sequence groups. Further, only the 5 groups with useful predictive models ($R^2 > 0.1$; Fig. 5.2B) were analysed.

For every sequence group we took k-mer features that were robust (present in all 10 CV models) and predictive (defined as having an average feature importance of at least 0.1; see Fig. 5.1D and F). For each of the selected features we also determined its directionality (positive or negative) from the shape of its partial dependence plot (see Materials and Methods, and Fig. 5.1E and G). We first sought to examine features that are consistently related to IRES activity across multiple sequence groups, i.e. common features, and thus focused on those k-mers that were predictive and robust in at least two groups. In Fig. 5.3A we show common k-mer count features separated into several classes based on their composition and effect; the remaining non-common features are shown in Suppl. Fig. 5.12.

Our predictive k-mer analysis recapitulates the findings from [1], as we also show

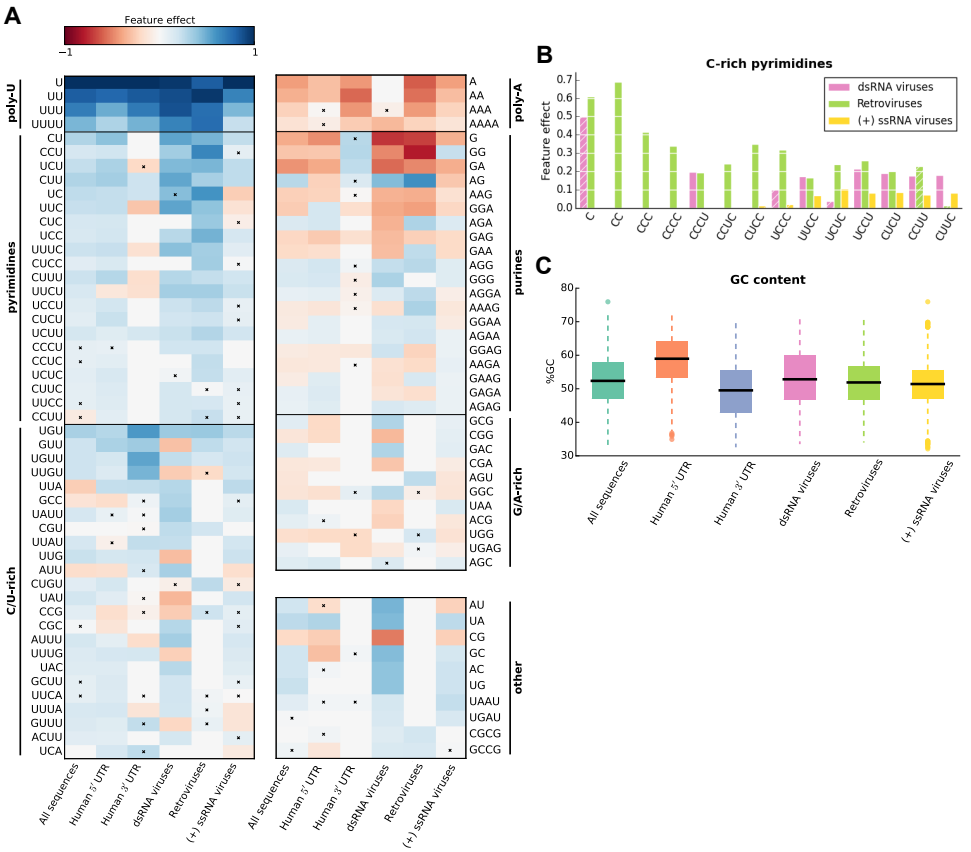

Figure 5.3: Overview of IRES global sequence features. (**A**) Robust and predictive global k-mer count features that appear in at least two IRES sequence groups; features were divided into classes based on their nucleotide composition and interpretation (vertical bars). For each feature, its effect (feature importance taken with sign "+" if the feature was classified as positive, and with sign "−" otherwise) is shown, and non-robust features are marked with a cross. (**B**) Comparison of C-rich pyrimidine tract feature importances across three viral sequence groups; non-robust features are shown with hatched bars. (**C**) Sequence GC content distribution for the defined sequence groups.

that k-mers presenting the poly-U motif are consistently selected in all sequence groups with poly-U k-mer presence being associated with increased IRES activity. However, in addition to the poly-U motif discussed in [1], we found that (i) k-mers representing pyrimidine (C/U) tracts are also strong determinants of IRES activity; and that (ii) these k-mers can equally contribute to the activity of IRESs from various positions on the transcripts and in various types of viruses.

Poly-A k-mers represent another group of features shared across models for different sequence groups. However, adenine tracts were not previously associated with decreased IRES activity in human cells. Selection of these k-mers by the trained models may be a consequence of an anti-correlation between the count of A/G and

U/C nucleotides in the measured sequences. However, Poly-G *k*-mer are generally not present in the trained models, suggesting that a mechanism specific to Poly-A tracts is involved in IRES-mediated translation. Similarly, the purine tract features, which are mostly associated with decreased IRES activity, can be explained by an anti-correlation between presence of purines and pyrimidines in sequences, and by an additional adenine tract specific mechanism.

Our results suggest that despite differences in model predictive power between sequence groups, robust and predictive global *k*-mer features are often shared by multiple groups, in which they agree on the effect they have on IRES activity (Figs. 5.3A and S7). However, we also sought to uncover features that are specific to a single sequence group or viral class. When reviewing features that were robust and predictive only for a single sequence group (Suppl. Fig. 5.12), we found that a number of pyrimidine tract features (C_{1-4} and UC_3) were uniquely selected for the retroviruses group. Interestingly, these features are all C-rich *k*-mers, whereas the common pyrimidine tract features, shared by multiple sequence groups, are not (Fig. 5.3A). This preference of retroviral IRESs for C-rich *k*-mers can be clearly seen from differences in feature importances of C-rich pyrimidine tract features across viral sequence groups (see Fig. 5.3B), which show that C-rich features are either uniquely used by the retroviral predictive models, or have the highest importance in those models. Furthermore, preference for C-rich *k*-mers within the group of retroviral sequences does not appear to be a consequence of GC-content bias, which is similar between retrovirus and (+) ssRNA virus groups (Wilcoxon rank-sum test, $p > 0.06$) and lower in retroviruses compared to dsRNA viruses (Wilcoxon rank-sum test, $p < 10^{-7}$; see Fig. 5.3C).

SYSTEMATIC MEASUREMENTS REVEAL THAT INCREASING THE NUMBER OF SHORT IRES ELEMENTS CAN LEAD TO ELEVATED IRES ACTIVITY

Collectively our *k*-mer count feature analyses (Figs. 5.2, 5.3A and S4) suggest that increasing the copy number of short "IRES elements" in an mRNA sequence would lead to increased IRES activity. In order to systematically test the effect of the number of short IRES elements on expression we designed synthetic oligos, in which we introduced the sequence of two short elements in 1-8 copies. We focused on short elements with reported IRES activity and high abundance of C/U nucleotides from the Tobacco Etch Virus (TEV, `UACUCCC`) and Poliovirus Type-2 (PV-2, `CGUCAAUUCCUUUA`) [51, 52]. To control for the effects of additional parameters varied between designed sequences, such as the distance of the site from the start `AUG`, the distance between two adjacent elements and the immediate flanking sequence in each position, we introduced each IRES element in all possible combinations of 1-8 sites at 8 predefined locations within two different backgrounds, resulting in a total of 1024 oligos (256 oligos for each element in each background; Fig. 5.4A). This set of sequences was measured for IRES activity as part of the 55,000 oligos library described before [1] in a high-throughput bicistronic assay using fluorescence activated cell sorting (FACS) and deep sequencing (see Materials and Methods).

We then computed the average IRES activity across all oligos with the same number of sites in each background sequence separately (Figs. 5.4B and C). To test the association between the number of short IRES elements and the measured IRES activity

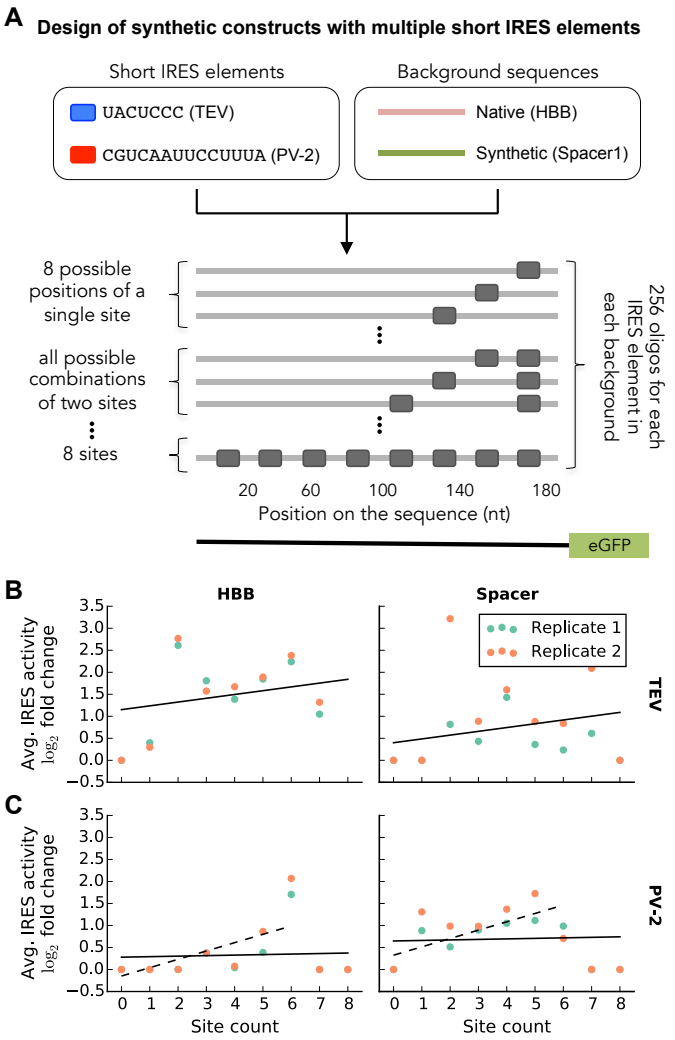

Figure 5.4: Design and results of experimental study of the effect of IRES element count on activity. (**A**) Two IRES elements (TEV and PV-2; coloured blocks) were placed in all possible combinations of 1-8 sites in predefined positions of two background sequences (native and synthetic; coloured lines) to generate synthetic oligos (grey blocks and lines), which were measured using the biscistronic IRES activity reporter assay. (**B**) Average measured IRES activity ($\log_2$ fold change over background levels) of these oligos shown for two biological replicates (green and orange circles) as a function of the number of the TEV IRES elements in synthetic sequences with two different backgrounds (left and right columns). The relationship between IRES activity and the number of sites is shown as a black line. (**C**) Same as B, but for the PV-2 IRES element. The relationship between IRES activity and the number of sites computed from data with up to 6 sites is shown as a dashed line.

we used a two-way ANOVA model. To increase the power we performed joint analyses

of the computed average IRES activity in the two backgrounds and biological replicates (green and orange circles in Figs. 5.4B and C). Notably, we observe a positive association between the number of TEV sites and IRES activity (Fig. 5.4B; F-test, $p < 0.01$). Moreover, we also found significant associations between IRES activity and background sequence type (F-test, $p < 0.05$), which suggests that sequences surrounding the introduced motif can determine whether or not the motifs will be functional. Examining the relationship between the number of sites and expression for the PV-2 IRES element reveals a more complex behaviour. A general trend of elevated IRES activity is obtained when increasing the number of elements up to six sites (F-test, $p < 10^{-3}$; Fig. 5.4C, dashed line). However, in both backgrounds a clear drop in expression is observed when placing 7 or 8 sites. One potential explanation for this is that higher numbers of elements, which result in smaller distances between adjacent sites, lead to steric hindrance as previously described for transcription factors binding sites when located in close vicinity [40]. Since the PV-2 IRES element is longer than the TEV element (14nt vs 7nt respectively) the resulting minimal distance between adjacent sites in the designed oligos is smaller (4nt for PV-2 sites vs. 11nt for TEV sites). Thus, it is possible that the synthetic PV-2 sequences were more sensitive to steric hindrance effects than the TEV sequences, so that the decrease in expression is not obtained for the latter.

k-MER POSITION IS A STRONG DETERMINANT OF IRES ACTIVITY

Having obtained a rendering of the global *k*-mer features predictive of IRES activity, we sought to expand our analysis of the effect that *k*-mer location may have on IRES activity. We were encouraged by the results of training models on different combinations of global and positional *k*-mer features (Fig. 5.2B) which showed that for all sequence groups models trained on positional features achieved highest performance, suggesting that *k*-mer position relative to the start AUG is a strong determinant of IRES activity.

To investigate this further we assessed the effect of positional *k*-mers as a function of their location in the sequence. We first focused on those positional *k*-mer features that were common to multiple sequence groups. To this end positional features were investigated only for those *k*-mers, which showed a robust location-specific signal (had at least two windows where the *k*-mer feature was selected in all CV folds), were predictive (had an average importance in those windows of at least 0.1) and were shared by several sequence groups (i.e. the windows were also robust and predictive for at least one more group). Common positional features in Fig. 5.5 are shown as heat maps depicting *k*-mer effect along the sequence and across sequence groups, which is summarised as a consensus effect, i.e. the largest effect at a particular position that is supported by multiple groups; the remaining positional features are shown in Suppl. Fig. 5.13.

Interestingly, nearly all predictive positional *k*-mers from Fig. 5.5 were also selected as robust and predictive global *k*-mer count features in Fig. 5.3. In particular the poly-U and pyrimidine *k*-mers are among the most predictive *k*-mers for both feature types. However, positional feature plots additionally show that effect strengths of these *k*-mers differ with their position relative to the start AUG. For example, the $U_{1\text{-}3}$ *k*-mers have an overall positive effect on IRES activity, which is largest if the *k*-mers are located about 50nt upstream of the start AUG.

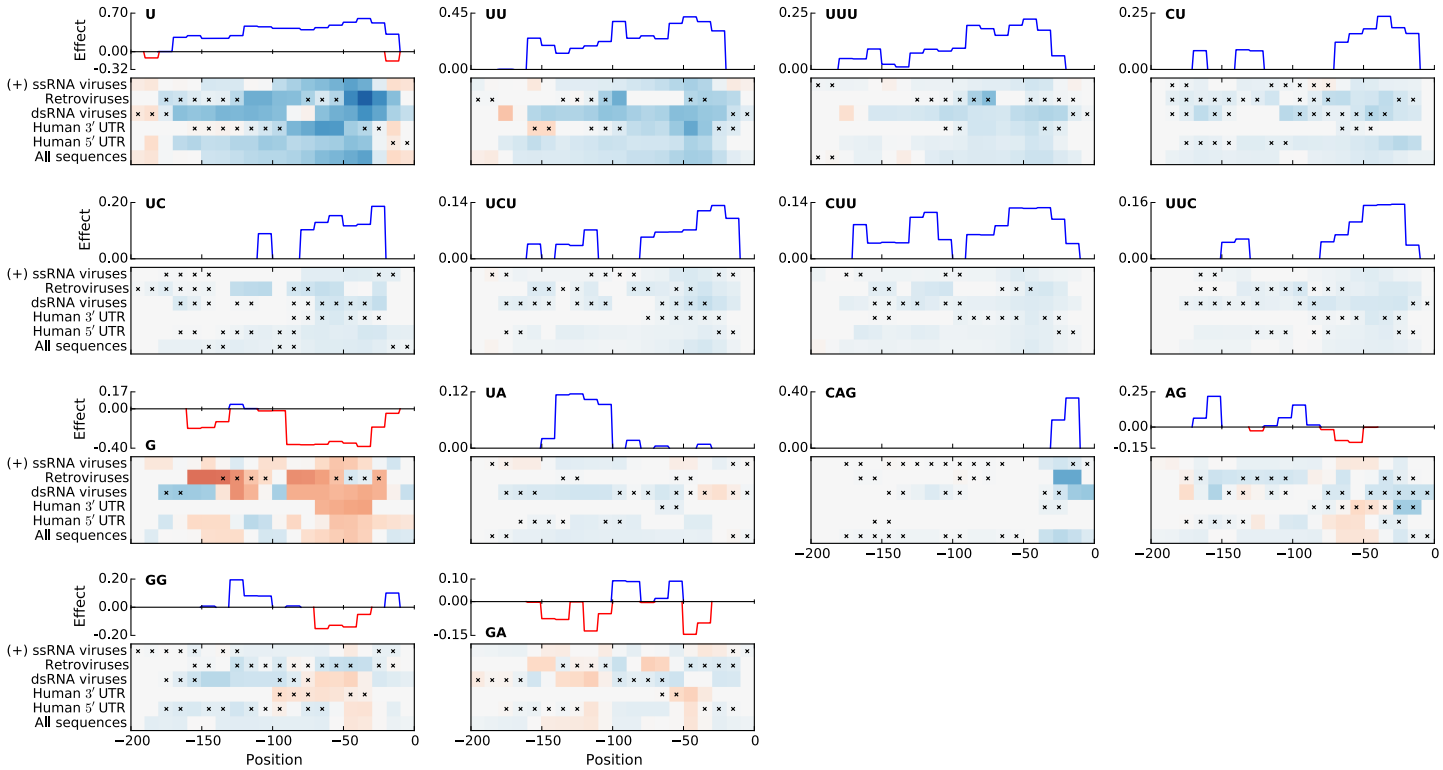

Figure 5.5: Robust and predictive positional features that appear in at least two of the analysed groups. For each feature, its effect along sequences is shown in a heat map (see Fig. 5.3), and summarised as a consensus effect (located above each of the heat maps) across several groups, chosen as the effect whose directionality and importance are confirmed by at least two groups. Horizontal axes show feature window position relative to the start AUG.

At the same time, many other features (e.g. CU, UUC, G and CAG) also show positions location-specific effects on IRES activity. Most notably, positional features of these *k*-mers tend to form "islands" from positions at which they have an effect on activity. These islands are consistently located around positions −50 (*k*-mers CU, UC, UCU, CUU, UUC, G, AG and GA) and −150 (*k*-mers G, UA, AG and GA). Interestingly, for the majority of presented *k*-mers, positions with the strongest effect are not located directly upstream of the start AUG. Further, congruence between optimal location for *k*-mers with negative effects (G, AG, GG, GA) and optimal locations for C/U-rich *k*-mers with positive effects further supports our interpretation of the poly-A, purine tract and G/A-rich *k*-mers as anti-correlated with the C/U-rich *k*-mers.

The CAG *k*-mer also shows distinct positional preferences for locations immediately upstream of the start codon. We further investigated its effect to determine whether it is a part of a larger motif, and whether there is a difference in splicing between sequences with and without the CAG *k*-mer. Our analyses (see Suppl. Text, page 145) indicate that the CAG *k*-mer may be related to RNA splicing in the group of dsRNA viruses, but not in Retroviruses.

In addition, a large number of *k*-mers are robust and predictive only for a single sequence group (Suppl. Fig. 5.13). Similar to the global *k*-mer features, the unique positional *k*-mers include C-rich *k*-mers C, CC, CUCC, UCC, CUC selected exclusively by the retroviral group. Interestingly, these *k*-mers show positional preferences different from those of the common positional *k*-mers, by forming islands around positions −50 and −200. Finally, we also found that a number of predictive positional *k*-mers are selected uniquely for the group of dsRNA viruses (e.g. AU, ACC, UG, AUU, UAC; Suppl. Fig. 5.13); these positional *k*-mers show little consistency in terms of preferred positions, suggesting a different mode of action of IRESs from dsRNA viruses.

5.4. DISCUSSION

In this work we provide the first in-depth computational analysis of thousands of IRESs from the human genome and different types of viruses. Analyses of this largest set of IRESs to date allowed us to decipher the effect of sequence features, their number and position relative to the AUG on IRES activity (summarised in Fig. 5.6A). To achieve this, we trained and interpreted Random Forest models that predict IRES activity from *k*-mer features of RNA sequences.

IDENTIFIED *k*-MERS RESEMBLE ITAF BINDING MOTIFS

Using the trained models, we identified robust and predictive *k*-mer features, which based on their composition could be divided into two classes: pyrimidine-rich elements, and purine-rich elements (Figs. 5.3A and 5.6A). Notably, *k*-mers from these classes are generally associated with the same kind of effect on IRES activity: pyrimidine-rich elements tend to have a positive effect on activity, whereas the purine-rich elements tend to have a negative effect.

Interestingly, sequences of predictive pyrimidine-rich *k*-mers resemble consensus binding motifs of known IRES *trans*-acting factors (ITAFs). The poly-U *k*-mers are consistent with the poly-U binding motif described for the hnRNP C1/C2 [56]

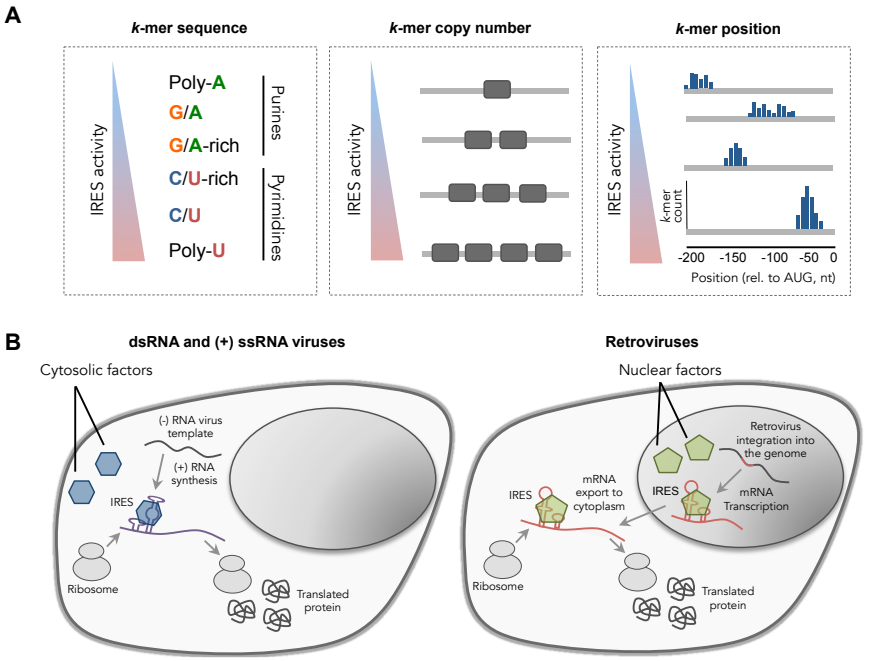

Figure 5.6: Summary of the sequence features associated with IRES activity. (**A**) Illustration of the sequence features found by our models and their association with IRES activity: (left) k-mer sequence, (middle) the number of sites of a *k*-mer, and (right) the position of the *k*-mer relative to the AUG start codon. (**B**) Illustration of the different life cycles of (left) dsRNA/(+) ssRNA viruses and (right) Retroviruses which may have led to differences in their IRESs sequence features. Retroviruses are integrated into the host genome and RNA-PolII transcribes their mRNA in the nucleus. Thus, their IRES elements are exposed to the nuclear environment including mRNA modifying enzymes (methylation, pseudouridylation etc) and nuclear specific ITAFs that can shuttle with the mRNA to the cytoplasm to facilitate cap-independent recruitment of the ribosome. In contrast, dsRNA and (+) ssRNA viruses that spend their entire replication cycle in the cytoplasm are exposed to cytosolic factors, which in turn can facilitate cap-independent recruitment of the ribosome.

RNA-binding proteins (RBPs), which were shown to be a part of the protein complex forming the XIAP IRES [57]. Whereas the pyrimidine-rich *k*-mers are consistent with the binding motifs of the PCBP-2 [58], PCBP-1 [59] and PTB-1 RBPs. The PCBP proteins were previously implicated in regulating IRES activity of the hepatitis C virus, poliovirus and rhinovirus IRESs [60], and the human proto-oncogene c-*myc* [61]. And the PTB-1 was previously shown to interact with many cellular and viral IRESs [25], and proposed as an universal ITAF [55]. The correspondence between ITAFs and pyrimidine-rich *k*-mer features, and the strong positive effect of the poly-U and pyrimidine tract *k*-mers on IRES activity (Fig. 5.3A), agree with the proposed role of ITAFs as RNA-binding proteins involved in cap-independent translation initiation.

In accordance with this interpretation, we observed that C/U-rich *k*-mers that contain a single non-C/U nucleotide tend to be associated with increased IRES activity. Given their similarity to the poly-U and pyrimidine tract *k*-mer features, interpreted as

potential ITAF binding sites, we propose that the C/U-rich *k*-mer features may represent imperfect binding sites of the PCBP and PTB proteins. This interpretation is supported by the observation that, compared to the perfect C/U-tract *k*-mers, features of this class tend to have a weaker effect on predicted activity.

Notably, systematic measurements of hundreds of fully designed oligos, in which the number of sites of the pyrimidine-rich TEV IRES element was carefully varied, support our finding of the positive relationship between the number of pyrimidine-rich elements and IRES activity. Thus, our study demonstrates the power of combining computational models with systematic measurements of synthetically designed oligos to decipher the principles governing IRES activity.

IRES ARCHITECTURES DIFFER BETWEEN VIRUS TYPES

Our results on common and unique sequence features uncover that poly-U and C/U-rich *k*-mers are shared among cellular and viral IRESs, including different families of viruses. This suggests that the involvement of ITAFs these *k*-mers represent in IRES-mediated translation initiation is not limited to a single viral class or location within human transcripts, but is shared across viral classes, as well as between viruses and eukaryotes. However, we also found that for IRESs originating form retroviral genomes, C-rich elements are stronger predictors of high IRES activity than for dsRNA and (+) ssRNA viruses (Fig. 5.3B) and have different positional preferences (Suppl. Fig. 5.13).

If pyrimidine tract *k*-mers indeed represent PCBP-1/2 and PTB binding sites, then while binding of these ITAFs to mRNA leads to increased IRES activity irrespective of its virus type, our results suggest that different virus types preferentially rely on different ITAFs for cap-independent translation initiation. The U/C-neutral *k*-mers are more consistent with the U[UC]U[UC]$_2$ and C$_2$U PTB binding motifs [55, 62] that have a weaker preference for cytosines, whereas the C-rich *k*-mers are more consistent with the UC$_3$U$_2$C$_3$U and U$_2$C$_6$AU PCBP-2 binding motifs [58] showing a stronger cytosine preference. Together this suggests that, compared to other sequence groups, retroviruses preferentially employ PCBP-1/2 RBPs for cap-independent translation initiation.

Interestingly, in contrast to most dsRNA and (+) ssRNA viruses, which spend their entire replication cycle in the cytoplasm, retroviruses are integrated into the host genome and their transcribed mRNA is exposed to the nuclear environment (Fig. 5.6B). Previous reports indicated that some IRESs require a “nuclear experience” in order to be functional [63–65]. It was suggested that nuclear specific events such as RNA modifications (by methylation, pseudouridylation and others) or the binding of exclusively nuclear ITAFs are required for certain IRESs. Our finding of retroviral IRESs preference for C-rich *k*-mers, presumably recognised by the PCBP ITAF, suggests that the mechanism by which IRES-mediated translation is accomplished, and consequently, IRES architecture, differ between viruses, which were evolved in differed cellular compartments and under different constraints. Taken together with numerous *k*-mer features, which were found to be predictive only for dsRNA IRESs (Suppl. Figs. 5.12 and S8), these results provide further support the proposition that viral IRESs arose independently several times in evolution [43].

ITAFS EXHIBIT DISTINCT LOCATION PREFERENCES

When considering positional *k*-mer features, we additionally found that many of the pyrimidine-rich features have a strong positional preference for location islands approximately 50nt and 150nt upstream of the start codon and a similar positive effect on the predicted IRES activity (Figs. 5.5 and 5.6A). The positive effect of these features, their similarity to ITAF binding motifs, and preference for distinct locations upstream of the start codon collectively suggest that ITAFs, whose (partial) binding motifs these *k*-mers describe, have multiple distinct optimal locations upstream of the start AUG at which they can contribute towards cap-independent translation initiation.

Intriguingly, predictive positions of the C-rich *k*-mers differ from that of the poly-U and U/C-neutral *k*-mers, and show a preference in retroviral IRESs for locations approximately 200nt upstream of the start codon. This further supports our proposition that IRESs originating from retroviral genomes rely more on PCBP-1/2 ITAFs for translation initiation, and suggests their optimal binding location.

RNA STRUCTURE AS A DETERMINANT OF IRES ACTIVITY

In our analyses we were unable to find a strong predictive relationship between RNA secondary structure and IRES activity (see Suppl. Text, p. 147), although RNA structure was previously shown to be functionally important for some viral IRESs. There are several possible reasons: First, the high-throughput assay conducted in [1] used designed synthetic oligonucleotides as the input sequence. Thus, the length of the tested sequences was limited to 174nt, which is shorter than some reported long structural viral IRESs [8]. It is possible that the identified IRESs do not form complex secondary structures as reported before (e.g. [66]), therefore limiting our ability to detect structural features in the current dataset. Second, it was shown that IRESs can form dynamic structures and that the binding of ITAFs can induce conformational changes that, in turn, facilitate IRES activity [67]. Thus, *in silico* prediction of RNA structure may differ considerably form the *in vivo* structures in the presence of ITAFs. In addition, computational predictions are limited in the ability to model complex tertiary structures such as pseudoknots. In order to investigate the relationship between RNA structure and IRES activity systematic measurements of secondary structures should be performed on the assayed sequences in cells. Recent advances in technology that facilitate high-throughput structural measurements *in vivo* [68] can shed light on this important layer of IRES regulation.

In this study we demonstrated that RNA sequence is predictive IRES activity, and proposed common and virus type-specific sequence *k*-mer features that may play a functional role in determining IRES activity, and could be used to predict IRESs *in silico*. Our results also yield a high-level IRES architecture of sequence features and their spatial organisation in RNA sequences, which suggests optimal positioning of ITAF binding sites upstream of the start AUG, and may be used to guide future synthetic IRES designs.

5.A. SUPPLEMENTARY INFORMATION

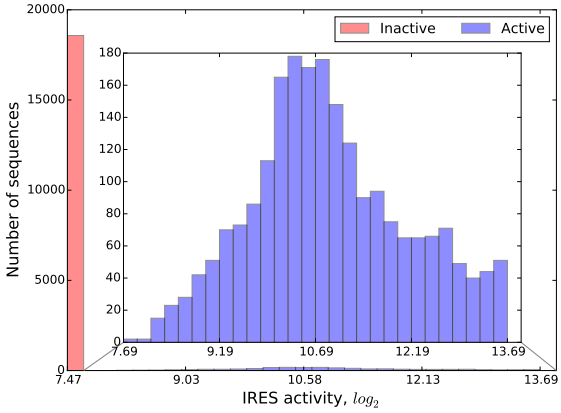

Figure 5.7: IRES activity distribution for all sequences remaining after filtering. Inset plot shows distribution of IRES activity in active sequences (IRES activity above background levels).

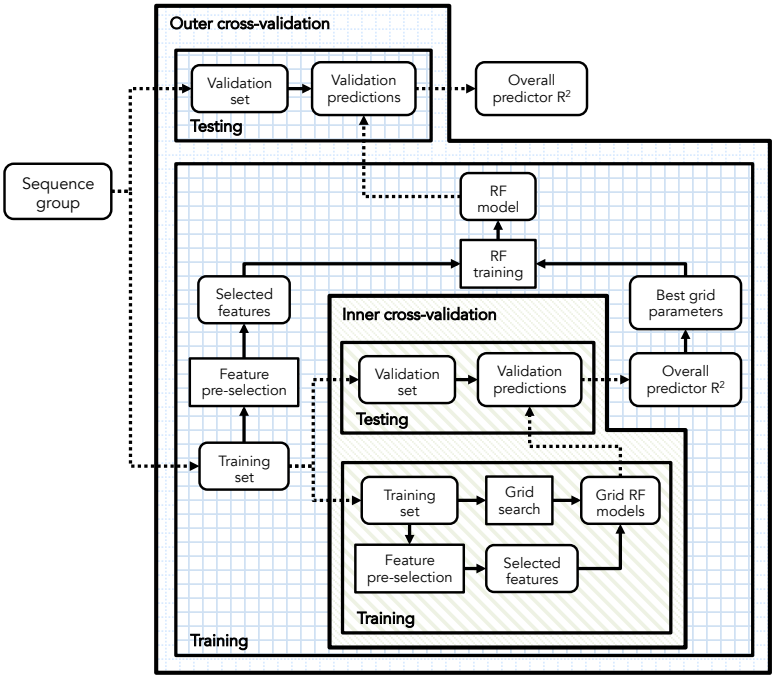

Figure 5.8: Cross-validation scheme employed for training RF models. Rectangular boxes denote actions or procedures, whereas round boxes are used denote their input or output (results); hatched boxes group items that belong to the same CV loop (outer or inner) or CV set (training or testing); arrows show how information flows through the CV procedure, with the arrows crossing CV loop/set boundaries drawn using dashed lines.

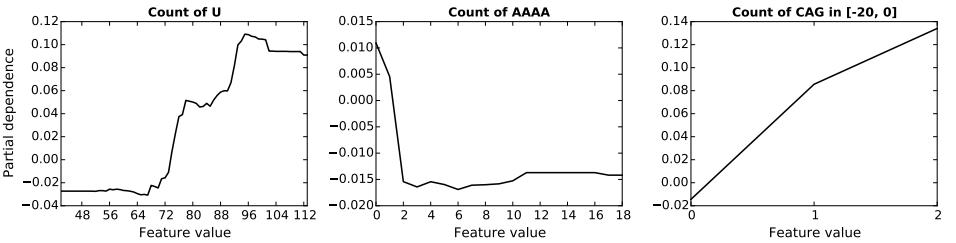

Figure 5.9: Representative examples of partial dependence plots. Three features from the dsRNA viruses models ($k = 4$, averaged over 10 CV folds): features U, AAAA and CAG in $[-20, 0]$ (as shown in the order from left to right) were respectively classified as positive, negative and positive.

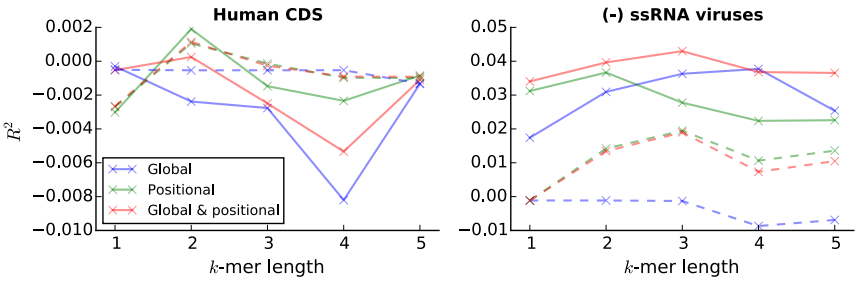

Figure 5.10: CV performance of k-mer count (solid lines) or presence (dashed lines) models trained on Human CDS and negative-sense ssRNA viruses sequence groups.

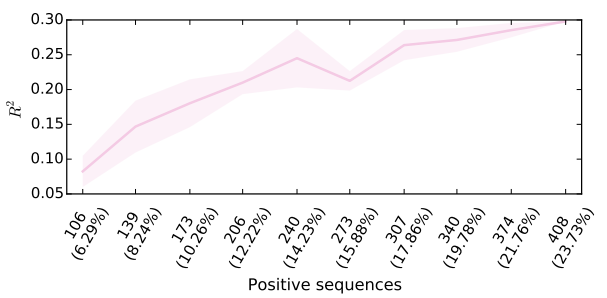

Figure 5.11: CV performance of models trained on subsamples of sequences from the group of dsRNA viruses. All models use global and positional k-mer counts ($k = 4$). Horizontal axis shows the number and the relative percentage of positive IRESs in the dataset, with the leftmost point (106 sequences) corresponding to the relative incidence of positive IRESs in the (−) ssRNA viruses group. Mean performance (solid line) and its standard deviation (shaded area) are shown for 5 random subsamples. These results indicate that small numbers of positive IRESs in a training set can limit predictive power of models trained on that set.

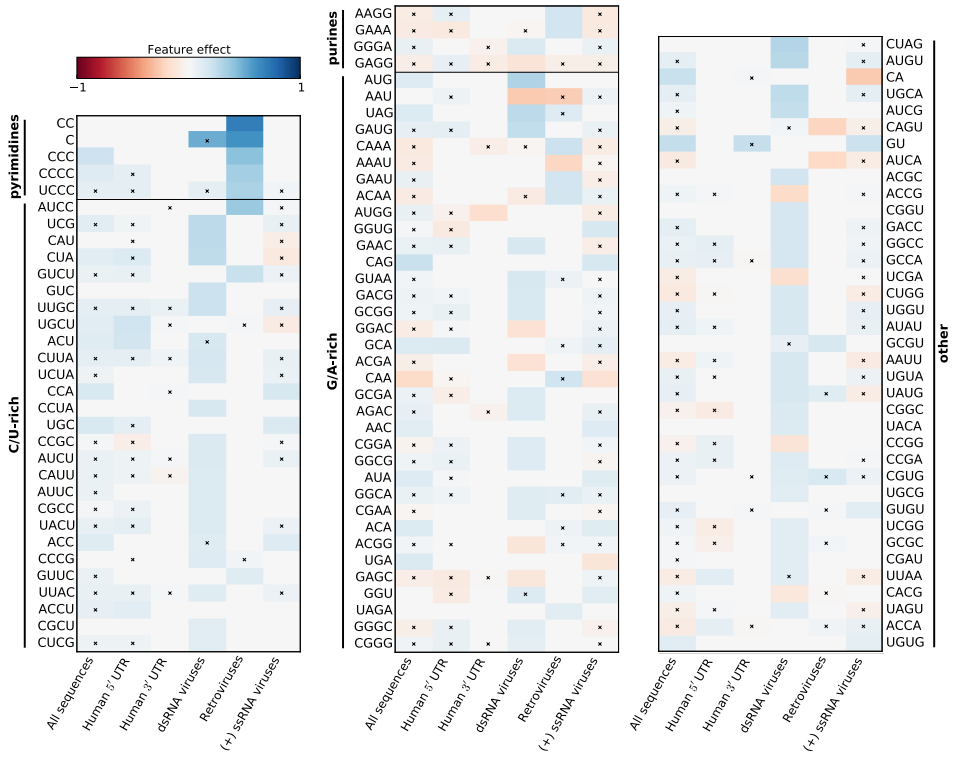

Figure 5.12: Robust and predictive global *k*-mer features that are uniquely selected by one sequence group.

Table 5.1: Sequences of oligos with no IRES elements (i.e. background sequences) used in synthetic designs.

Name	Sequence
HBB	CTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGAGGA GAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCC TGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGT
Spacer1	TCTGACATTTCTGACATTTCTGACATTTCTGACATTTCTGACATTTCTGACATTTCTGACATTTC TGACATTTCTGACATTTCTGACATTTCTGACATTTCTGACATTTCTGACATTTCTGACATTTCTG ACATTTCTGACATTTCTGACATTTCTGACATTTC

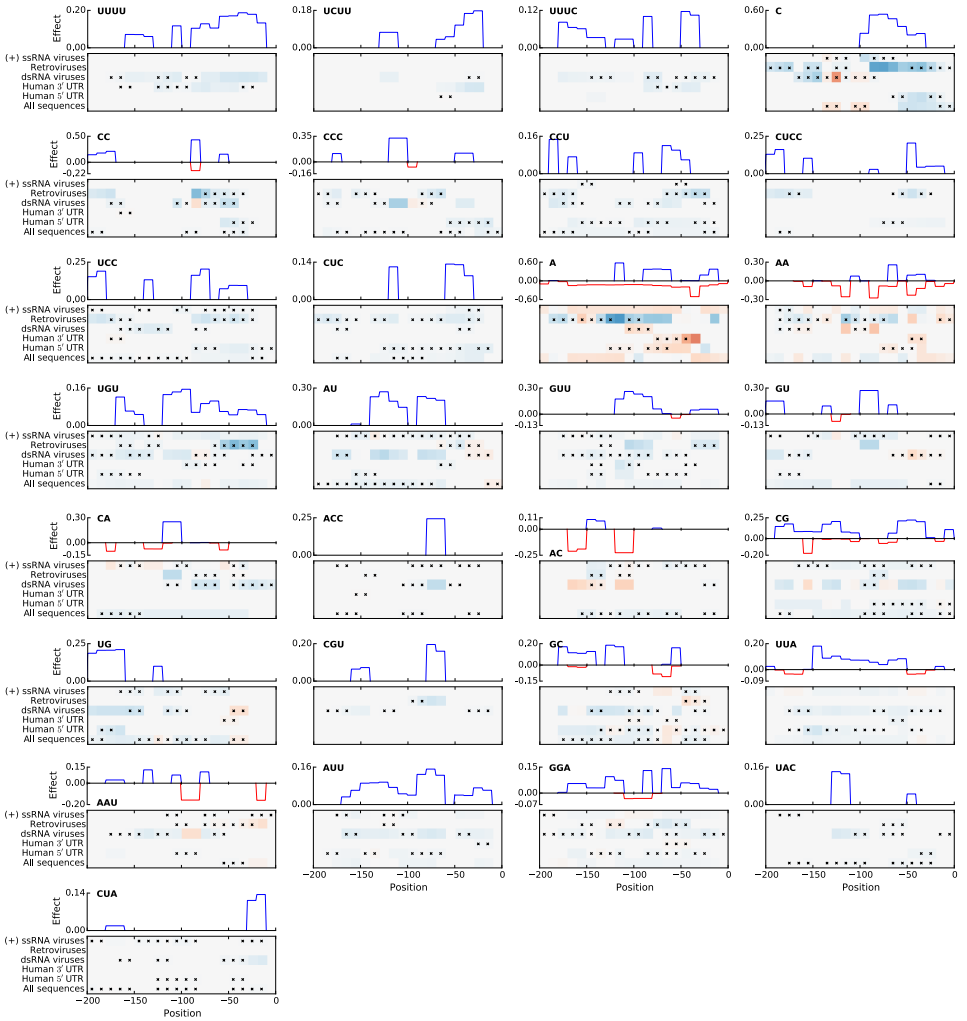

Figure 5.13: Robust and predictive positional k-mer features that are uniquely selected by one sequence group.

DATA PRE-PROCESSING

IRES activity measurements analysed in this work are complemented by high-throughput measurements of splicing activity and promoter activity [1]. As in the original manuscript, these additional measurements were used to filter out unreliable sequences, i.e. sequences whose eGFP expression was likely to be a result of cap-dependent translation due to (i) the mRFP and the assayed sequence being spliced out using a splice acceptor site present in the assayed sequence, or due to (ii) independent transcription of the eGFP from a cryptic promoter in the assayed sequence. To this end, following Weingarten-Gabbay *et al.*, all oligos with splicing scores below -2.5 or promoter activity above 0.2 were removed from the analyses. To further reduce the fraction of oligos, for which eGFP translation could be a result of splicing, we additionally removed all positive sequences (IRES activity above background levels) for which splicing activity could not be measured.

Several filtering and averaging steps were taken in order to obtain more reliable estimates and to increase robustness of the learned sequence models. First, for all analyses measured IRES activities were $\log_2$-transformed and averaged across the two replicates. Then, IRES sequences that had background IRES activity levels in only one of the replicates, and sequences that could be measured in at least one of the replicates were filtered out. Finally, to reduce the affect of outlier sequences with very high IRES activity on the learned predictive models, IRES activities were capped at the 99.5% percentile.

Further, because sequences outside of the 174nt variable region can affect IRES activity (e.g. by forming secondary structure with the variable region), for our analyses we extended the variable region by 84nt downstream and 74nt upstream as shown in Fig. 5.1A (solid filling).

RANDOM FOREST PARAMETER GRID SEARCH

When learning random forests, parameters were chosen using a grid search performed on the inner CV loop that evaluated all possible parameter combinations. The learning rate r, minimum number of leaf node training samples m and subsampling fraction f parameters were chosen in this way from grids $[0.001, 0.002, 0.004, 0.008]$, $[5, 25, 125]$ and $[0.9, 0.7]$ respectively.

DETAILED ANALYSIS OF THE UPSTREAM CAG FEATURE

The CAG k-mer in Fig. 5.5B does not share positional preferences of other features for locations around -50 or -150; instead its effect is strongest when it is located close to the start `AUG` at positions $[-30, 0]$. We expected that if this k-mer is a part of the optimal translation initiation context or splicing signal, then it would show further position or reading frame preferences within the $[-30, 0]$ window. To check this, we analysed CAG position preferences, sequence around CAG, and splicing score difference between sequences with and without CAG for the groups of dsRNA viruses and retroviruses. These groups were chosen as they are the two most specific sequence groups for which this feature was consistently selected across all CV folds and had a strong effect.

First, we compared position distributions for CAG within the $[-20, 0]$ window between positive and negative dsRNA virus sequences. Fig. 5.14A shows a strong

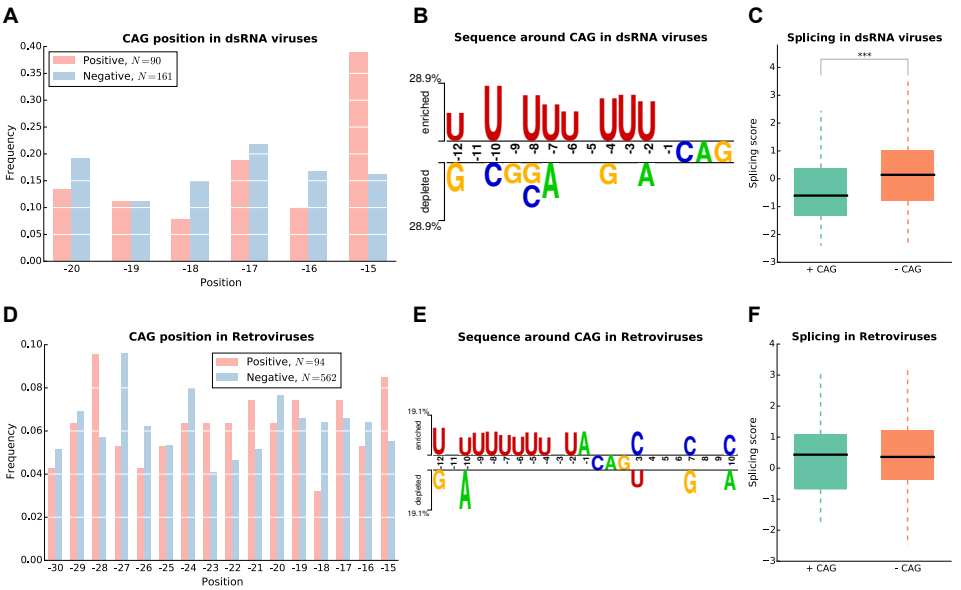

Figure 5.14: Detailed analyses of the CAG k-mer within the $[-20,0]$ window of the dsRNA viruses group (top row, **A**-**C**) and within the $[-30,-10]$ window of the retroviruses group (bottom row, **D**-**F**). (**A**, **D**) position preference; (**B**, **E**) enriched motif around the k-mer; and (**C**, **F**) distribution of splicing scores for positive sequences with (+CAG) and without (−CAG) the k-mer in the corresponding window.

preference of the CAG k-mer in dsRNA virus IRES sequences for position −15, i.e. the end of the variable part of the assayed sequences (positions $[-12;0]$ are the same for all sequences; see Fig. 5.1A). We then sought to determine whether this k-mer is a part of a larger sequence motif and checked for position-specific nucleotide enrichment between the sets of positive and negative dsRNA virus sequences with a CAG in the $[-20,0]$ window. Fig. 5.14B shows a significant (Binomial test $p < 0.05$; visualised using the Two Sample Logo website, Vacic *et al.* [69]) enrichment for Us upstream of the CAG k-mer; the downstream part was not included in the analyses due to the strong preference of the CAG for positions right before the constant part of the sequences. Remarkably, the enriched sequence resembles the canonical splice acceptor motif of poly-U followed by N[CT]AGG [37], suggesting that the CAG k-mer may be a part of a splicing site located at the end of analysed IRES sequences.

Presence of such a splicing site may lead to the loss of mRFP and the assayed IRES sequences in spliced mRNAs and result in translation of the eGFP protein through classical cap-dependent initiation mechanisms. To confirm that this is indeed what may be happening, we compared distributions of splicing scores from Weingarten-Gabbay *et al.* [1], which are indicative of the $\log_2$ splice-in ratios for the assayed sequences, between positive dsRNA virus sequences with a CAG in the $[-20,0]$ window and without it. Fig. 5.14C shows that IRES sequences with a CAG k-mer in the given window tend to have significantly smaller splicing scores than the sequences without it (Mann-Whitney

U-test, $p < 0.001$), suggesting that the +CAG sequences are spliced more often.

We repeated the above analyses for the retroviral group, and found that it only partially recapitulates the results obtained for the group of dsRNA viruses. In particular, while we found a similar poly-U enrichment upstream of the CAG, there was no longer a strong preference for position −15, and the difference in splicing scores between −CAG and +CAG sequences was not present. Neither the reason for differences in position preferences between dsRNA viruses and retroviruses, nor a possible mechanism that could link CAG −15 position preference and splicing activity, are clear to us.

Presence of active splicing signals in IRES sequences is problematic for the IRES activity assay, as its measurements may be inflated by eGFP produced via cap-dependent translation mechanisms. However, our analyses of predictive RNA sequence features across different groups of sequences suggest that splicing signals may only moderately effect IRES activity measurements, since only a handful of presented sequence features could be linked to the splicing mechanism. Moreover, the sequence overlap between the splicing acceptor motif and the hnRNAP C1/C2 ITAF binding motifs, both of which require the presence of a poly-U stretch, suggests that co-occurrence of splicing and IRES activity is a general phenomenon. This is supported by the fact that most of the known ITAFs have also been implicated in pre-mRNA splicing [43]; and by the existence of IRESs, such as XIAP, which are known to contain splice sites [70].

RNA SECONDARY STRUCTURE FEATURES

Because RNA structure is considered to be a major determinant of IRES activity in known IRESs, we sought to incorporate it in our prediction models. To this end several features describing RNA structure and accessibility were calculated for all sequences and, after applying the same feature pre-selection as in the case of k-mer counts, were used as predictive model features.

RNA ACCESSIBILITY AND REGION INTERACTION

First, we attempted to describe RNA structures in terms of *accessibility* and *region interactions*. To account for local context effects, secondary structures were predicted for sequences with regions flanking them in the reporter construct (Fig. 5.1A). RNA base pairing probabilities were computed using the Vienna RNA package [71] with default settings.

We defined RNA accessibility of a region as the expected number of unpaired nucleotides in this region. The intuition behind this definition is that if a region is highly paired, it is unavailable for interactions with RNA-binding proteins (RBPs) required to initiate translation or with the ribosome itself. This measure was calculated as region length minus the sum of base pairing probability matrix (BPPM) columns corresponding to that region. Similarly, to capture high-level RNA secondary structure, we defined the RNA interaction measure of two regions as the expected number of paired nucleotides between the two regions; and calculated it as the sum of elements in the BPPM located at the intersection of rows corresponding to the first region and columns corresponding to the second region.

RNA accessibility and interactions features were calculated for 10nt moving windows. These features showed weak (Spearman $|\rho| \leq 0.13$), but consistent

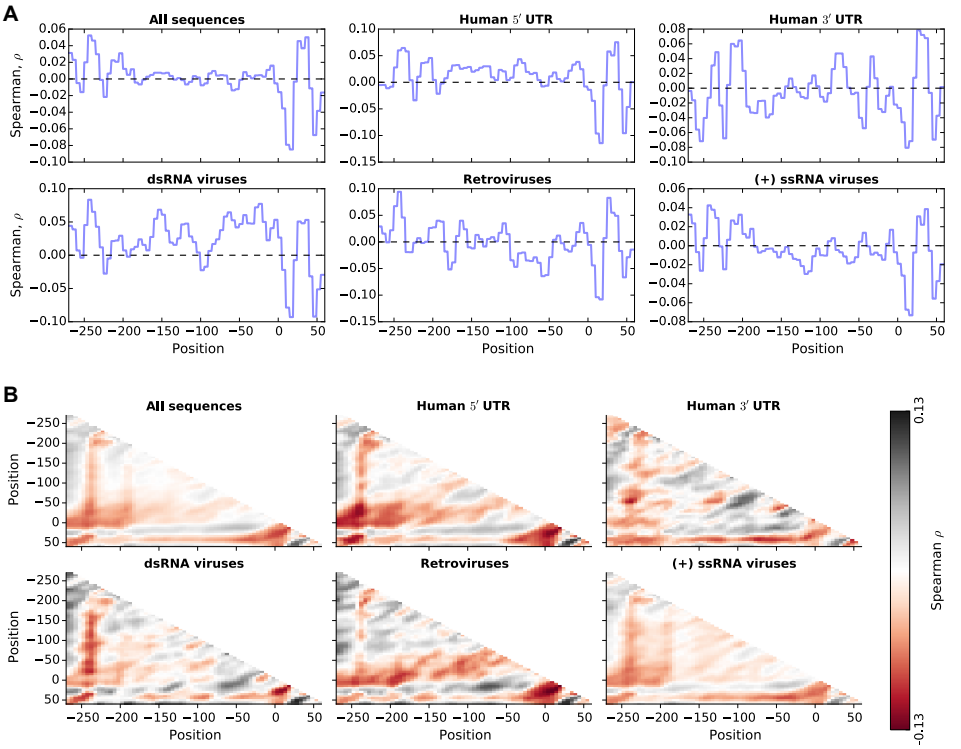

Figure 5.15: (**A**) RNA accessibility correlations for 10nt moving windows with a step of 5nt; and (**B**) RNA region interaction correlations for moving 10nt × 10nt regions with a step of 5nt computed for different sequence groups. Correlations for overlapping windows (regions) were averaged.

correlations across different sequence groups. Specifically, RNA accessibility shows a reproducible pattern of negative-positive-negative correlation with IRES activity in region $[0,50]$ and a similar, although weaker, correlation pattern for region $[-270,-230]$ (Fig. 5.15A). Because we expected that RNA accessibility correlations would be easier to interpret as a product of individual region interactions, we also computed correlations between RNA region interaction features and IRES activity (Fig. 5.15B). Correlation pattens observed for region interaction features suggest that (i) pairing between the region located immediately upstream of position -250 or the region located immediately downstream of position 50 with any other region negatively correlates with IRES activity (predominantly red columns and rows are observed around these positions across all sequence groups); (ii) interactions of regions around the start `AUG` with nearby regions show strongest correlations with IRES activity (as can be readily seen from the dark grey/red spots around the origin for the Retroviruses and Human 5′ UTR sequence groups in Fig. 5.15B). These correlations suggest that the RNA structure formed by the three mentioned regions may play a role in the mechanism of IRES-mediated translation.

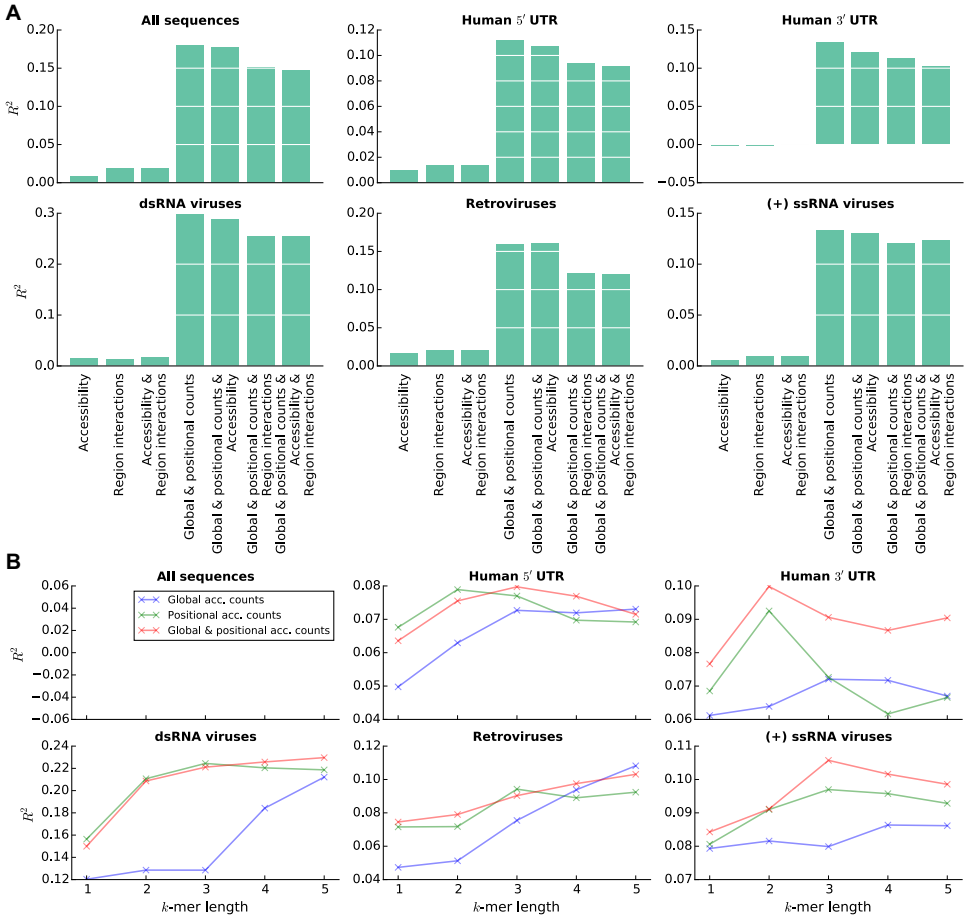

Figure 5.16: Performance for predictors trained with RNA structure. CV performance of models for difference groups of sequences trained on combinations of (**A**) k-mer count for $k = 4$, accessibility and region interaction features; and (**B**) accessible k-mer count features.

Given these observed correlations, we sought to improve our Random Forest models by including RNA accessibility and region interaction features. We followed the same feature pre-selection procedure as described for k-mer features in the main text, and considered different feature combinations, but did not observe any improvement in predictor accuracy beyond what could be achieved using k-mer features alone (see Fig. 5.16A).

ACCESSIBLE k-MER COUNTS

Having observed good predictive power of k-mer features and no improvement in predictor performance when naïvely combining RNA structure or accessibility features with k-mer features, we sought to combine the two feature descriptions in a more

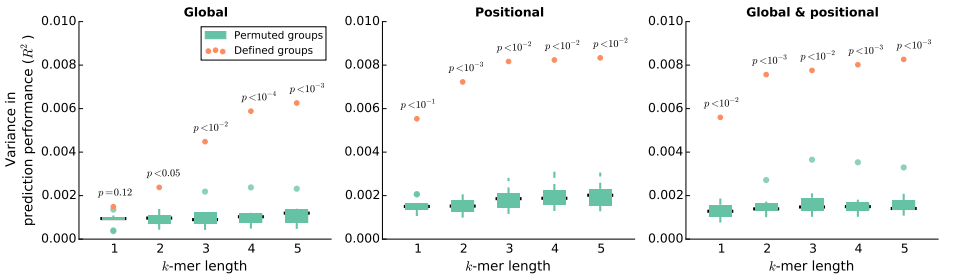

Figure 5.17: Variation in predictive power (R^2) between groups for randomly permuted (green) and defined (orange) groups for different feature combinations (left, middle and right sub-figures) and k-mer lengths. Observed variations in the defined groups are annotated with the corresponding p-values.

in a more direct manner. To this end we modified k-mer count features to produce counts of accessible k-mers by summing k-mer accessibilities instead of occurrences. k-mer accessibilities were calculated as RNA accessibility measurements for regions occupied by k-mer occurrences and normalised by k-mer length. To include accessible k-mer count features in our models, we followed the same feature pre-selection and combination procedure as described for k-mer count features. Unfortunately, as in the case of RNA accessibility and interaction features, we did not observe an increase of model predictive power beyond what can be achieved by k-mer count features alone (see Fig. 5.16B).

GROUP SEQUENCE PERMUTATION

Separation of sequences into $n = 7$ groups based on their species and origin resulted in differences in predictive power between groups. These differences may arise due to group-specific IRES mechanisms being captured by the learned models, or due to group structure (i.e. the number of positive and negative sequences). To see whether the observed variation in the defined groups is higher than the variation one would expect from group structure alone, we performed 10 permutation experiments. In each experiment positive and negative sequences were independently permuted across groups, thus preserving group structure, and models were learned on the permuted groups for each combination of features and k-mer lengths as before. CV predictive power of models learned on the permuted groups were used to obtain 10 samples of variation that can be expected due to group structure alone (green boxplots in Fig. 5.17). These samples were used to arrive at p-values for the variation observed in defined groups (orange dots in Fig. 5.17) by assuming that they follow a scaled χ^2 distribution with $n-1$ degrees of freedom and scaling factor $\frac{n-1}{\sigma^2}$, where σ^2 is the unknown true variance estimated as mean of the 10 permutation variances [72]. Fig. 5.17 shows that variation observed in defined groups is significantly higher than what can be expected due to group structure alone for the majority of feature and k-mer length combinations ($p < 0.05$ for $k > 2$), suggesting that the sequence groups we defined in the main text capture group-specific mechanisms of IRES translation.

FEATURE IMPORTANCE AND PARTIAL DEPENDENCE

When training a RF, tree node variables and splits are chosen to maximise the reduction in weighted variance between the node itself and the two children produced by the node split. Formally, if p, l and r are respectively the current node, and its left and right children; and $S^n_{v,s} = \{(x,y)\}$ are the sets of training samples assigned to nodes $n = p,l,r$ created for feature v and split s, and given as (feature vector, IRES activity) pairs, then feature v and split s are chosen for node p (concisely written as $V(p) = v$ and $S(p) = s$) by maximising

$$C^p = \mathrm{Var}_{(x,y)\in S^p}(y)\cdot|S^p| - \left[\mathrm{Var}_{(x,y)\in S^l}(y)\cdot\left|S^l\right| + \mathrm{Var}_{(x,y)\in S^r}(y)\cdot|S^r|\right],$$

where $\mathrm{Var}_{(x,y)\in S^n}(y)$ gives the variance off all IRES activity values in S^n, and $|S^n|$ gives the number of elements in S^n. Intuitively, the more a variable v is used in the RF trees, and the higher the values C^p are for nodes associated with this variable, the more predictive of IRES activity it is. For our analysis we used *feature importance* as defined in Hastie *et al.* [73], which captures this intuition by accumulating values C^p for all RF trees $t \in T$ and all nodes p assigned to variable v when calculating its importance I_v:

$$I_v = \frac{1}{|T|}\sum_{t\in T}\frac{\sum_{p\in\{p|p\in t\wedge V(p)=v\}} C^p}{\sum_u\left[\sum_{p\in\{p|p\in t\wedge V(p)=u\}} C^p\right]}.$$

These feature importances were additionally normalised by the maximum I_v to allow for comparison of feature importances between models trained on different sequence groups:

$$\tilde{I}_v = \frac{I_v}{\max_u I_u}.$$

A Random Forest $f(x) = f([x^1 \ldots x^M])$ trained on samples $\{(x_j, y_j)|j = 1\ldots N\}$ and M features can be used to investigate the relationship between each its features and the RF prediction. In order to understand the relationship between the i^{th} variable and the prediction $f(x)$ we considered its *partial dependence* on the RF prediction function f, as described in Hastie *et al.* [73]:

$$f_i\left(x^i\right) = \mathbb{E}_{x^i}\left[f\left(x^1,\ldots,x^i,\ldots,x^M\right)\right],$$

which for RFs this function can be efficiently estimated using the training samples x_j as

$$\hat{f}_i\left(x^i\right) = \frac{1}{N}\sum_{j=1}^{N}\left[f\left(x^1_j,\ldots,x^i,\ldots,x^M_j\right)\right].$$

We used the latter estimation in our model interpretation analyses.

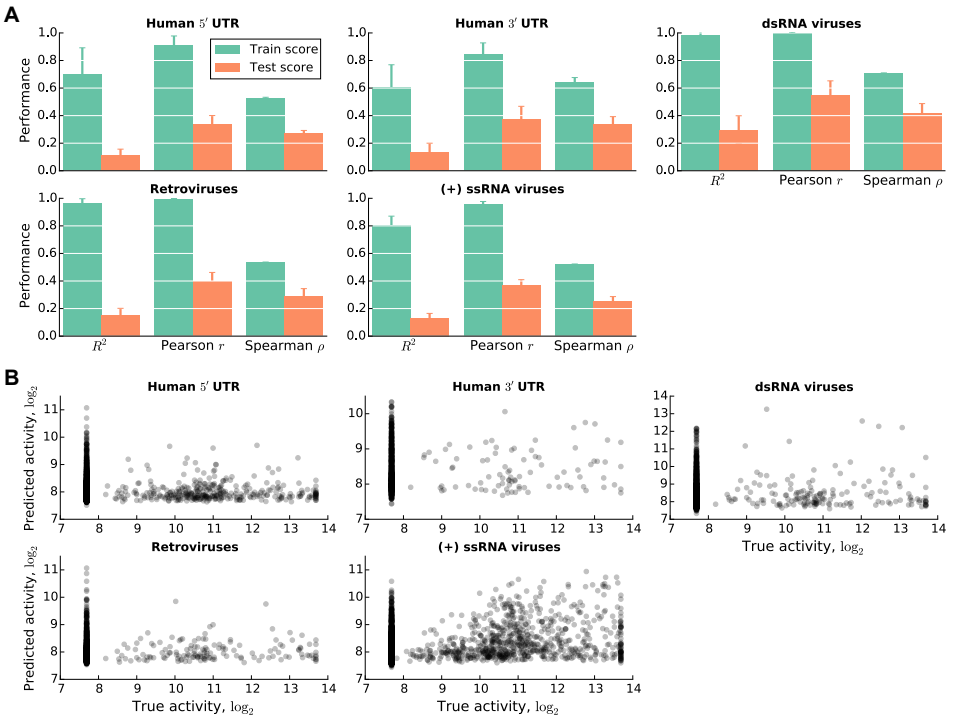

Figure 5.18: (**A**) Training and test performance of selected models for each of the sequence groups. (**B**) Scatter plots of the true vs. predicted IRES activity for these models.

REFERENCES

[1] S. Weingarten-Gabbay, S. Elias-Kirma, R. Nir, A. A. Gritsenko, N. Stern-Ginossar, Z. Yakhini, A. Weinberger, and E. Segal, *Systematic discovery of cap-independent translation sequences in human and viral genomes*, Science **351**, aad4939 (2016).

[2] F. Poulin and N. Sonenberg, *Mechanism of translation initiation in eukaryotes*, Proceedings of the National Academy of Sciences (2000).

[3] M. Bhat, N. Robichaud, L. Hulea, N. Sonenberg, J. Pelletier, and I. Topisirovic, *Targeting the translation machinery in cancer*, Nature Reviews Drug Discovery **14**, 261 (2015).

[4] W. C. Merrick, *Cap-dependent and cap-independent translation in eukaryotic systems*, Gene **332**, 1 (2004).

[5] J. W. Hershey, N. Sonenberg, and M. B. Mathews, *Principles of translational control: an overview*, Cold Spring Harbor Perspectives in Biology **4**, a011528 (2012).

[6] I. N. Shatsky, S. E. Dmitriev, I. M. Terenin, and D. Andreev, *Cap-and IRES-independent scanning mechanism of translation initiation as an alternative to the concept of cellular IRESs*, Molecules and Cells **30**, 285 (2010).

[7] J. Pelletier and N. Sonenberg, *Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA*, Nature **334**, 320 (1988).

[8] M. Mokrejš, T. Mašek, V. Vopálenskỳ, P. Hlubuček, P. Delbos, and M. Pospíšek, *IRESite - a tool for the examination of viral and cellular internal ribosome entry sites*, Nucleic Acids Research **38**, D131 (2010).

[9] P. J. Lukavsky, *Structure and function of HCV IRES domains*, Virus Research **139**, 166 (2009).

[10] A. Brasey, M. Lopez-Lastra, T. Ohlmann, N. Beerens, B. Berkhout, J.-L. Darlix, and N. Sonenberg, *The leader of human immunodeficiency virus type 1 genomic RNA harbors an internal ribosome entry segment that is active during the G2/M phase of the cell cycle*, Journal of Virology **77**, 3939 (2003).

[11] P. S. Ray, R. Grover, and S. Das, *Two internal ribosome entry sites mediate the translation of p53 isoforms*, EMBO Reports **7**, 404 (2006).

[12] M. Holčík, C. Lefebvre, C. Yeh, T. Chow, and R. G. Korneluk, *A new internal-ribosome-entry-site motif potentiates XIAP-mediated cytoprotection*, Nature Cell Biology **1**, 190 (1999).

[13] K. W. Sherrill, M. P. Byrd, M. E. Van Eden, and R. E. Lloyd, *BCL-2 translation is mediated via internal ribosome entry during cell stress*, Journal of Biological Chemistry **279**, 29066 (2004).

[14] M. Holčík and N. Sonenberg, *Translational control in stress and apoptosis,* Nature Reviews Molecular Cell Biology **6**, 318 (2005).

[15] N. Sonenberg and A. G. Hinnebusch, *Regulation of translation initiation in eukaryotes: mechanisms and biological targets,* Cell **136**, 731 (2009).

[16] M. D. Faye and M. Holčík, *The role of IRES trans-acting factors in carcinogenesis,* Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms **1849**, 887 (2015).

[17] X. Du, J. Wang, H. Zhu, L. Rinaldo, K.-M. Lamar, A. C. Palmenberg, C. Hansel, and C. M. Gomez, *Second cistron in CACNA1A gene encodes a transcription factor mediating cerebellar development and SCA6,* Cell **154**, 118 (2013).

[18] S. Cornelis, Y. Bruynooghe, G. Denecker, S. Van Huffel, S. Tinton, and R. Beyaert, *Identification and characterization of a novel cell cycle–regulated internal ribosome entry site,* Molecular Cell **5**, 597 (2000).

[19] C. H. Herbreteau, L. Weill, D. Décimo, D. Prévôt, J.-L. Darlix, B. Sargueil, and T. Ohlmann, *HIV-2 genomic RNA contains a novel type of IRES located downstream of its initiation codon,* Nature Structural & Molecular Biology **12**, 1001 (2005).

[20] M. Candeias, D. Powell, E. Roubalova, S. Apcher, K. Bourougaa, B. Vojtesek, H. Bruzzoni-Giovanelli, and R. Fåhraeus, *Expression of p53 and p53/47 are controlled by alternative mechanisms of messenger RNA translation initiation,* Oncogene **25**, 6936 (2006).

[21] S. Xue, S. Tian, K. Fujii, W. Kladwang, R. Das, and M. Barna, *RNA regulons in Hox* $5'$ *UTRs confer ribosome specificity to gene regulation,* Nature **517**, 33 (2015).

[22] A. B. Sachs, P. Sarnow, and M. W. Hentze, *Starting at the beginning, middle, and end: translation initiation in eukaryotes,* Cell **89**, 831 (1997).

[23] D. A. Costantino, J. S. Pfingsten, R. P. Rambo, and J. S. Kieft, *tRNA-mRNA mimicry drives translation initiation from a viral IRES,* Nature Structural & Molecular Biology **15**, 57 (2008).

[24] L. Balvay, R. S. Rifo, E. P. Ricci, D. Decimo, and T. Ohlmann, *Structural and functional diversity of viral IRESes,* Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms **1789**, 542 (2009).

[25] H. King, L. Cobbold, and A. Willis, *The role of IRES trans-acting factors in regulating translation initiation,* Biochemical Society Transactions **38**, 1581 (2010).

[26] A. A. Komar and M. Hatzoglou, *Cellular IRES-mediated translation: the war of ITAFs in pathophysiological states,* Cell Cycle **10**, 229 (2011).

[27] M. Stoneley and A. E. Willis, *Cellular internal ribosome entry segments: structures, trans-acting factors and regulation of gene expression,* Oncogene **23**, 3200 (2004).

[28] S. A. Mitchell, K. A. Spriggs, M. J. Coldwell, R. J. Jackson, and A. E. Willis, *The Apaf-1 internal ribosome entry segment attains the correct structural conformation for function via interactions with PTB and unr,* Molecular Cell **11**, 757 (2003).

[29] B. M. Pickering, S. A. Mitchell, K. A. Spriggs, M. Stoneley, and A. E. Willis, *Bag-1 internal ribosome entry segment activity is promoted by structural changes mediated by poly (rC) binding protein 1 and recruitment of polypyrimidine tract binding protein 1,* Molecular and Cellular Biology **24**, 5595 (2004).

[30] E. Martínez-Salas, A. Pacheco, P. Serrano, and N. Fernandez, *New insights into internal ribosome entry site elements relevant for viral gene expression,* Journal of General Virology **89**, 611 (2008).

[31] P. Kafasla, N. Morgner, T. A. Pöyry, S. Curry, C. V. Robinson, and R. J. Jackson, *Polypyrimidine tract binding protein stabilizes the encephalomyocarditis virus IRES structure via binding multiple sites in a unique orientation,* Molecular Cell **34**, 556 (2009).

[32] M. Schüler, S. R. Connell, A. Lescoute, J. Giesebrecht, M. Dabrowski, B. Schroeer, T. Mielke, P. A. Penczek, E. Westhof, and C. M. Spahn, *Structure of the ribosome-bound cricket paralysis virus IRES RNA,* Nature Structural & Molecular Biology **13**, 1092 (2006).

[33] M. E. Filbin and J. S. Kieft, *Toward a structural understanding of IRES RNA function,* Current Opinion in Structural Biology **19**, 267 (2009).

[34] E. Martínez-Salas, R. Francisco-Velilla, J. Fernandez-Chamorro, G. Lozano, and R. Diaz-Toledano, *Picornavirus IRES elements: RNA structure and host protein interactions,* Virus Research **206**, 62 (2015).

[35] S. Weingarten-Gabbay and E. Segal, *The grammar of transcriptional regulation,* Human Genetics **133**, 701 (2014).

[36] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, *Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning,* Nature Biotechnology (2015).

[37] A. B. Rosenberg, R. P. Patwardhan, J. Shendure, and G. Seelig, *Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences,* Cell **163**, 698 (2015).

[38] M. A. Cleary, K. Kilian, Y. Wang, J. Bradshaw, G. Cavet, W. Ge, A. Kulkarni, P. J. Paddison, K. Chang, N. Sheth, *et al., Production of complex nucleic acid libraries using highly parallel in situ oligonucleotide synthesis,* Nature Methods **1**, 241 (2004).

[39] E. M. LeProust, B. J. Peck, K. Spirin, H. B. McCuen, B. Moore, E. Namsaraev, and M. H. Caruthers, *Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process,* Nucleic Acids Research **38**, 2522 (2010).

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

[41] X. Cheng, N. Virk, W. Chen, S. Ji, S. Ji, Y. Sun, and X. Wu, *CpG usage in RNA viruses: data and hypotheses,* PLoS One (2013).

[42] M. S. Benleulmi, J. Matysiak, D. R. Henriquez, C. Vaillant, P. Lesbats, C. Calmels, M. Naughtin, O. Leon, A. M. Skalka, M. Ruff, *et al., Intasome architecture and chromatin density modulate retroviral integration into nucleosome,* Retrovirology **12**, 13 (2015).

[43] G. Hernandez, *Was the initiation of translation in early eukaryotes IRES-driven?* Trends in Biochemical Sciences **33**, 58 (2008).

[44] L. Zhang, S. Kasif, C. R. Cantor, and N. E. Broude, *GC/AT-content spikes as genomic punctuation marks,* Proceedings of the National Academy of Sciences of the United States of America **101**, 16855 (2004).

[45] R. C. Spitale, R. A. Flynn, Q. C. Zhang, P. Crisalli, B. Lee, J.-W. Jung, H. Y. Kuchelmeister, P. J. Batista, E. A. Torre, E. T. Kool, *et al., Structural imprints in vivo decode RNA regulatory mechanisms,* Nature (2015).

[46] J. H. Friedman, *Stochastic gradient boosting,* Computational Statistics & Data Analysis **38**, 367 (2002).

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine learning in Python,* Journal of Machine Learning Research **12**, 2825 (2011).

[48] R. Tibshirani, *Regression shrinkage and selection via the LASSO,* Journal of the Royal Statistical Society. Series B (Methodological) , 267 (1996).

[49] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, *1-norm Support Vector Machines,* Advances in Neural Information Processing Systems **16**, 49 (2004).

[50] J. H. Friedman, *Greedy function approximation: a gradient boosting machine,* Annals of Statistics , 1189 (2001).

[51] V. Zeenko and D. R. Gallie, *Cap-independent translation of tobacco etch virus is conferred by an RNA pseudoknot in the* 5′*-leader,* Journal of Biological Chemistry **280**, 26813 (2005).

[52] R. Nicholson, J. Pelletier, S. Le, and N. Sonenberg, *Structural and functional analysis of the ribosome landing pad of poliovirus type 2: in vivo translation studies.* Journal of Virology **65**, 5886 (1991).

[53] B. T. Baranick, N. A. Lemp, J. Nagashima, K. Hiraoka, N. Kasahara, and C. R. Logg, *Splicing mediates the activity of four putative cellular internal ribosome entry sites,* Proceedings of the National Academy of Sciences **105**, 4733 (2008).

[54] S. A. Chappell, G. M. Edelman, and V. P. Mauro, *A 9-nt segment of a cellular mRNA can function as an internal ribosome entry site (IRES) and when present in linked multiple copies greatly enhances IRES activity,* Proceedings of the National Academy of Sciences **97**, 1536 (2000).

[55] S. A. Mitchell, K. A. Spriggs, M. Bushell, J. R. Evans, M. Stoneley, J. P. Le Quesne, R. V. Spriggs, and A. E. Willis, *Identification of a motif that mediates polypyrimidine tract-binding protein-dependent internal ribosome entry,* Genes & Development **19**, 1556 (2005).

[56] M. Görlach, C. G. Burd, and G. Dreyfuss, *The determinants of RNA-binding specificity of the heterogeneous nuclear ribonucleoprotein C proteins.* Journal of Biological Chemistry **269**, 23074 (1994).

[57] M. Holčík, B. W. Gordon, and R. G. Korneluk, *The internal ribosome entry site-mediated translation of antiapoptotic protein XIAP is modulated by the heterogeneous nuclear ribonucleoproteins C1 and C2,* Molecular and Cellular Biology **23**, 280 (2003).

[58] R. A. Flynn, L. Martin, R. C. Spitale, B. T. Do, S. M. Sagan, B. Zarnegar, K. Qu, P. A. Khavari, S. R. Quake, P. Sarnow, *et al., Dissecting noncoding and pathogen RNA–protein interactomes,* RNA **21**, 135 (2015).

[59] K. Choi, J. H. Kim, X. Li, K. Y. Paek, S. H. Ha, S. H. Ryu, E. Wimmer, and S. K. Jang, *Identification of cellular proteins enhancing activities of internal ribosomal entry sites by competition with oligodeoxynucleotides,* Nucleic Acids Research **32**, 1308 (2004).

[60] L. Wang, K.-S. Jeng, and M. M. Lai, *Poly (C)-binding protein 2 interacts with sequences required for viral replication in the hepatitis C virus (HCV)* 5′ *untranslated region and directs HCV RNA replication through circularizing the viral genome,* Journal of Virology **85**, 7954 (2011).

[61] J. R. Evans, S. A. Mitchell, K. A. Spriggs, J. Ostrowski, K. Bomsztyk, D. Ostarek, and A. E. Willis, *Members of the poly (rC) binding protein family stimulate the activity of the c-myc internal ribosome entry segment in vitro and in vivo,* Oncogene **22**, 8012 (2003).

[62] Y. Xue, Y. Zhou, T. Wu, T. Zhu, X. Ji, Y.-S. Kwon, C. Zhang, G. Yeo, D. L. Black, H. Sun, *et al., Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping,* Molecular Cell **36**, 996 (2009).

[63] S. R. Thompson, *So you want to know if your message has an IRES?* Wiley Interdisciplinary Reviews: RNA **3**, 697 (2012).

[64] M. Stoneley, T. Subkhankulova, J. P. Le Quesne, M. J. Coldwell, C. L. Jopling, G. J. Belsham, and A. E. Willis, *Analysis of the c-myc IRES; a potential role for cell-type specific trans-acting factors and the nuclear compartment,* Nucleic Acids Research **28**, 687 (2000).

[65] B. L. Semler and M. L. Waterman, *IRES-mediated pathways to polysomes: nuclear versus cytoplasmic routes,* Trends in Microbiology **16**, 1 (2008).

[66] P. J. Lukavsky, I. Kim, G. A. Otto, and J. D. Puglisi, *Structure of HCV IRES domain II determined by NMR,* Nature Structural & Molecular Biology **10**, 1033 (2003).

[67] M. Majumder, I. Yaman, F. Gaccioli, V. V. Zeenko, C. Wang, M. G. Caprara, R. C. Venema, A. A. Komar, M. D. Snider, and M. Hatzoglou, *The hnRNA-binding proteins hnRNP L and PTB are required for efficient translation of the Cat-1 arginine/lysine transporter mRNA during amino acid starvation,* Molecular and Cellular Biology **29**, 2899 (2009).

[68] R. A. Flynn, Q. C. Zhang, R. C. Spitale, B. Lee, M. R. Mumbach, and H. Y. Chang, *Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE,* Nature Protocols **11**, 273 (2016).

[69] V. Vacic, L. M. Iakoucheva, and P. Radivojac, *Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments,* Bioinformatics **22**, 1536 (2006).

[70] A. Riley, L. E. Jordan, and M. Holčík, *Distinct* 5′ *UTRs regulate XIAP expression under normal growth conditions and during cellular stress,* Nucleic Acids Research **38**, 4665 (2010).

[71] R. Lorenz, S. H. Bernhart, C. H. Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, I. L. Hofacker, *et al., ViennaRNA Package 2.0.* Algorithms for Molecular Biology **6**, 26 (2011).

[72] K. Knight, *Mathematical Statistics,* 1st ed., Chapman & Hall/CRC Texts in Statistical Science (Chapman and Hall/CRC, 1999).

[73] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, *The elements of statistical learning: data mining, inference and prediction,* The Mathematical Intelligencer **27**, 83 (2005).

6

DISCUSSION

The marked drop in the costs of DNA sequencing and programmable DNA synthesis made it possible for individual labs to (re-)design DNA sequences in a targeted manner, and prompted novel applications in biotechnology, metabolic engineering (ME) and synthetic biology (SB). This thesis presented several algorithms for the analysis of second generation sequencing data from the above disciplines. It includes methods developed for improving microbial genome assemblies through scaffolding (biotechnology), for optimising protein expression by synonymously re-writing genes (ME and SB), for whole-cell modelling of protein translation (ME and molecular biology; MB), and for uncovering RNA sequence features governing IRES-mediated translation (MB). These methods were motivated by open biological questions that can directly affect industrial applications, and aim at bridging the gap between basic and translational research.

Below, we discuss current challenges in sequencing data analysis (Section 6.1) and translation modelling (Section 6.2) addressed in this thesis, followed by an outlook into the future opportunities for these fields (Section 6.3).

6.1. CHALLENGES IN GENOME SCAFFOLDING

The availability of a high-quality reference genome sequence is a prerequisite for genetic engineering and synthetic biology techniques employed in metabolic engineering. Good reference genome sequences of the studied organism (i.e. same strain) provide a better backbone for genetic engineering. Decreasing costs of second generation sequencing made it an attractive technology for *de novo* sequencing and re-sequencing of host organisms used in biotechnology. Our association with the *de novo* sequencing of the *Saccharomyces cerevisiae* laboratory yeast strain (CEN.PK 113-7D; Nijkamp *et al.* [1]) prompted us to develop the GRASS assembly scaffolder (described in Chapter 2), which can use existing reference sequences and paired read sequencing information to improve fragmented microbial genome assemblies.

IMPROVED OPTIMISATION STRATEGIES

GRASS is based on a strict mathematical formulation that combines the contig order, distance and orientation constraints in a single optimisation objective. However, the resulting optimisation problem is computationally intractable and solving it requires the use of approximation techniques. In Chapter 2 we described an expectation-maximisation (EM) strategy that can be used to address this optimisation problem. However, our approach has several limitations: first, the EM strategy uses stochastic algorithms for solving some of the sub-problems within the EM approach. This leads to undesired variation in the optimisation results, and complicates scaffolder use within larger pipelines. Second, due to the complexity of the EM approach, it is unclear whether GRASS can be used to produce competitive results on larger (e.g. plant or mammalian) genome assemblies. Third, the approximate optimisation strategy does not allow for estimating the gap between the current solution and the unknown optimal solution, which complicates solution quality assessment.

Although further research is required to overcome these limitations, they could potentially be addressed by alternative optimisation strategies that (i) provide solution quality guarantees, and (ii) do not rely on stochastic optimisation. Generally, the trade off between a mathematically rigorous problem description and the ability to solve the resulting problem should be considered during the design of scaffolding algorithms to avoid similar limitations in the future.

MODULAR INFORMATION EXTRACTORS

The mathematical formulation employed by GRASS allowed it to use a wide range of information sources (e.g. paired reads, restriction maps, related genomes) for scaffolding. However, its practical application as a generic scaffolder was limited by the ability of its *linker* module implementation to derive contig links from the various information sources. The pace of technology development makes keeping an up-to-date version of the scaffolder targeting a variety of information sources impractical without a dedicated team. This challenge can be partially alleviated by sharing this responsibility with the research community that uses the scaffolder, which could be achieved by providing an implementation and instructions that allow for easy addition of new modules for contig link generation.

Alternatively, this goal could be achieved by standardising the different scaffolding steps (e.g. contig linking and scaffold optimisation), as already done for genome assembly within the AMOS (A Modular, Open-Source whole genome assembler; Pop *et al.* [2]) consortium. The mathematical description of contig links provided in Chapter 2 can serve as a starting point for a generic description of the contig relationship information, and for the development of a standardised storage format of that information. Independent modules can be used to process the individual information sources that, after a merging step, would be fed to one of the interchangeable scaffolding modules. Such an approach would not only increase the power of existing scaffolding algorithms, but would additionally allow for their direct comparison and easy incorporation into larger genome sequencing pipelines.

6.2. CHALLENGES IN TRANSLATION MODELLING

Despite the limited understanding of the mechanisms behind it, codon optimisation (CO) is routinely employed to increase heterologous expression of genes in new host organisms. In Chapter 3 we described the development of a predictive model for yeast *Saccharomyces cerevisiae* that combined existing sequence features typically used for CO, and which could be used to guide CO. In an *addendum* to Chapter 3 we described an attempt to codon optimise the *Arabidopsis thaliana PAL1* gene for expression in yeast using this model.

This result led us to believe that the process of protein synthesis is too complex to be tackled by a naïve data-driven approach, and requires a different type of modelling before directing codon optimisation. The complexity of CO is further aggravated by the fact that translation contains numerous recently discovered and debated stages and mechanisms that do not fit within the simplistic three-stage model of translation (see Chapter 1). These include:

- ribosome recycling, in which a ribosome in the translation termination stage can efficiently re-initiate translation of the same mRNA [3, 4];
- tRNA re-use, which presumably allows for efficiently (quickly) reusing tRNA molecules that were used to translate one of the recent codons [5];
- translation of upstream reading frames (uORFs) in the 5′ UTR that impacts initiation efficiency of the main coding sequence [6, 7];
- initiation of translation using mechanisms that do not rely on the 5′ cap structure or the scanning for a start codon, such as IRES-mediated translation [8, 9], and translation via the 3′- and 5′-CITEs (Cap-Independent Translation Elements; Miller *et al.* [10], Shatsky *et al.* [11]);
- inhibition of translation through miRNA targeting [12, 13];
- stop codon read-through [14, 15];
- reading frame shifting during the elongation phase [16];

and many others. Currently, no unified models covering all of the above mechanisms are available. However, models of varying detail addressing individual mechanisms or combinations thereof, have been proposed in this thesis (Chapters 4-5) and elsewhere.

WHOLE-CELL MODELS OF TRANSLATION

Some of the central cellular processes, such as stress response or mRNA localisation [17] are regulated translationally, with evidence suggesting that dysregulation of translation may lead to disease [18, 19]. Modelling of global effects of translation regulation does not require detailed description of all of the constituent translation mechanisms, and can be captured by high-level models addressing the three phases of translation. Difficulties in measuring translation kinetics make it challenging for existing models to accurately model the underlying process. The recent advent of ribosome profiling (RP; Ingolia [20]) measurements provided the information necessary for constructing translation models with data-derived parameters. However, construction of such models requires reconciling the noisy RP data with stochastic simulation techniques. In Chapter 4 we described the first method addressing this challenge. We analysed RP data at several scales to capture as much information as the noisy low-coverage measurements

would allow, and developed a statistical framework for comparing stochastic simulations to multi-scale measurements, which together allowed us to fit whole-cell models of translation to *Saccharomyces cerevisiae* yeast RP data that compared favourably to exiting models.

HIGH-RESOLUTION ACCURATE DATA

Low-coverage unreliable RP data makes capturing changes in ribosome density along genes challenging. To address this, in Chapter 4 we devised a multi-scale representation of this data, which pooled reads for more reliable estimates by considering the average ribosome density of gene regions at resolutions determined by the local coverage depth. While necessary for processing low-coverage data, this approach potentially limits the model fitting procedures sensitivity to changes in global translation parameters. This limitation could be overcome by fitting models to high-coverage RP data that allows for reliable single-codon measurements.

However, recent studies showed that even high coverage data may not be suitable for single codon level analyses [21–23]. It was suggested that differences in experimental protocols, such as the pre-treatment of cells using cycloheximide or its used concentration, may prevent translation elongation from halting immediately on some codons. Continued elongation "smears" codon-specific signals across several downstream codons, thus reducing measurement accuracy. While it may be possible to develop bioinformatic algorithms for deconvolving smeared signals (e.g. by fitting Gaussian mixture models with codon-specific parameters onto the RP data), a better approach would be to generate data that does not require deconvolution by eliminating the cycloheximide pre-treatment step and carrying out RP measurements in cryogenic conditions.

Nevertheless, given the evident sensitivity of RP to protocol-specific conditions, even with the necessary protocol adjustments, quality control should become a routine step of RP. It should at least include checking for continued translation elongation, but could also involve Ribo-seq spike-ins from well-characterised bacterial species or *in vitro* translation systems [24].

DATA NORMALISATION

It has been shown that Ribo-seq and mRNA-seq measurements are subject to sequence (e.g. pausing at certain codons) and sequencing biases, which affect local read depth of transcripts. While sequence biases represent the sought signal, the sequencing biases, PCR amplification, RNA digestions, or other readout biases may overshadow the sought signal. To remove the effect of sequencing biases, we averaged ribosome and mRNA read density across longer regions in Chapter 4. While this approach was justified for low coverage datasets, alternative techniques are required for analysing single-codon resolution data. These approaches will either have to manually classify the detected signals into sequencing biases (e.g. signals related to PCR amplification and endonuclease digestion biases likely to be located towards the ends of sequencing reads), or require the development of normalisation methods based on the assumption of shared biases between the Ribo-seq and mRNA-seq reads included in RP measurements [21]. Alternatively, RP measurements could be normalised using

spike-ins of reference RNA added to the Ribo-seq and mRNA-seq samples, as was previously proposed for RNA-seq data [25–27].

RIBOSOME CONFORMATIONS

Existing RP data analysis pipelines, including the one described in Chapter 4, are typically restricted to analysis of the canonical ≈ 28nt ribosome protected fragments. It was suggested that different fragment lengths correspond to distinct elongating ribosome conformations [28] that can be stabilised by different chemicals. It is possible that the choice of chemical (and thus conformation) creates biases in single-codon RP data analyses by allowing ribosomes to transition from their current conformation to the conformation they can be stabilised in. Such biases would effect downstream single-codon analyses of the data and should be avoided whenever possible. It would be interesting to augment ribosome profiling experiments by employing simultaneous treatment using multiple chemicals to stabilise translating ribosomes in their current conformation. Footprint length differences produced by the augmented RP should allow for separating the different conformations, and thus provide the data necessary for modelling this phenomenon *in silico*.

ALTERNATIVELY SPLICED TRANSCRIPTS

The models introduced in Chapter 4 do not support alternatively spliced transcripts. This is not a significant limitation for organisms that do not extensively use mRNA splicing (such as the bakers yeast *Saccharomyces cerevisiae*; Ast [29]). However, simulation of alternatively spliced transcripts of the same gene is necessary for modelling translation in higher eukaryotes, such as human or the fruit fly.

Currently, no whole-cell protein synthesis models exist that support alternatively spliced transcripts. Simulation of such models would only require treating such transcript isoforms as independent molecules that share the gene-specific initiation rate parameter. However, fitting such models would additionally require algorithmic improvements that address the following points:

- Comparison of simulated density from transcript isoforms to the same measured mRNA-seq and Ribo-seq profiles in a way that accounts for isoform abundances and inclusion/exclusion of introns and exons.
- Fitting of the gene-specific initiation rates shared by all isoforms, which may hamper the use of the initiation rate approximation scheme we introduced in Chapter 4.

SHARED MOLECULE POOL

Over-expression of tRNAs or heterologous genes [30], hijacking of the cells gene expression machinery by viral infections [31], or subjection of cells to stress or different growth conditions [32], are examples of events that can dramatically change the numbers of available cellular molecules involved in translation (e.g. ribosomes, mRNA transcripts or tRNAs), thus also changing the kinetics of protein synthesis. Models from Chapter 4 were fitted on the RP data measured in normal conditions, and thus do not generalise to situations with significantly different quantities of translation molecules.

Models that describe transcript abundances and keep track of the (un)bound ribosomes and tRNAs (e.g. SMoPT, Shah *et al.* [30]), can adjust translation initiation and elongation rates in accordance with these quantities, and consequently can be used to model translation in unseen cellular conditions. However, the use of global quantities creates inter-dependencies between modelled mRNAs, and thereby prevents their efficient simulation through parallel mRNA simulations. Since fitting of the models from Chapter 4 was already computationally challenging, it is likely that, if approached naïvely, fitting of models with global quantities onto the RP data would be computationally intractable.

While fitting models with global quantities requires further research, it could be tackled by separating the model simulation into two stages: (i) a stage in which the steady-state values of the global quantities of the (un)bound ribosomes and tRNAs are found; and (ii) a stage in which these values are used to independently simulate model transcripts as in Chapter 4. Specifically, the steady-state values from stage (i) could be efficiently obtained by simulating an *equivalent* coarse-grained version of the original model that describes molecule pool states, but is not concerned with exact locations of ribosomes or tRNAs on transcripts.

MODELS OF CAP-INDEPENDENT TRANSLATION INITIATION

Although currently available knowledge and data do not allow for the construction of unified translation models encompassing the entirety of translation mechanisms, some of these mechanisms could be modelled using measurements obtained from specifically designed functional genomics assays. In Weingarten-Gabbay *et al.* [33] we developed an IRES activity assay (see also Chapter 1) that we used to measure the ability and strength of 55,000 sequences to initiate translation in an IRES-dependent manner. In Chapter 5 we described the development of Random Forest (RF) models predicting IRES activity from RNA sequence, which uncovered sequence determinants of IRES translation.

MODEL INTERPRETATION AND EXPERIMENTAL VALIDATION

Sequence models of IRES activity described in Chapter 5 are based on the k-mer feature description of the RNA sequence. By interpreting these models we found that k-mer features associated with increased IRES activity resembled (partial) ITAF binding motifs. This association is in line with the proposed role of ITAFs as RNA-binding proteins involved in IRES-mediated translation, and suggests that their binding is a part of the IRES mechanism employed by the analysed sequences. However, additional research is required to confirm this.

This hypothesis could be further tested by constructing a sequence model of their activity that is based on the description of these sequences in terms of their similarity to known ITAF binding motifs [34, 35]. Known binding motifs of ITAFs and other RNA binding proteins (RBP) can be used to detect and score potential binding sites in RNA sequences; these scores could then be used to devise a feature description of the sequences that is used to learn models of IRES activity as described in Chapter 5. Combined with model feature selection, this approach can then be employed to further test whether a predictive relationship between ITAF binding sites and IRES activity exists.

However, a model based on RBP binding motifs still would not prove that (i) ITAFs

indeed bind to the studied RNA sequences, and that (ii) their binding influences IRES activity. Additional validation experiments are required to confirm these hypotheses. The former hypothesis could be validated by performing a CLIP-seq assay [36] that reads out the unique barcodes present in the IRES sequences analysed in Chapter 5, whereas the latter hypothesis would require mutating the predicted (or identified via CLIP-seq) binding sites in these sequences and measuring the effect on IRES activity.

RNA STRUCTURE

IRESs are believed to execute their function using a combination of specific RNA primary sequence and secondary structure, which makes it surprising that we were unable to find a strong RNA structure signal predictive of IRES activity. In Chapter 5 we discussed the possible reasons for this outcome, and proposed additional experiments that would help uncovering the role of RNA structure in IRES translation. Here, we propose computational approaches that could be used to detect a relationship between IRES activity and RNA structure.

To our knowledge only two studies [37, 38] previously reported the ability to predict IRESs. Both studies are based on the idea of comparing predicted RNA structures to a small number of known IRES RNA structures. While it is unclear whether the results reported in these studies are based on independent validation sets, and whether the predictive performance estimates obtained using small datasets are reliable, it would be interesting to extend this idea to the large set of IRESs measured by Weingarten-Gabbay *et al.* [33].

Largely, the above approach could be seen as a special case of describing RNA sequences or structures in terms of their similarity to a target set of structures, and using the results of this similarity to classify the original RNA sequence as an IRES or not. This general description suggests a straightforward integration of RNA structure into the analysis pipeline described in Chapter 5, in which structure similarity scores would be used as additional features of the RF models. Further, several target sets and similarity scores could be used:

- The set of RNA structures of known IRESs (e.g. obtained from Rfam [39]) with an RNA structure alignment-based similarity as in Wu *et al.* [37], Hong *et al.* [38];
- The set of known atomic resolution IRES 3D structures [40] with a similarity measure based on homology modelling (i.e. approximation of the unknown 3D structure based on sequence similarity to RNAs with known structures; Saxena *et al.* [41]) of the new RNA sequences;
- The set of motifs from the RNA 3D Motif Atlas [42] with a similarity measure provided by JAR3D [43] or a similar tool.

The latter *target set - similarity* combination is particularly interesting, as it was recently reported [44] to improve binding affinity prediction of the PTBP (Pyrimidine Tract Binding Protein), a known ITAF whose involvement in IRES activity we also suggested in Chapter 5.

CODON OPTIMISATION

Chapter 4 contains several results important for codon optimisation. First, our finding that predictions made by existing translation models are heavily biased by mRNA

transcript levels suggests that the CO model from Chapter 3 may have suffered from the same limitations, especially since we chose to use protein abundance and unnormalised ribosome density as prediction targets. This could be remedied by using alternative targets, as discussed in the *addendum* to Chapter 3.

Second, the observation that protein production is limited at initiation rather than elongation (presumably targeted by CO), suggests that stronger (over-)expression effects could be gained by optimising sequences that affect the initiation step. This could be achieved by constructing sequence models of the 5′ UTR or through the incorporation of alternative initiation mechanisms (e.g. IRES) into the optimised sequence. The former would ideally allow for detecting and quantifying the relationship between the sequence around the initiation site and a quantitative measure of translation initiation (e.g. as obtained by FACS-seq, GTI-seq or QTI-seq; Noderer *et al.* [45], Lee *et al.* [46], Gao *et al.* [47]).

Finally, the context-dependence of codon elongation times suggests that existing codon adaption indices (e.g. the CAI or tAI) do not adequately capture the mechanisms complexity, and that more explicit models (e.g. models from Chapter 4 that take codon context into account) would be more suitable for CO.

TAKING A STEP BACK

The complex nature of gene regulation, the results discussed above, and known cases of failed CO attempts [48, 49] raise the question of whether translation elongation is indeed the aspect of translation that is improved through CO. Synonymous changes in the CDS can alternatively affect transcript stability, translation initiation or protein folding, which would all manifest themselves in the expression or activity of the protein in question. The lack of a clear-cut selection of mechanisms affected by CO, their complex interplay, and the sheer size of the space of possible synonymous versions of the same gene, all presumably contribute to making the CO approach pioneered by DNA 2.0 (Menlo Park, CA) successful. In their approach [50] multiple gene re-designs are considered, and assayed to be used for construction of predictive models employed to design the next batch of re-designs. This process is repeated until a re-design with the sought properties is obtained.

While being considerably more time-consuming and cost-intensive than a single-shot optimisation approach, the iterative strategy of DNA 2.0 allows for learning the protein- and condition-specific net effect of the various translation mechanisms, without untangling them or modelling their individual effects. It is likely that in the near future, until a better understanding of translation emerges, this CO approach will remain most reliable.

CONDITION SPECIFICITY

Although protein expression levels are determined by the net effect of multiple regulation mechanisms, it is likely that under some conditions this effect is dominated by a single mechanism. For example, under the conditions of strong transcript over-expression, small differences in its codon bias can lead to large differences in the resulting protein expression [30]. In those conditions, optimisation techniques targeting the limiting aspect of translation could be successful, whereas for the conditions where

no one mechanism is dominant, the gene- and condition-specific iterative approach discussed above would be better suited.

Conditions (e.g. medium, cell type, transcript expression levels) appropriate for using a particular optimisation approach can be identified by performing simulations of whole-cell translation models with shared molecule pools. More generally, a strategy that encompasses translation bottleneck identification through whole-cell modelling, followed by bottleneck-specific gene optimisation, could provide the necessary intermediate step between the currently employed mechanism-unaware CO approaches and future approaches, that would likely rely on accurate modelling of the entire gene expression process.

6.3. TECHNOLOGICAL INNOVATION AS A DRIVER OF BIOLOGICAL RESEARCH

The past decade brought about several disruptive technological advancements and methodological innovations in measurement and genetic engineering techniques, such as high-throughput sequencing or genome editing using engineered nucleases [51]. These techniques revolutionised biological research by providing the means to interrogate previously inaccessible aspects of living organisms. Today, continuing maturation of established technologies, and improvement of cutting-edge techniques, such as single molecule sequencing or the CRISPR-Cas system [52], are creating new opportunities in basic research and prompting their application in metabolic engineering and synthetic biology.

THIRD GENERATION SEQUENCING TECHNOLOGIES

Current third generation sequencing technologies, consisting of the Pacific Biosciences (Menlo Park, CA) and Oxford Nanopore Technologies (Oxford, UK) single molecule sequencing platforms, generate reads that are tens of thousands of nucleotides long [53, 54]. Their long reads allow sequencing through repeat regions that would be problematic for second and first generation sequencing platforms, making 3rd generation technologies particularly suited for *de novo* genome sequencing. Similarly, by sequencing through RNA splice junctions, long reads also alleviate problems associated with transcript assembly and isoform identification in mRNA sequencing [55].

WHOLE-GENOME SEQUENCING AND SCAFFOLDING

Data from these technologies can be used to create near-finished *de novo* genome assemblies with contigs spanning entire chromosomes, thus making the problem of genome scaffolding, and methods like GRASS (Chapter 2), obsolete for future whole-genome sequencing projects of low complexity genomes. However, due to the high error rate of these platforms (20% to 38%; Ross *et al.* [53], Laver *et al.* [54]), the costs associated with generating enough coverage depth to create such assemblies are still prohibitively high for most labs, as for consortia sequencing large cohorts (e.g. clinical trials or population studies). So for the foreseeable future, until the sequencing cost and accuracy of these platforms improve, hybrid strategies that combine the low cost high-throughput 2nd generation sequencing with long read 3rd generation sequencing,

are likely to be favoured. In these strategies short accurate reads are used either alone or in combination with long reads to create contigs that are then scaffolded using the (very) long 3rd generation reads in a post-processing step. Scaffolding using long reads can also be used to improve genomes previously assembled from shorter reads. However, there are several challenges specific to the use of 3rd sequencing reads for scaffolding:

- Correctly using constraints that span multiple contigs (several contigs aligning to the same read), and should be simultaneously "on" or "off" for all of them;
- Ensuring that overhanging and overlapping long reads in the scaffold agree on distance and sequence;
- Handling of non-unique or uncertain alignments of contigs to error-prone long reads;
- Handling long chimeric reads that falsely join different DNA molecules.

Solving these challenges will require re-visiting the formulation of the contig scaffolding problem, and innovation in the computational approaches addressing it.

However, it is likely that for some genomes scaffolding will remain necessary for obtaining finished genomes even after the cost and accuracy of 3rd generation improve. Even the long reads of these technologies are unable to resolve megabase-long segment duplications present in some genomes [56], and require scaffolding using additional information, such as optical mapping data [57], to resolve such repeats. Interestingly, current 3rd generation sequencing read lengths are limited by the size of input DNA molecules rather than the physical characteristics of the instruments [54]. It is thus possible that new versions of instrument chemistry and library preparation protocols will further alleviate the necessity of downstream assembly scaffolding.

Overall, it can be expected that ongoing improvements in DNA sequencing technologies will relieve most *de novo* assembly challenges and allow for accurate chromosome-length genome assemblies necessary for genetic and metabolic engineering efforts.

TRANSLATION NOISE AND TRANSCRIPT HETEROGENEITY

Third generation sequencing technologies allow for single-read transcript sequencing, which simplifies alternative splicing analyses, and allows for mapping noisy transcription start and stop sites. For the moment, sequencing costs prevent generating enough sequencing data to allow for quantitative analyses of these effects. However, once the sequencing costs drop, deep single-molecule sequencing could be used to obtain quantitative isoform measurements.

A recent study used deep second generation sequencing of *polysome fractions*[1] to reconstruct transcript isoforms for each of the fractions [58]. This approach, if combined with quantitative single-molecule sequencing, could allow for studying the quantitative relationship between mRNA isoforms and translation. Unlike ribosome profiling, which gives the average ribosome density across all transcripts of the same gene, this approach would also allow for quantifying translation noise from the distribution of isoforms across polysome fractions; and for studying its determinants, as was previously done for transcriptional noise [59, 60]. Ultimately, data and insights learned from these studies

[1] mRNA molecules with the same number of ribosomes bound to them.

should facilitate the creation the next generation of translation models described in Chapter 4, which may need to model heterogenous families of mRNA transcripts of the same gene with transcript-specific initiation rates.

As accuracy of third generation sequencing improves, single-read transcript sequencing could be used to devise an assay analogous to the CLIP-seq and PAR-CLIP techniques [36, 61], in which transcript positions bound by ribosomes would be modified or mutated in a way that is detectable with subsequent single-molecule sequencing. Such an assay would yield locations of ribosomes bound to the same transcript, thus providing the measurements required for studying translation heterogeneity. This data would allow for further improving translation models by constraining the kind of situations that may occur on simulated transcripts. Alternatively, techniques for studying translation heterogeneity could come from advances in single-cell second generation sequencing approaches. For example, a single-cell equivalent of ribosome profiling would allow for mapping cell-to-cell variability of gene density profiles and, thus, for studying dynamics of translation regulation.

Irrespective of the assay or technology that will be used to study translation noise, it is likely that these data will itself be noisy (e.g. missing ribosomes due to failed cross-linking) and sparse. Their analysis will require the development of statistical methods that are robust to missing and noisy data, which could be based on the idea of pooling sparse single-transcript or single-cell data to tune statistical model parameters, and then using the tuned models to analyse single cells and transcripts [62].

SYSTEMATIC STUDIES OF GENE REGULATION

Gene expression regulation is a highly multi-factorial process, whose individual aspects can be difficult to tease apart due to their influence on each other. For example, sequence features that promote mRNAs translation initiation may also extend mRNA half-life by making it less accessible to degradation machinery through increased ribosome occupancy [63]. This regulatory crosstalk makes studying determinants of gene expression from native genes in their native context very challenging. Systematic approaches that vary a single aspect of the expression machinery while keeping all other aspects constant have proven instrumental for uncovering determinants of transcriptional regulation [64]. However, systematic studies of translation regulation have lagged behind despite their potential to deepen the understanding of this process.

In Weingarten-Gabbay *et al.* [33] and Chapter 5 we demonstrated that systematic measurements of cap-independent translation can be used to uncover mechanisms involved in IRES-mediated translation. This approach, relying on synthesis of short oligonucleotide sequences that are later assayed for their effect on expression, can be used to study other aspects of translation that are known to occur, but whose mechanisms are poorly understood. For example, synthetic libraries combined with appropriate reporter constructs can be used to study stop codon read-through, which was shown to be extensively used by viruses and *Drosophila*, and was identified in human transcripts [14]; or to study programmed frame-shifting, which was identified in viruses, as well as more complex organisms [16]. Insight into the regulation of these processes could lead to better understanding of diseases and ultimately to the

identification of novel therapeutic targets; as well as provide new "knobs" for adjusting gene expression for metabolic engineering and biotechnology.

In the spectrum of translation regulation mechanisms amenable to systematic measurements, translation elongation would benefit the most from such approach. Assaying expression of synonymous gene encodings, while controlling for all other variation, would allow for quantifying the effect that these substitutions have on expression. Similar measurements were previously carried out in *Escherichia coli* [65], and used to identify a strong contribution of RNA secondary structure around the start codon towards gene expression. However, these measurements assessed only reporter protein levels, and thus could not separate contributions of transcription and translation on the measured expression. The two levels of regulation could be disentangled by performing ribosome profiling measurements on the pool of cells with synonymously encoded genes (one per cell), and using bioinformatic processing to recover translation efficiency measurements from Ribo- and mRNA-seq reads that may align non-uniquely to the synonymous genes. Ultimately, such systematic measurements of synonymous codon substitutions would be particularly suited for developing predictive models for codon optimisation described in Chapter 3, and could lead to accurate data-driven codon optimisation approaches.

EPITRANSCRIPTOMICS - A NEW LEVEL OF REGULATION

Decreasing costs of second generation sequencing prompted the development of numerous high-throughput functional genomics assays, some of which allow for interrogating biochemical RNA modifications. These modifications (e.g. nucleotide methylation or pseudouridinilation, m^6A, m^5C; Sun *et al.* [66]), do not alter the RNA sequence, but modify its chemical properties. Similar to DNA methylation or histone modifications in epigenetics, RNA modifications comprising the *epitranscriptome*, can alter gene expression, thus providing an additional layer of expression regulation. Compared to the dozens of known DNA modifications, RNA modifications have a richer repertoire with more than 140 alternative forms [67], suggesting a strong potential of these modifications for transcriptional and translational control.

Owing to the recent availability of whole-genome epitranscriptomic maps, we are only now beginning to understand the effect of RNA modifications on cellular processes. However, accumulating evidence suggests that these modifications can affect miRNA targeting [68], RNA folding [69], tRNA selection in translation elongation [70], IRES-mediated translation [71], translation of methylated mRNA [72], mRNA localisation, stability [73] and splicing [74]. It may thus be necessary to re-visit previously developed models and mechanism of gene regulation and to consider them in the context of these recent findings. For example, the TASEP model in Chapter 4 could be extended to include codon- and modification- specific elongation rates, and to assess the effect of this extension on model predictions. More broadly, any approach combining sequence information and ribosome profiling data, could be complemented by epitranscriptome measurements and used to test the effect of RNA modifications on translation. Further, our IRES activity determinants study (Chapter 5), if complemented by epitranscriptome measurements, could be modified to check whether the epitranscriptome is predictive of cap-independent translation. Notably,

the first attempts at answering these questions would not require investments into new measurements, and could be undertaken today by combining publicly available datasets.

Further, epitranscriptome data could be used to determine sequence specificity of the various RNA modifications by learning sequence models that predict local modification abundance along the genome. The genome-scale sample sizes available for learning these models, and the inherent structure of linear sequences makes the problem of predicting RNA modifications from sequence particularly amenable to modelling using Deep Learning techniques [75].

DATA INTEGRATION

The explosion of high-throughput functional genomics assays in recent years allowed for interrogating previously inaccessible cellular mechanisms (see Chapter 1), however, it also made it difficult for data integration efforts to keep up with the rapid growth in available measurements. This has to do with the lag between the introduction of new measurement techniques and the development of analysis and integration methods for them.

Integration of data from existing functional genomics assays has a vast potential for improving existing analyses. For example, *in vivo* measurements of RNA structure [76] could be used to extend the IRES translation model described in Chapter 5, as it would allow for detecting the expected relationship between IRES activity and RNA structure suggested by previous studies. Initial attempts to integrate these data could use the base pairing probabilities measured by RNA structure probing methods as model features, however, more advanced approaches that reconstruct the (family) of likely *in vivo* structures [77] could better capture structural features and mechanism. Reconstruction approaches should focus on predicting pseudoknotted RNA structures that are often implicated in translation regulation, and are challenging to predict using current methods [78, 79]. Considering multiple possible structures would require new algorithms that correctly handle such information. These methods could draw from Multiple Instance Learning (MIL, Babenko [80]) approaches used in Machine Learning. Similar algorithms would also be required for extending the translation models from Chapter 4, which would have to consider the effect of dynamically changing RNA structures on model kinetics.

Data integration can also be used to test new hypothesis of gene expression regulation. For example, combining epitranscriptomic and RNA structure measurements would allow for studying the effects of RNA modifications on RNA folding. As another example, a combination of ribosome profiling, epitranscriptomic and chromatin conformation measurements can be used to test whether genes that are co-localised in the genome (i) receive similar RNA modifications; or (ii) tend to be translated with similar efficiency. These examples represent only a small fraction of the questions that could be answered by integrating different types of measurements. Due to the heterogeneity of the available measurements, it is likely that development of problem-specific methods would be required for every new question asked. Tackling this challenge would require data integration to become a priority of the broader scientific community, much like generation of new data is.

CO-TRANSLATIONAL FOLDING

Once translation modelling allows for reliable prediction of translation efficiency of RNA sequences, the next big challenge is to quantify the link between translation and protein folding (i.e. co-translational folding; Yu *et al.* [81]). Isolated examples [48, 49] suggest that synonymous codon substitutions can cause protein misfolding and reduced enzymatic activity, presumably by not allowing sufficient time for correct co-translational folding of the nascent peptide. However, this process has not been systematically studied, and how exactly translation pausing affects protein structure is not fully understood. Progress in this direction is hampered by the lack of assays that can simultaneously assess transcription, translation and folding (or enzymatic activity) of a reporter gene. Development of such assays or alternative methods for measuring protein (mis-)folding will likely remain challenging in the near future, but once solved, their data could be used to devise structure-aware models of translation. These models would require novel algorithms that are capable of calculating the probability space of possible protein folds from the simulated RNA sequence.

6.4. CONCLUDING REMARKS

SYNTHETIC DESIGNS ARE THE NEXT STEP

This thesis described algorithms and models that address several challenges in Biotechnology, ME and MB. We have developed models addressing different aspects of protein translation, including whole-cell modelling, context-dependence of elongation and cap-independent translation. The next step is to validate these models through design of synthetic sequences with sought translation activity. Additionally, validation results of model-designed sequences can be used for improving the models, as is usually done within the systems biology cycle [82]. Finally, models and approaches presented in this thesis can also be applied in other fields. For example, they could be used to predict the effects of synonymous or non-coding (e.g. IRES) mutations on expression and to investigate their links to diseases.

RECONCILING DIFFERENT REGULATORY MECHANISMS REQUIRES A "MODEL OF MODELS"

The biggest challenge for synthetic biology and metabolic engineering lays perhaps not in solving the immediate questions of codon or pathway optimisation, but rather in closing the gap between basic and applied research. These disciplines would benefit greatly from the ability to express user-designed proteins at user-desired levels when designing DNA sequences, which is hampered by the incomplete understanding of gene regulation mechanisms. It is, thus, desirable that, when writing DNA sequences, we make use of the novel regulation layers as soon as they are discovered. Due to the cross-talk between different levels of regulation, application of new mechanisms in synthetic designs would require re-visiting the previously discovered mechanisms. With the current pace of discoveries, such an approach is entirely infeasible, and would leave practical applications of new research lagging far behind. An alternative way is to create a modular framework, a "model of models", which would allow easy incorporation of new predictive models based on recent discoveries as building blocks. Black-box methods (e.g. late classifier integration) could be used to combine the individual models. However, for the sake of interpretability and reduced complexity, explicit integration

should be used whenever possible. For example, the IRES activity models (Chapter 5) could be integrated within the TASEP model (Chapter 4), in which rates of codon elongation, translation initiation and internal ribosome entry are context-dependent and provided by the corresponding "building block models". Such a modular approach is also advantageous from a cost point of view, as constructing several smaller models would require exponentially less data than constructing the overall model directly.

LARGER DATASETS SHOULD LEAD TO MORE ACCURATE MODELS

Construction of reliable models requires large amounts of high-quality data. However, most basic research experiments are designed to test *a priori* hypotheses and do not generate data of quality or coverage suitable for model construction. For example, in Chapter 5 we described analysis of a library of 55,000 synthetic sequences of 172nt in length. This is the largest library with measured IRES activity available to date, however, it represents less than $\frac{1}{10^{95}}$% of all possible sequences of that length. While datasets with low coverage of the problem space can also be used to derive predictive models, larger datasets are required for achieving prediction confidence high enough for model integration, and for model-driven ME and SB that would not require multiple design-measurement iterations. Ideally, future modelling efforts should be complemented by (iterations of) data generation to facilitate model construction.

Overcoming current bioinformatic challenges will rely on algorithmic and methodological improvements of *in silico* analyses. It will likely also prompt further technological innovation, which, in turn, is bound to lead to new bioinformatic challenges and opportunities, thus creating a "flywheel" spinning ever-faster to accelerate the pace of biological discovery. This positive feedback loop will guarantee a bright future for bioinformatics and computation biology for the years to come.

REFERENCES

[1] J. F. Nijkamp, M. van den Broek, E. Datema, S. de Kok, L. Bosman, M. A. Luttik, P. Daran-Lapujade, W. Vongsangnak, J. Nielsen, W. H. Heijne, *et al., De novo sequencing, assembly and analysis of the genome of the laboratory strain Saccharomyces cerevisiae CEN.PK113-7D, a model for modern industrial biotechnology,* Microbial Cell Factories **11**, 1 (2012).

[2] M. Pop, A. Phillippy, A. L. Delcher, and S. L. Salzberg, *Comparative genome assembly,* Briefings in Bioinformatics **5**, 237 (2004).

[3] E. Nürenberg and R. Tampé, *Tying up loose ends: ribosome recycling in eukaryotes and archaea,* Trends in Biochemical Sciences **38**, 64 (2013).

[4] D. J. Young, N. R. Guydosh, F. Zhang, A. G. Hinnebusch, and R. Green, *Rli1/ABCE1 Recycles Terminating Ribosomes and Controls Translation Reinitiation in 3′ UTRs In Vivo*, Cell **162**, 872 (2015).

[5] G. Cannarozzi, N. N. Schraudolph, M. Faty, P. von Rohr, M. T. Friberg, A. C. Roth, P. Gonnet, G. Gonnet, and Y. Barral, *A role for codon order in translation dynamics,* Cell **141**, 355 (2010).

[6] S. J. Andrews and J. A. Rothnagel, *Emerging evidence for functional peptides encoded by short open reading frames,* Nature Reviews Genetics **15**, 193 (2014).

[7] S. K. Young, J. A. Willy, C. Wu, M. S. Sachs, and R. C. Wek, *Ribosome Reinitiation Directs Gene-specific Translation and Regulates the Integrated Stress Response,* Journal of Biological Chemistry **290**, 28257 (2015).

[8] R. J. Jackson, C. U. Hellen, and T. V. Pestova, *The mechanism of eukaryotic translation initiation and principles of its regulation,* Nature Reviews Molecular Cell Biology **11**, 113 (2010).

[9] A. A. Komar and M. Hatzoglou, *Cellular IRES-mediated translation: the war of ITAFs in pathophysiological states,* Cell Cycle **10**, 229 (2011).

[10] W. Miller, Z. Wang, and K. Treder, *The amazing diversity of cap-independent translation elements in the 3′-untranslated regions of plant viral RNAs,* Biochemical Society Transactions **35**, 1629 (2007).

[11] I. N. Shatsky, S. E. Dmitriev, I. M. Terenin, and D. Andreev, *Cap-and IRES-independent scanning mechanism of translation initiation as an alternative to the concept of cellular IRESs,* Molecules and Cells **30**, 285 (2010).

[12] S. Djuranovic, A. Nahvi, and R. Green, *miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay,* Science **336**, 237 (2012).

[13] A. Wilczynska and M. Bushell, *The complexity of miRNA-mediated repression,* Cell Death & Differentiation **22**, 22 (2015).

[14] G. Loughran, M.-Y. Chou, I. P. Ivanov, I. Jungreis,

M. Kellis, A. M. Kiran, P. V. Baranov, and J. F. Atkins, *Evidence of efficient stop codon readthrough in four mammalian genes,* Nucleic Acids Research **42**, 8928 (2014).

[15] I. Jungreis, M. F. Lin, R. Spokony, C. S. Chan, N. Negre, A. Victorsen, K. P. White, and M. Kellis, *Evidence of abundant stop codon readthrough in Drosophila and other metazoa,* Genome Research **21**, 2096 (2011).

[16] J. D. Dinman, *Programmed ribosomal frameshifting goes beyond viruses: organisms from all three kingdoms use frameshifting to regulate gene expression, perhaps signaling a paradigm shift,* Microbe (Washington, DC) **1**, 521 (2006).

[17] J. Kong and P. Lasko, *Translational control in cellular and developmental processes,* Nature Reviews Genetics **13**, 383 (2012).

[18] P. Reynolds, *In sickness and in health: the importance of translational regulation,* Archives of Disease in Childhood **86**, 322 (2002).

[19] J. Darnell, *Defects in translational regulation contributing to human cognitive and behavioral disease,* Current Opinion in Genetics & Development **21**, 465 (2011).

[20] N. T. Ingolia, *Ribosome profiling: new views of translation, from single codons to genome scale,* Nature Reviews Genetics **15**, 205 (2014).

[21] C. G. Artieri and H. B. Fraser, *Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation,* Genome Research **24**, 2011 (2014).

[22] J. A. Hussmann, S. Patchett, A. Johnson, S. Sawyer, and W. H. Press, *Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast,* PLoS Genetics **11**, e1005732 (2015).

[23] D. E. Weinberg, P. Shah, S. W. Eichhorn, J. A. Hussmann, J. B. Plotkin, and D. P. Bartel, *Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation,* bioRxiv , 021501 (2015).

[24] S. Uemura, C. E. Aitken, J. Korlach, B. A. Flusberg, S. W. Turner, and J. D. Puglisi, *Real-time tRNA transit on single translating ribosomes at codon resolution,* Nature **464**, 1012 (2010).

[25] S. C. Baker, S. R. Bauer, R. P. Beyer, J. D. Brenton, B. Bromley, J. Burrill, H. Causton, M. P. Conley, R. Elespuru, M. Fero, *et al., The external RNA controls consortium: a progress report,* Nature Methods **2**, 731 (2005).

[26] J. Lovén, D. A. Orlando, A. A. Sigova, C. Y. Lin, P. B. Rahl, C. B. Burge, D. L. Levens, T. I. Lee, and R. A. Young, *Revisiting global gene expression analysis,* Cell **151**, 476 (2012).

[27] D. Risso, J. Ngai, T. P. Speed, and S. Dudoit, *Normalization of RNA-seq data using factor analysis of control genes or samples,* Nature Biotechnology **32**, 896 (2014).

[28] L. F. Lareau, D. H. Hite, G. J. Hogan, and P. O. Brown, *Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments,* eLife **3**, e01257 (2014).

[29] G. Ast, *How did alternative splicing evolve?* Nature Reviews Genetics **5**, 773 (2004).

[30] P. Shah, Y. Ding, M. Niemczyk, G. Kudla, and J. B. Plotkin, *Rate-limiting steps in yeast protein translation,* Cell **153**, 1589 (2013).

[31] R. Toribio and I. Ventoso, *Inhibition of host translation by virus infection in vivo,* Proceedings of the National Academy of Sciences **107**, 9837 (2010).

[32] S. E. Wohlgemuth, T. E. Gorochowski, and J. A. Roubos, *Translational sensitivity of the* Escherichia coli *genome to fluctuating tRNA availability,* Nucleic Acids Research **41**, 8021 (2013).

[33] S. Weingarten-Gabbay, S. Elias-Kirma, R. Nir, A. A. Gritsenko, N. Stern-Ginossar, Z. Yakhini, A. Weinberger, and E. Segal, *Systematic discovery of cap-independent translation sequences in human and viral genomes,* Science **351**, aad4939 (2016).

[34] K. B. Cook, H. Kazan, K. Zuberi, Q. Morris, and T. R. Hughes, *RBPDB: a database of RNA-binding specificities,* Nucleic Acids Research **39**, D301 (2011).

[35] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, *Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning,* Nature Biotechnology (2015).

[36] R. B. Darnell, *HITS-CLIP: panoramic views of protein-RNA regulation in living cells,* Wiley Interdisciplinary Reviews: RNA **1**, 266 (2010).

[37] T.-Y. Wu, C.-C. Hsieh, J.-J. Hong, C.-Y. Chen, and Y.-S. Tsai, *IRSS: a web-based tool for automatic layout and analysis of IRES secondary structure prediction and searching system in silico,* BMC Bioinformatics **10**, 1 (2009).

[38] J.-J. Hong, T.-Y. Wu, T.-Y. Chang, and C.-Y. Chen, *Viral IRES prediction system-a web server for prediction of the IRES secondary structure in silico,* PloS One **8**, e79288 (2013).

[39] P. P. Gardner, J. Daub, J. G. Tate, E. P. Nawrocki, D. L. Kolbe, S. Lindgreen, A. C. Wilkinson, R. D. Finn, S. Griffiths-Jones, S. R. Eddy, *et al., Rfam: updates to the RNA families database,* Nucleic Acids Research **37**, D136 (2009).

[40] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, *The protein data bank,* Nucleic Acids Research **28**, 235 (2000).

[41] A. Saxena, R. S. Sangwan, and S. Mishra, *Fundamentals of homology modeling steps and comparison among important bioinformatics tools: An overview,* Science International **1**, 237 (2013).

[42] A. I. Petrov, C. L. Zirbel, and N. B. Leontis, *Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas,* RNA **19**, 1327 (2013).

[43] C. L. Zirbel, J. Roll, B. A. Sweeney, A. I. Petrov, M. Pirrung, and N. B. Leontis, *Identifying novel sequence variants of RNA 3D motifs,* Nucleic Acids Research , gkv651 (2015).

[44] S. Zhang, J. Zhou, H. Hu, H. Gong, L. Chen, C. Cheng, and J. Zeng, *A deep learning framework for modeling structural features of RNA-binding protein targets,* Nucleic Acids Research , gkv1025 (2015).

[45] W. L. Noderer, R. J. Flockhart, A. Bhaduri, A. J. D. de Arce, J. Zhang, P. A. Khavari, and C. L. Wang, *Quantitative analysis of mammalian translation initiation sites by FACS-seq,* Molecular Systems Biology **10**, 748 (2014).

[46] S. Lee, B. Liu, S. Lee, S.-X. Huang, B. Shen, and S.-B. Qian, *Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution,* Proceedings of the National Academy of Sciences **109**, E2424 (2012).

[47] X. Gao, J. Wan, B. Liu, M. Ma, B. Shen, and S.-B. Qian, *Quantitative profiling of initiating ribosomes in vivo,* Nature Methods **12**, 147 (2015).

[48] C. Gustafsson, S. Govindarajan, and J. Minshull, *Codon bias and heterologous protein expression,* Trends in Biotechnology **22**, 346 (2004).

[49] E. Angov, *Codon usage: nature's roadmap to expression and folding of proteins,* Biotechnology Journal **6**, 650 (2011).

[50] M. Welch, S. Govindarajan, J. E. Ness, A. Villalobos, A. Gurney, J. Minshull, and C. Gustafsson, *Design parameters to control synthetic gene expression in Escherichia coli,* PloS One **4**, e7002 (2009).

[51] T. Gaj, C. A. Gersbach, and C. F. Barbas, *ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering,* Trends in Biotechnology **31**, 397 (2013).

[52] J. D. Sander and J. K. Joung, *CRISPR-Cas systems for editing, regulating and targeting genomes,* Nature Biotechnology **32**, 347 (2014).

[53] M. G. Ross, C. Russ, M. Costello, A. Hollinger, N. J. Lennon, R. Hegarty, C. Nusbaum, and D. B. Jaffe, *Characterizing and measuring bias in sequence data,* Genome Biology **14**, R51 (2013).

[54] T. Laver, J. Harrison, P. O'Neill, K. Moore, A. Farbos, K. Paszkiewicz, and D. Studholme, *Assessing the performance of the Oxford Nanopore Technologies MinION,* Biomolecular Detection and Quantification **3**, 1 (2015).

[55] A. E. Minoche, J. C. Dohm, J. Schneider, D. Holtgräwe, P. Viehöver, M. Montfort, T. R. Sörensen, B. Weisshaar, and H. Himmelbauer, *Exploiting single-molecule transcript sequencing for eukaryotic gene prediction,* Genome Biology **16**, 1 (2015).

[56] C. M. O'Neill, D. Baker, G. Bennett, J. Clarke, and I. Bancroft, *Two high linolenic mutants of Arabidopsis thaliana contain megabase-scale genome duplications encompassing the fad3 locus,* The Plant Journal **68**, 912 (2011).

[57] L. Faino, M. F. Seidl, E. Datema, G. C. van den Berg, A. Janssen, A. H. Wittenberg, and B. P. Thomma, *Single-Molecule Real-Time Sequencing Combined with Optical Mapping Yields Completely Finished Fungal Genome,* mBio **6**, e00936 (2015).

[58] S. N. Floor and J. A. Doudna, *Tunable protein synthesis by transcript isoforms in human cells,* eLife , e10921 (2016).

[59] L. B. Carey, D. Van Dijk, P. M. Sloot, J. A. Kaandorp, and E. Segal, *Promoter sequence determines the relationship between expression level and noise,* PLoS Biology **11**, e1001528 (2013).

[60] E. Sharon, D. van Dijk, Y. Kalma, L. Keren, O. Manor, Z. Yakhini, and E. Segal, *Probing the effect of promoters on noise in gene expression using thousands of designed sequences,* Genome Research **24**, 1698 (2014).

[61] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano, A.-C. Jungkamp, M. Munschauer, *et al., PAR-CliP - a method to identify transcriptome-wide the binding sites of RNA binding proteins,* Journal of Visualized Experiments **41**, 2034 (2010).

[62] J. D. Buenrostro, B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, and W. J. Greenleaf, *Single-cell chromatin accessibility reveals principles of regulatory variation,* Nature **523**, 486 (2015).

[63] S. Dvir, L. Velten, E. Sharon, D. Zeevi, L. B. Carey, A. Weinberger, and E. Segal, *Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast,* Proceedings of the National Academy of Sciences **110**, E2792 (2013).

[64] S. Weingarten-Gabbay and E. Segal, *The grammar of transcriptional regulation,* Human Genetics **133**, 701 (2014).

[65] G. Kudla, A. Murray, D. Tollervey, and J. Plotkin, *Coding-sequence determinants of gene expression in Escherichia coli,* Science **324**, 255 (2009).

[66] W.-J. Sun, J.-H. Li, S. Liu, J. Wu, H. Zhou, L.-H. Qu, and J.-H. Yang, *RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data,* Nucleic Acids Research , gkv1036 (2015).

[67] H. Grosjean, *RNA modification: the Golden Period 1995–2015,* RNA **21**, 625 (2015).

[68] A. M. Burroughs, Y. Ando, M. J. de Hoon, Y. Tomaru, T. Nishibu, R. Ukekawa, T. Funakoshi, T. Kurokawa, H. Suzuki, Y. Hayashizaki, *et al., A comprehensive survey of 3′ animal miRNA modification events and a possible role for 3′ adenylation in modulating miRNA targeting effectiveness,* Genome Research **20**, 1398 (2010).

[69] M. Helm, *Post-transcriptional nucleotide modification and alternative folding of RNA,* Nucleic Acids Research **34**, 721 (2006).

[70] J. Choi, K.-W. Ieong, H. Demirci, J. Chen, A. Petrov, A. Prabhakar, S. E. O'Leary, D. Dominissini, G. Rechavi, S. M. Soltis, *et al., N^6-methyladenosine in mRNA disrupts tRNA selection and translation-elongation dynamics,* Nature Structural & Molecular Biology (2016).

[71] S. R. Thompson, *So you want to know if your message has an IRES?* Wiley Interdisciplinary Reviews: RNA **3**, 697 (2012).

[72] D. Dominissini, S. Nachtergaele, S. Moshitch-Moshkovitz, E. Peer, N. Kol, M. S. Ben-Haim, Q. Dai, A. Di Segni, M. Salmon-Divon, W. C. Clark, *et al., The dynamic N^1-methyladenosine methylome in eukaryotic messenger RNA,* Nature (2016).

[73] Y. Fu, D. Dominissini, G. Rechavi, and C. He, *Gene expression regulation mediated through reversible m^6a RNA methylation,* Nature Reviews Genetics **15**, 293 (2014).

[74] J. Karijolich, A. Kantartzis, and Y.-T. Yu, *RNA modifications: a mechanism that modulates gene expression,* in *RNA Therapeutics* (Springer, 2010) pp. 1–19.

[75] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning,* Nature **521**, 436 (2015).

[76] R. A. Flynn, Q. C. Zhang, R. C. Spitale, B. Lee, M. R. Mumbach, and H. Y. Chang, *Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE,* Nature Protocols **11**, 273 (2016).

[77] M. Kertesz, Y. Wan, E. Mazor, J. L. Rinn, R. C. Nutter, H. Y. Chang, and E. Segal, *Genome-wide measurement of RNA secondary structure in yeast,* Nature **467**, 103 (2010).

[78] M. Hajiaghayi, A. Condon, and H. H. Hoos, *Analysis of energy-based algorithms for RNA secondary structure prediction,* BMC Bioinformatics **13**, 1 (2012).

[79] A. Krokhotin and N. V. Dokholyan, *Chapter Three - Computational Methods Toward Accurate RNA Structure Prediction Using Coarse-Grained and All-Atom Models,* Methods in Enzymology **553**, 65 (2015).

[80] B. Babenko, *Multiple instance learning: algorithms and applications*, Tech. Rep. (Department of Computer Science and Engineering, University of California, San Diego, USA, 2004).

[81] C.-H. Yu, Y. Dang, Z. Zhou, C. Wu, F. Zhao, M. S. Sachs, and Y. Liu, *Codon usage influences the local rate of translation elongation to regulate co-translational protein folding,* Molecular Cell **59**, 744 (2015).

[82] H. Kitano, *Systems biology: a brief overview,* Science **295**, 1662 (2002).

A WORD OF THANKS

I remember one of the first meetings I had with my advisors Dick and Marcel about a month into my PhD. They were wondering whether I was already making friends in Delft and within the research group. I responded that I prefer to keep work and friendships separate. I am happy to admit that I was wrong.

Throughout my PhD I have met countless amazing people, many of which I proudly consider friends. Without them, neither my work, nor my journey through the PhD would have been the same. And I'd like to use this opportunity to extend my gratitude to the people that made the past 6 years so special. It's impossible to list everyone, but I will do my best!

Dick, **Marcel**, first of all, thank you for taking a chance on me and giving me an opportunity to pursue a PhD in Delft. It's been a rewarding experience. The lessons I've learned during the PhD go far beyond academia, and are something that I will carry through life. I consider myself lucky to have had the two of you as supervisors: your guidance, coaching, vast scientific knowledge, ability to switch from one complicated topic to another in minutes, endless drive and passion for scientific discovery, the ability to challenge, criticise and motivate – all at the same time – inspired me to also try to go the extra mile and push further, like you do.

I'm also very grateful for the freedom and flexibility that you granted me in my PhD. Whether it was the choice of research topics, learning to do wet lab work, going abroad for 5 months, or waiting for me to finish the last bits of the thesis next to a full-time job - you always found a way to make it work. I feel that this really made the PhD my own, and I would not want to have it any other way.

Dick, remember the "Mastering Your PhD" book you gave to me before I even started? I think I am one of the few people who actually read it. It talks a lot about the relationship between the PhD student and the PhD supervisor, and about how special and unique it is. While it is difficult to put the specifics into words, I believe that our relationship fits this description quite accurately. And I hold this very dear.

I had the privilege of carrying out my PhD within the Pattern Recognition and Bioinformatics (PRB) group, alongside some of the most intelligent people I know. Many contributed to the research presented in this thesis, but I want to highlight just a few. **Marc**, **Jurgen**, **Wynand**, **Bastiaan**, **Sjoerd**, **Erik**, **Jasper**, your insights in (bio)informatics, mathematics, biology and statistics can be found back in different chapters, and shaped some of them significantly. **Marco**, **Lodewyk**, **Jeroen**, **Laurens**, **Thomas**, **Roeland**, **Emile**, **David**, while we did not have a chance to work as closely, I still learned a great deal from you!

But intelligence is perhaps something that PhD research selects for. That's why I'm happy to have finished the PhD with a lot of good stories to tell about many of colleagues, including some that are better left out of the acknowledgements ;-). **Marc**, our road trip after ISMB left some of the brightest memories, and allowed me to realise how much I actually love travelling. **Amin**, your work attitude and ethics served as an inspiration, and your sense of humour and openness allowed for filling the times when I didn't feel like working with the most interesting conversations. **Thies**, **Wouter**, I hope to some day be able to tell your jokes! **Erik**, **Erdogan**, **Ahmed**, I don't know how is it that our coffee machine chats drifted into the most bizarre topics so often, but I blame you. **Ekin**, **Veronika**, **Alex**, **Laura**, Thursday borrels with you and their city center extensions were immensely fun and provided the necessary distraction from PhD student troubles. **Christian**, **Joana**, **Marcel v.d. Broek**, **Joske**, I always fondly enjoyed the random chats that would spark in our room. **Sepideh**, **Gorkem**, **Roy**, I recall many pleasant conversations with you that left me feeling that I can get through it, no matter what "it" meant at the time.

I also had a unique opportunity to work together with experimental biologists from Industrial Microbiology (IMB) group in Delft. And to do some of the experimental work myself too! **Jean-Marc**, **Frank**, **Jules**, I discovered an entirely new world of scientific research through our collaboration that turned out to be invaluable at later stages of my PhD. **Barbara**, **Irina**, **Beth**, you helped me get from a point where I could not hold a pipette, to a point where I could independently carry out simple experiments - it is a testament to your patience! **Tim**, **Robbert**, **Harmen**, **Niels**, thank you for answering my endless (stupid) questions, and for doing that in a way that I could actually make sense of.

Spending 5 months of my PhD in the Segal Lab at the Weizmann Institute of Science in Israel was an amazing experience. Thank you for making me feel so much at home there, that I had to seriously consider the possibility of staying longer. **Eran**, many thanks for making the research visit to your lab possible; the insights and learnings I took away from this visit are invaluable. **Shira**, the breadth of your scientific knowledge, and your enthusiasm for research never ceased to amaze and inspire me. **Martin**, it was a great pleasure working with, but I enjoyed exploring Israel together even more. **Ilya**, it was great having you as a roommate and to be able to discuss fitness alongside science. **Dudi**, **Tal**, thanks for introducing me to Tubi 60, too bad I can't find it anywhere. And many thanks to all other friends I made at the Weizmann Institute - my time there would not have been the same without you.

In my experience, PhD is an endeavour that also takes a toll on the people close to you. And I'm grateful that my family stood by me even during the most difficult periods. Мама, папа, ничего этого не было бы возможно без вашего примера и поддержки. Именно благодаря вам я считаю, что девиз нашей семьи "никогда не сдавайся". **Alina**, I really appreciate the small things things you do - be it a cup of coffee or drawings that you made, they all mean a lot. **Katia**, meeting you was one of the best things that happened to me. I don't think I would be able to get to the end of the

PhD without your unconditional love, support, limitless patience and understanding. I also owe a debt of gratitude to your family, who has accepted me as their own. **Laurinda**, **Agosthino**, obrigado pela sua hospitalidade, e por toda a comida deliciosa.

Alexey A. Gritsenko
Delft, December 2016

CURRICULUM VITÆ

Alexey Alexeevich Gritsenko (Russian: Алексей Алексеевич Гриценко) was born on the 6th of September, 1989 in Barnaul, Altai krai, USSR. Before the age of 4, he had already lived in Siberia, Kazakhstan, and Kaliningrad, where his family eventually settled.

At the age of 5 he started pre-university education at the Kaliningrad Lyceum № 18, where he later chose to focus on mathematics and physics. After graduating in 2005, he gained admission to the Applied Mathematics and Computer Science programme of the Immanuel Kant Baltic Federal University. During his studies he became passionate about Computer Science and algorithm development. This passion found its expression in the participation in the ACM ICPC programming contests, which culminated with him competing in the World Finals of 2011 in Orlando, FL, USA. However, despite his affinity to Computer Science, in his yearly thesis projects Alexey concentrated on the more theoretical problem of breaking elliptic curve cyphers.

In 2009 Alexey was admitted to the joint Leiden-Delft Bioinformatics master programme, arranged to continue his studies in Russia remotely, and moved to Oegstgeest, the Netherlands to commence his MSc in Leiden University. He carried out his final thesis project under the supervision of dr.ir. Dick de Ridder and Jurgen Nijkamp at TU Delft. After graduating with distinction from both Leiden University and the Immanuel Kant Baltic Federal University, Alexey continued working with dr.ir. de Ridder and prof.dr.ir. Marcel J.T. Reinders as a PhD student at the Pattern Recognition and Bioinformatics group of the TU Delft.

Alexey carried out his PhD within the Platform Green Synthetic Biology consortium aimed at developing biotechnological methods for the production of plant-derived chemicals in microbes. His research evolved to be centred around the more fundamental question of understanding and modelling protein synthesis in living cells. In 2015 he went on a 5-month research visit to the Segal lab of Computational Biology at the Weizmann Institute of Science in Rehovot, Israel. Since May 2016, Alexey works as a Data Scientist at Booking.com in Amsterdam, the Netherlands.

LIST OF PUBLICATIONS

5. **A.A. Gritsenko**[*], S. Weingarten-Gabbay[*], S. Elias-Kirma, R. Nir, D. de Ridder, and E. Segal, *Sequence features of viral and human Internal Ribosome Entry Sites predictive of their activity*, submitted for publication.

4. S. Weingarten-Gabbay, S. Elias-Kirma, R. Nir, **A.A. Gritsenko**, N. Stern-Ginossar, Z. Yakhini, A. Weinberger, and E. Segal, *Systematic discovery of cap-independent translation sequences in human and viral genomes*, Science, **351**, aad4939 (2016).

3. **A.A. Gritsenko**, M. Hulsman, M.J.T. Reinders, and D. de Ridder, *Unbiased Quantitative Models of Protein Translation Derived from Ribosome Profiling Data*, PLoS Computational Biology, **11**, e1004336 (2015).

2. **A.A. Gritsenko**, M.J.T. Reinders, and D. de Ridder, *Using predictive models to engineer biology: a case study in codon optimization*, in *Pattern Recognition in Bioinformatics*, pp. 159–171 (Springer, Nice, France, 2013).

1. **A.A. Gritsenko**, J.F. Nijkamp, M.J.T. Reinders, and D. de Ridder, *GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies*, Bioinformatics, **28**, 1429 (2012).

[*] These authors contributed equally.