Jørgen Skancke

The beginning and the end of gene expression

Doctoral thesis
for the degree of philosophiae doctor

Trondheim, May 2016

Norwegian University of Science and Technology
The Faculty of Natural Sciences and Technology
Department of Chemical Engineering

NTNU
Innovation and Creativity

NTNU

Norwegian University of Science and Technology

Doctoral thesis
for the degree of philosophiae doctor

The Faculty of Natural Sciences and Technology
Department of Chemical Engineering

ISBN 978-82-326-0390-9 (printed version)
ISBN 978-82-326-0391-6 (electronic version)
ISSN 1503-8181

Doctoral theses at NTNU, 2016:8

Printed by Skipnes kommunikasjon

Summary

Gene expression – the synthesis of RNA and protein – requires most of the cell's energy and is a highly regulated process at all its levels. In this thesis, four studies are presented which focus on the regulation of three different levels of gene expression. Two are centered on the regulation of transcription initiation; the other on translation initiation; while the third looks at RNA post-transcriptional processing. The studies are integrated in the way that they illustrate how DNA sequences directly affect regulation at these three different levels. The studies rely on different experimental data, however, and require different theoretical methods of analysis.

The studies of transcription initiation focus on how the RNA polymerase (RNAP) takes the step from promoter binding to promoter escape. Before reaching promoter escape, RNAP may undergo abortive initiation, in which the nascent RNA transcript is released from RNAP. At many promoters, RNAP undergoes repeated abortive initiation, known as abortive cycling, before promoter escape is reached. The DNA sequence composition of the 20 first transcribed basepairs has been known to affect the efficiency of promoter escape, presumably by affecting the probabilities of abortive initiation, but the mechanism for this effect was not known. Using a thermodynamic model of transcription initiation, we have shown that the manner in which RNAP translocates during initial transcription explains the observed variation in promoter escape efficiency on the N25 promoter. The key variable for linking translocation of RNAP during initiation and promoter escape efficiency was the sequence of the RNA $3'$ dinucleotide, and not the free energy of the DNA bubble which had been postulated by others. We proceed to verify our findings with follow-up experiments, in which we used our thermodynamic model to construct N25 promoter variants with predicted promoter escape efficiencies. The experiments agreed well with the predictions, making a strong argument for translocation during initial RNA synthesis as the major determinant of promoter escape efficiency. While this study sheds light on the sequence specificity of initial transcription, it

does not reveal the dynamics of the process. This therefore is the focus on the second piece of work on initial transcription in this thesis. In the work on kinetics, we combine two lines of separate experimental evidence (single-molecule and bulk transcription studies) to identify the rate constants for the key steps in initial transcription: the nucleotide addition cycle, backtracking, and unscrunching and abortive RNA release. The most important finding in this work was that the speed of pause-free transcription during initiation is the same for initial transcription as reported for transcription elongation. This tells us that scrunching and the expanded DNA bubble do not seem to affect the kinetics of the combined steps of nucleotide addition, translocation, and pyrophosphorolysis.

The study on translation initiation focuses on the problem of heterologous expression of the human *inf-α2b* gene in *E. coli.* A known bottleneck for heterologous gene expression is the binding of the ribosome to a folded translation initiation region on a messenger RNA. Therefore, we introduced silent mutations in the first 9 codons of *inf-α2b* that were predicted to reduce the free energy of RNA secondary structures at the ribosome binding site to different degrees. This approach produced some *inf-α2b* variants which showed an increased amount of *inf-α2b* transcript, indicating that translation initiation is likely a strong barrier for the expression of this gene in an *E. coli* host.

In the final study, we focused on the process of $3'$ cleavage and polyadenylation of eukaryote RNAs. Cleavage and polyadenylation make up a part of pre-RNA processing of many eukaryote RNAs, and is required for mRNA stability and transport from the nucleus to the cytoplasm. To study $3'$ cleavage and polyadenylation, we analyzed RNA sequencing data to identify polyadenylation sites in transcripts from different cell compartments across 12 human cell lines. We found over 160.000 polyadenylation sites, 80% of which were previously not annotated. In addition we found an unexpected enrichment of polyadenylation sites in intronic regions of nuclear RNA. We offer a discussion on how these sites may be signals of polyadenylation-related nuclear degradation.

We have employed two different approaches in our investigation of the three different aspects of gene regulation: a gene-centric approach for transcription and translation initiation, and a genome-centric approach for $3'$ cleavage and polyadenylation. This has resulted in different types of research questions being asked, different computational challenges, and different sorts of results, with the gene-centric results being more mechanism oriented, and the genome-centric being more general.

The main conclusion of the thesis as a whole is that kinetic modelling

and free energy calculations of RNA and DNA nucleotide chains have been successfully applied in combination with traditional molecular biology techniques; we base this on the results from our investigations of the relationship between gene sequence and gene expression for RNAP-dependent transcription initiation and ribosome-dependent translation initiation.

Acknowledgements

I have many people to thank for helping me get through the past years in good shape. First I owe a great thanks to my main advisor, Nadav Bar, for taking me on as a student and giving me a chance to see what the world of research is all about.

Second of all I would like to thank my fellow PhD students at NTNU for their help and friendship. From the Department of Chemical Engineering I thank Deeptanshu for all the meaningful discussions and culinary experiences, Johannes for talking about window managers and climbing, Bjørn-Tore for many, many climbing sessions, Ivan for teaching me *rude*mentary Italian while writing up his thesis, Olaf for insisting that I start using Linux from day one, and Magnus, Maryam and Ismael for good company and discussions. From Biotechnology I thank Rahmi for good advice and suggesting I should check out Python; I thank Veronika for friendship and good collaboration, and Simone and Friederike for many interesting discussions. From the Department of Biology I thank Aravind and Konika for many conversations and for sharing very good Indian food during lunch. An extra thanks to Konika for making it so that I could stay with her family in Delhi when I was there – it was a blast! Finally, I thank Deeptanshu, Aravind, and Shalini for the many shared dinners, cabin trips and good conversations. You made this journey easier!

Thirdly I owe a great *moltes gràcies* to Roderic Guigó and his group at the CRG in Barcelona. I came to you during a challenging time in my PhD work and you took me along from the very beginning. My stay in Barcelona was an eye-opener, as I observed and partook in how a high-level, productive and professional research environment like the CRG operates. Thanks to Roderic, Pedro, Rory, Sarah, Angelika, David, Maik, and all the rest for teaching me about bioinformatics and genomics.

Fourthly I thank Martin Kuiper for becoming my co-supervisor during the last years of my PhD and for helping me getting through to the end. Your help was especially appreciated when I moved from the CRG back to

NTNU and also in the final stages of writing the thesis.

Fifthly, I want to give a very large *thank you* to Lilian M. Hsu. Through so many emails Lilian has constructively aided my education of the intricacies of transcription initiation. I thank you for your help and your collaboration. I especially want to thank you for trusting my modeling results enough to perform experiments using the ITS-sequences I had suggested. Waiting for the outcome those experiments was definitely the most exciting part of my PhD.

I am also happy to thank my mother and my father in Tromsø and my sister in Oslo. You have always supported me and made my life easy so that I could focus on the things I wanted to do. Thank you so much for that.

Finally and most importantly I thank my wonderful Itziar for her positive spirits, happy smiles, and for reminding me time and time again about what is important in life. Thank you for listening to my ramblings about my latest fancy in molecular or evolutionary biology and for patiently explaining all the things I never learned. Thank you for making life easy and happy and for inspiring me with your imagination and creativity!

List of papers

The work on this thesis has resulted in the main authorship by the thesis author of two papers and the co-authorship of two papers. Below are listed the papers (page number where they are found in this thesis in parenthesis) and the authors and their contributions. For the papers where the thesis author has a co-authorship, a paragraph describing his direct contribution is also given.

Design and Optimization of Short DNA Sequences That Can Be Used as 5′ Fusion Partners for High-Level Expression of Heterologous Genes in *Escherichia coli* (page 111)

Veronika Kucharova planned the work and experiments, performed experiments, analysed data, and wrote the paper
Jørgen Skancke performed bioinformatic analysis and gave input to experiments; gave input to the writing of the paper (20 % of work)
Trygve Brautaset planned the work and wrote the paper
Svein Valla planned the work and wrote the paper

The thesis authors' main contribution to this paper is the directed computational method for screening for highly expressing 5′ variants. This approach yielded consistently elevated transcript levels compared to the original gene, showing that directed computational methods on the gene sequence level can be successfully used to enhance transcript levels, which in this study sidestepped the time-consuming step of generating and screening a physical sequence library by random mutagenesis. The thesis author contributed directly Figure 1 and the sequences tested for transcript and protein levels as shown in Table 2.

Thermodynamic modeling of initial transcription elucidates the sequence dependence of promoter escape efficiency (page 55)

Jørgen Skancke developed the idea and theory behind the paper; performed data analysis; developed, implemented, and simulated the model; applied the model to generate nucleotide sequences for experimental model validation; wrote the paper and created the figures (80 % of work)
Nadav Bar participated in discussions and gave input to writing
Martin Kuiper particiapted in discussions and gave input to writing
Lilian M. Hsu performed all wet lab experiments; participated in discussions; wrote the paper

Kinetic Model of Initial Transcription Suggests the Same Average Speed for Promoter Bound and Elongating RNA Polymerase (page 57)

Jørgen Skancke developed the idea; performed data analysis; developed, implemented, and simulated the model; created figures, and wrote the paper (90 % of work)
Nadav Bar participated in discussions, created figures, and wrote the paper

Landscape of transcription in human cells (page 101)

The list of authors can be read on page 101. Due to the large number of authors, only the contribution by the thesis author is listed below.

Jørgen Skancke analysed data; provided results on polyadenylation sites; gave input to the writing of the paper (5 % of work)

The thesis author's contribution to this paper is the outcome of a one year stay at the biomedical research institution CRG (Center for Genomic Regulation), where he contributed to a collaborative effort that led up to this publication. The thesis author's direct contributions are the results on polyadenylation sites given in the section "Alternative transcription initiation and termination". His contribution is important as it complements the work by other authors on transcription start sites. The methods and background study that led up to the thesis author's published results are presented in chapter 6 on page 59.

Contents

List of Figures

List of Tables

Chapter 2

Introduction

2.1 Overview

The aim for this doctoral work was to use computational tools in combination with empirical experiments to study regulation of gene expression due to variation in deoxynucleic acid (DNA) sequences. This aim has been met by using calculations of free energy changes for base pairing of ribonucleic acid (RNA) and DNA molecules to study the mechanisms of transcription initiation and translation initiation, as well as using DNA-pattern recognition to identify polyadenylation sites at the $3'$ end of human messenger RNA (mRNA). The work on transcription initiation has further been complemented by a kinetic study. The unifying theme in this thesis work is the effect of sequence variation in DNA and RNA leads on different aspects of gene expression.

RNA and DNA molecules are polymers of nucleotides denoted by G, A, C, and T (U instead of T for RNA). Due to their polymer nature, RNA and DNA are also called nucleotide chains. It is today introductory text book material that these nucleotide chains can anneal together by base pairing to form complementary double stranded DNA and RNA, as well as RNA-DNA hybrids. Now taken for granted, the self-complementarity of RNA and DNA was once a tremendous discovery and remains the very basis for life itself: the sole reason why the genetic material can be replicated and expressed is because the self-complementarity of double stranded nucleotides. This feature enables the replication of DNA, the synthesis of RNA (transcription), and the synthesis of protein (translation).

The role of double stranded DNA in DNA replication is well expressed in the famous quote from the study by Watson and Crick [1] on the discovery of the structure of double stranded DNA: "(...) the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material". In other words, the DNA helix consists of two complementary molecules, and this complementarity enables the accurate base-to-base replication of the genetic material.

In transcription, base pairing plays an important role through the 9-10 nucleotide hybrid of nascent RNA and template DNA within RNA polymerase during RNA synthesis [2]. After incorporation, each new RNA nucleotide becomes part of both the growing RNA chain and the RNA-DNA hybrid that anchors the RNA to DNA. The RNA-DNA hybrid ensures that template DNA and nascent RNA are kept together in the RNA polymerase active site; the hybrid is therefore essential for the accurate base-by-base replication of the genetic information from DNA to RNA.

In translation, RNA-RNA base pairing plays an important part: for the amino acid chain to grow, a nucleotide triplet codon on mRNA must bind by base pairing to a complementary RNA anticodon on the incoming amino-acid charged tRNA. Additionally, the catalytic component of peptide bond formation in the ribosome consists only of RNA [3], making translation a highly RNA-dependent process.

Other biological processes which depend on base pairing include the formation of secondary and tertiary structures of folded RNA; micro-RNA regulation by hybridization to full-length RNA; and the binding of the 16S RNA of the bacterial ribosome to the Shine-Dalgarno sequence of mRNA to initiate protein synthesis.

These naturally evolved mechanisms of interactions between RNA and DNA show how nucleotide chain base pairing and hybridization are fundamental to cellular life. Additionally, nucleotide chain base pairing and hybridization also underlie many of the experimental techniques employed in molecular biology today. One example is polymerase chain reaction (PCR), which relies on the temperature-dependence of the melting and annealing of double stranded DNA. Another example is gene silencing, in which small interfering RNA (siRNA) can be constructed to base pair with coding regions of mRNA to interfere with protein synthesis. Another example are DNA microarrays, which rely on constructing optimal RNA probes for hybridizing with a target DNA sequence. Yet another example is the CRISPR/Cas DNA editing technology, which relies on a complementary RNA-DNA hybrid for precise excision of DNA fragments.

Regardless of the biological or technical function mediated by nucleotide base pairing, the function can be greatly influenced by the strength of the base pair bonds. In general, GC-rich sequences form stronger base pairs than sequences that are AT-rich, mainly due to differences in the stacking of the GC and AT nucleotides [4]. Look-up tables for the strength of RNA-DNA, DNA-DNA, and RNA-RNA dinucleotide pairs (in terms of Gibbs free energy of formation, ΔG) are available in the literature, and have a long history of use in computational biology, especially in the field of predicting RNA secondary structures [5].

Nucleotide chain interactions, as mentioned, play key roles in gene expression; from transcription initiation to translation. However, although the genetic code of more and more organisms has been accounted for, we still only have a limited understanding of how DNA and RNA sequence affects transcription and translation. In this thesis we have contributed to that understanding by investigating three aspects of gene expression that are affected by the binding strength between nucleotide chains and the binding strength between nucleotide chains and protein: transcription initiation, translation initiation, and post-transcriptional RNA processing.

We have studied translation initiation (Chapter 3) through calculations of RNA secondary structures using methods that rely on RNA-RNA base pairing to predict how secondary structures affect the binding of the bacterial ribosome to mRNA. By experimentally translating RNA sequences with different predicted secondary structure, this work has contributed to a greater understanding of several protein expression systems.

In the first study of initial transcription in this thesis (Chapter 4), we have used calculations of the binding strengths of double stranded DNA and the RNA-DNA hybrid in an equilibrium model of translocation to investigate the sequence dependence of productive transcription from a promoter. This led to the discovery of an association between the translocation of RNAP on DNA and the production of abortive RNA during transcription initiation. Based on this finding, transcription initiation experiments were performed which supported the computational findings. This has resulted in new knowledge about how the sequence of the $3'$ terminal end of the RNA affects promoter escape in bacteria.

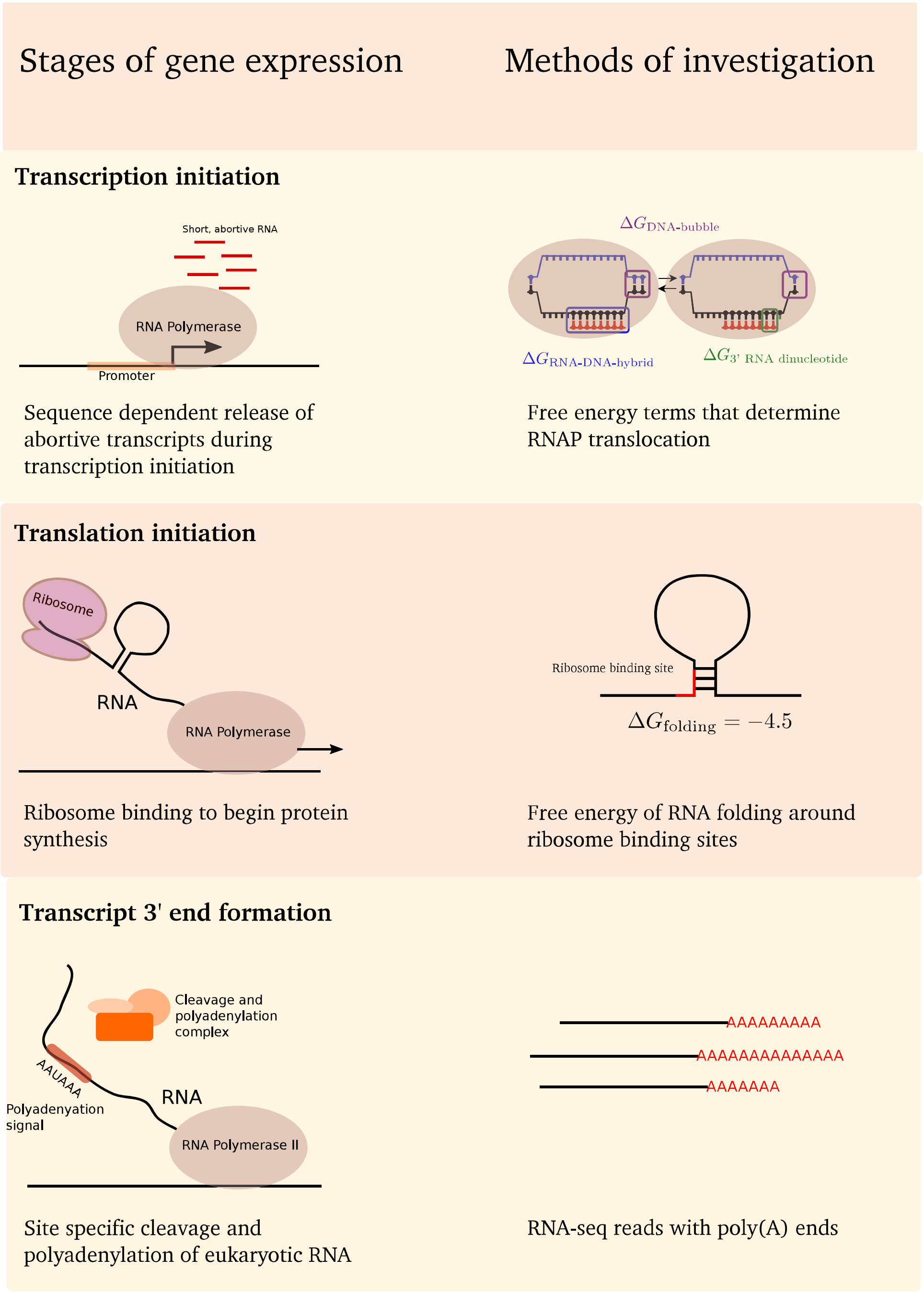
In the second study of initial transcription (Chapter 5, we have used a kinetic model of initial transcription to study the dynamical nature of abortive cycling. To inform the rate constant of backtracking, we used abortive probabilities found from bulk transcription experiments. By comparing model outcome to single-molecule experiments of abortive cycling, we found that the speed of initial transcription was the same as what has been reported for transcription elongation. This has contributed new knowledge to the dynamics of initial transcription.

In the final study in this thesis (Chapter 6), we used a genome-wide approach to study the regulation of gene expression by polyadenylation of

RNA. This work involved the analysis of nucleotide sequences; however, instead of using free energies, the focus was on analysing poly(A) sequences from high-throughput RNA sequencing (RNA-seq) data to study pre-RNA processing in the form of cleavage and polyadenylation. This work has broadened the view of the thesis by investigating the regulation of gene expression from a genome-wide perspective, as opposed to the gene-centric approach of the two first studies.

Figure 2.1 summarizes the different aspects of gene expression that are investigated in this thesis. On the left in this figure are the aspects of gene expression that have been studied, and on the right are the main computational methods that have been used to study them.

When studying the different aspects of gene expression we have employed two different research approaches: for the studies with free energy and secondary structure, a traditional gene-centric, single-molecule approach has been used; while for the study of $3'$ cleavage and polyadenylation a top-down, genome-wide approach has been taken. The utilization of these two different approaches, which can also be referred to as hypothesis-driven and data-driven, is interesting in its own right and will make up part of the final discussion at the end of the thesis.



Stages of gene expression

Methods of investigation

Transcription initiation

Short, abortive RNA
RNA Polymerase
Promoter
$\Delta G_{\text{DNA-bubble}}$
$\Delta G_{\text{RNA-DNA-hybrid}}$
$\Delta G_{\text{3' RNA dinucleotide}}$

Sequence dependent release of abortive transcripts during transcription initiation

Free energy terms that determine RNAP translocation

Translation initiation

Ribosome
RNA
RNA Polymerase
Ribosome binding site
$\Delta G_{\text{folding}} = -4.5$

Ribosome binding to begin protein synthesis

Free energy of RNA folding around ribosome binding sites

Transcript 3' end formation

Site specific cleavage and polyadenylation of eukaryotic RNA

RNA-seq reads with poly(A) ends

Figure 2.1

2.2 Transcription initiation in bacteria

Two articles by Skancke et al. cover the kinetics of transcription by promoter bound RNAP (chapter 5) and sequence-determinants for the efficiency of promoter escape (chapter 4). In this section we review the computational and biological aspects relevant for the articles. We cover promoter binding and transcription initiation in detail, and then briefly discuss the elongation phase of transcription. At the end of the section, we review computational modeling of transcription, both for initiation and elongation.

2.2.1 Biochemical background

Transcription is the synthesis of a complementary RNA molecule from a DNA template. The macromolecule that performs transcription is the DNA-dependent RNA polymerase (RNAP). See Figure 2.2 for an overview of transcription elongation by RNA polymerase.

In bacteria, RNAP consists of the subunits β, β', ω, and two α subunits and has a molecular mass of around 400 kDa. The structure and function of bacterial RNAPs are conserved in comparison with archaea and eukaryotes [6], showing that the fundamental mechanism of RNA synthesis is the same for all cell types.

Here, we will review the role of RNAP in initiation, elongation, and termination of transcription in bacteria. Special care will be given to the topic of abortive transcription initiation and how RNAP moves along DNA during transcription.

Promoter localization

Before RNAP can start transcription of a gene it must localize a transcription start site (TSS). These sites are advertised by promoters, stretches of DNA which are typically located from around -60 to +20 basepairs relative to the TSS [7, 8]. RNAP cannot bind strongly to promoters directly, but must first associate with a regulatory protein called sigma (σ) factor; the σ factor in turn only binds strongly to promoters when in complex with RNAP [9], making the RNAP-σ complex essential for transcription initiation. In *E. coli* there are seven types of σ factors, where each σ factor regulates the binding of RNAP to promoters associated with a particular gene family [10]. For example, σ^{70} recognizes promoters for housekeeping genes, while σ^{32} recognizes promoters for genes that are activated in conditions of heat shock. Most promoters are of the σ^{70} type (meaning that σ^{70} binds to them), and we will here focus on this type of promoter. The most

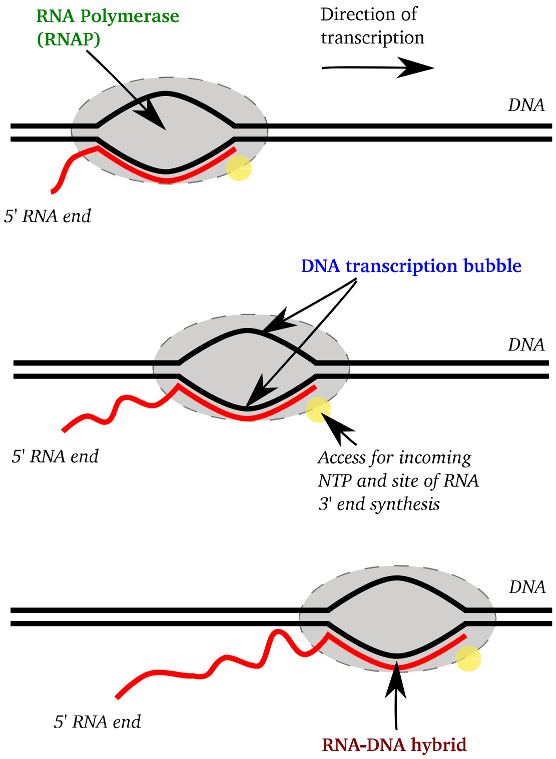


Figure 2.2 – Transcription elongation by RNA polymerase. The nascent RNA transcript grows at the 3′ end where NTP binds and is incorporated into the RNA nucleotide chain. NTP binds in a complementary fashion to bases on the single stranded template DNA which are exposed due to the $\sim$ 14/15 bp DNA transcription bubble.

prominent σ-binding elements in σ^{70} promoters are the -35 and -10 hexamer (six-nucleotide) motifs, with the consensus sequences TTGACA (-35) and TATAAT (-10) [11]. The major determinant for how strongly σ^{70} binds to a promoter is how much the -35 and -10 motifs vary from the consensus, as well as the distance in nucleotides between these elements. A third sequence that can affect promoter-RNAP-σ^{70} binding is an AT rich element upstream -35 called, simply, the upstream element [7, 12] (see Figure 2.3). In general, the stronger σ^{70} binds to a promoter the higher the transcription initiation rate will be, which again may lead to higher expression of the downstream gene. However, we will shortly see that the initial transcribed sequence, defined as the region from +1 to +20 realtive to the TSS (Figure 2.3), may also influence the expression from a promoter.

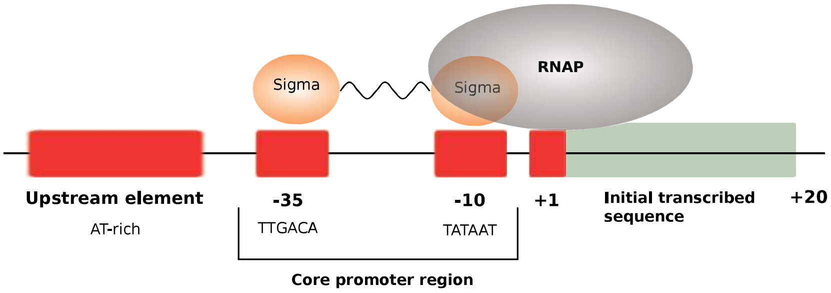


Figure 2.3 – RNAP and sigma bound to promoter elements. Positions along the DNA template are indicated relative to the transcription start site.

Transcription initiation

Once the RNAP-σ complex has bound to a promoter, transcription initiation may commence. First, σ mediates the opening of the double stranded DNA from -11 to +2 [6]. The unwound DNA helix is referred to as the DNA bubble, or the transcription bubble. Within the DNA bubble, bases on the template DNA strand are exposed and ready to base pair with incoming NTP. After having opened the DNA bubble, RNAP is positioned so that the nucleotide at the TSS is at RNAP's active site (the active site is the part of RNAP where RNA synthesis occurs). When transcription begins, the two first nucleotides bind in the active site and a phosphodiester bond is formed between them [13]. This constitutes the first dinucleotide of the nascent RNA. In order to incorporate the third nucleotide, RNAP must first move its active site relative to downstream DNA to make room for the incoming NTP. However, RNAP is bound to the promoter, making it immobile. It can therefore not translocate downstream as it would during transcription elongation. Instead, translocation during initiation results in RNAP pulling DNA into itself in a process that has been labeled "scrunching" [14, 15]. By pulling in downstream DNA, RNAP's active site translocates relative to template DNA, making the active site available for incoming NTP, even though RNAP is still bound to the upstream promoter sequence. A consequence of scrunching is that the DNA bubble increases in size with 1bp for each scrunching step, since one basepair of DNA is opened downstream without the complementary closing of an upstream basepair as for transcription elongation. The increasingly large DNA bubble that results from scrunching has been suggested to play a role for both promoter escape and abortive RNA release [8, 14].

Transcription initiation proceeds in a cycle of scrunching, NTP binding, and phosphodiester bond formation. With each step, the DNA bubble grows with one basepair and the nascent RNA grows with one nucleotide. When the nascent RNA reaches a length of around 10-12 nt, promoter escape may occur if σ dissociates from the promoter [8] (see Figure 2.4).

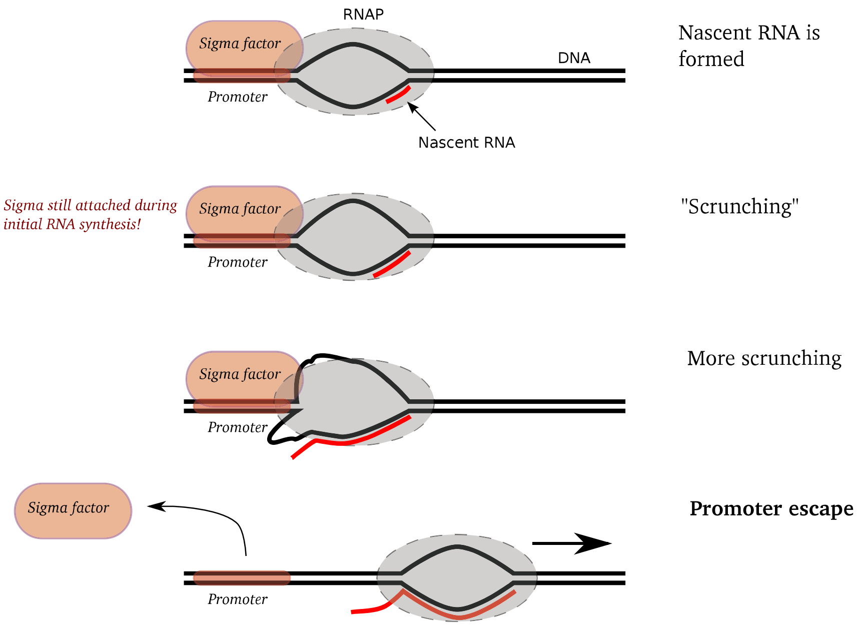


Figure 2.4 – Promoter escape involves scrunching of DNA.

Abortive initiation

Promoter escape is, however, only one outcome of a transcription initiation attempt. Another outcome is that the initiation attempt fails, which means that the nascent RNA is released prematurely before promoter escape can occur [16]. The starting point of a failed initiation attempt is a reversal of the translocation-scrunching reaction, a step which is referred to as backtracking [17]. When backtracking occurs, the “unschrunching“ of DNA forces the 3′ end of the nascent RNA into the entry channel for NTP, leading to a shortened RNA-DNA hybrid [17]. Eventually the nascent RNA is released from this complex, possibly because the shortened RNA-DNA hybrid is unstable. RNA release is assumed to be accompanied by a release of the rest of the scrunched DNA from RNAP, since it is known that after an abortive RNA release RNAP can fall back to the open complex formation, where it may restart transcription initiation *de novo* [8] (see Figure 2.5).

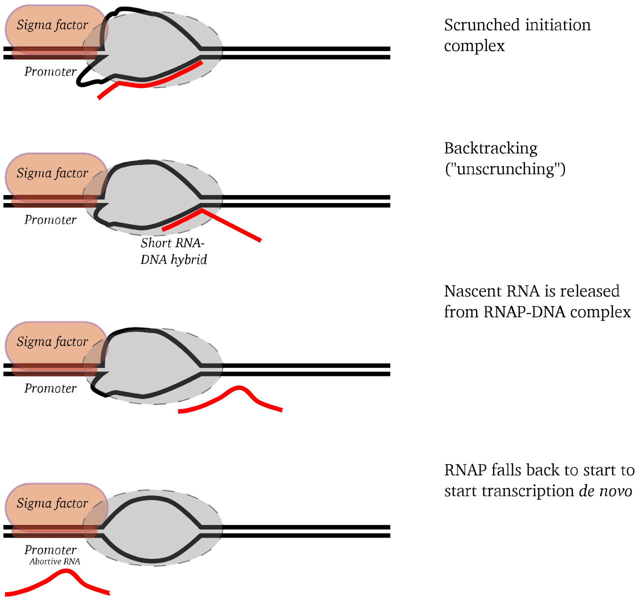

Sigma factor
Promoter
Scrunched initiation complex
Sigma factor
Promoter
Short RNA-DNA hybrid
Backtracking ("unscrunching")
Nascent RNA is released from RNAP-DNA complex
Sigma factor
Promoter
RNAP falls back to start to start transcription *de novo*
Sigma factor
Promoter
Abortive RNA

Figure 2.5 – Abortive initiation caused by a backtracking-“unscrunching” reaction.

The term for unscrunching and RNA release during transcription initiation is abortive initiation. A related term is abortive cycling, which refers to repeated initiation attempts that end with abortive RNA release (see Figure 2.6). The aborted RNA fragments are generally no longer than 15 nucleotides, but lengths up to 20 have been observed [18]. In *in vitro* experiments abortive initiation is the rule rather than the exception for many promoters [8]. Some promoters have a calculated *in vitro* ratio of aborted to successful transcription initiation attempts of over 300 [17]. This indicates that RNAP may on average spend considerable time in abortive cycling before achieving promoter escape.

Experimental detection of abortive initiation

To quantify abortive initiation, one must obtain a measure of the molar abundance of each abortive RNA while ensuring that abortive RNAs of a given length are quantified separately from any possible cleavage product of the same length (since a backtracked RNA can be cleaved at the RNAP active site before abortive release occurs). A method that fulfills these two requirements is radioactive labeling using nucleotide γ-phosphates [19]. Each abortive RNA will only contain one γ-phosphate, namely the one at

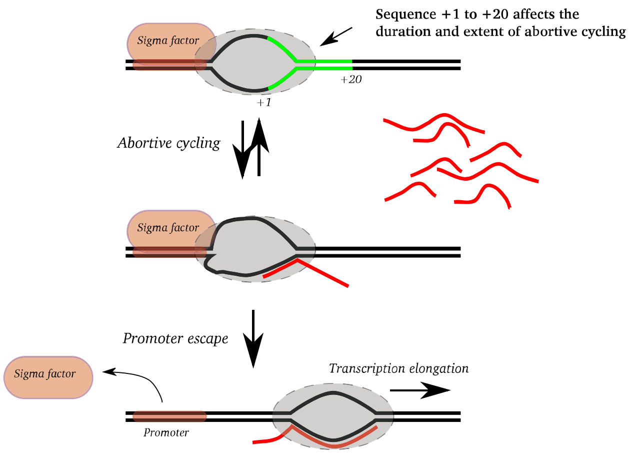


Figure 2.6 – Repeated abortive RNA release and *de novo* transcription initiation is called abortive cycling.

the 5′ end. All other γ-phosphates are cleaved off during transcription and released in pyrophosphate (PPi). By separating abortive RNAs in terms of their length on a single-nucleotide resolution gel, one can then use a phosphoimager to quantify the amount of each abortive product [19]. By this method, one can infer at which nucleotide positions abortive initiation occurs and to what extent [20]. Although cleaved RNA is contained in the gel, the lack of a radioactive phosphate means it will not affect the quantification of abortive RNAs.

Calculation of abortive probability

The quantification of abortive RNAs of different lengths permits the calculation of abortive probability (AP) at each template position, as described by Hsu [21]. Briefly, the probability to abort a transcript of length i is given as

$$\frac{R_i}{P_i},$$

where R_i is the number of abortive RNAs of length i and P_i is the number of polymerases that transcribe to reach an RNA of length i. This can be written as

$$\frac{R_i}{P_i} = \frac{r_i R}{p_i P}.$$

Here, r_i is the fraction of RNAs of length i in terms of total RNA (R), and p_i is the fraction of RNAPs transcribing an RNA of length i in terms of total

RNAPs (P). We assume that each observed aborted RNA corresponds to one transcribing RNAP, so that $R = P$. Therefore, we find the abortive probability as

$$AP_i = \frac{r_i}{p_i}.$$

The fraction of abortive RNAs of each length is found directly from the phosphoimaging quantification. The fraction of RNAP obtaining an RNA of length i is found iteratively in the following way: 100% of RNAPs obtain an RNA of length 2; if for example 40% of all abortive RNA is of length 2, then we know that 60% of RNAPs obtain an RNA of length 3; then, if for example 15% of all abortive RNA is of length 3, then we know that 45% of RNAPs obtain an RNA of length 4; and so on.

Promoter escape and the initial transcribed sequence

An early discovery by Kammerer et al. [22] was that the promoter sequence downstream +1 can have a strong effect on promoter strength. Kammerer et al. found this effect by comparing the promoter strength of the N25 phage promoter with a variant of N25 constructed by changing C for A, T for G and vice versa for N25's +1 to +20 sequence. This variant was called N25/anti, and has later been referred to as N25$_{\text{anti}}$ [**hsu_vitro>_2003**]. Later, it was found that the cause for the difference in promoter strength was due to a large increase in both the amount and length of abortive transcripts produced from N25/anti compared to N25 [**hsu_vitro>_2003**, 23]. To comprehensively map the effect of the +1 to +20 sequence on promoter escape, Hsu et al. [17] investigated *in vitro* the abortive properties of a library of 43 promoter variants which had randomized +1 to +20 initial transcribed sequences (ITSs). They showed that sequence variation in the ITS could result in a 20-fold difference in promoter escape efficiencies. In the study by Hsu et al. promoter escape efficiency is defined as the fraction of full length transcript (i.e., not abortive) to the total amount of transcript (full length and abortive) produced from a promoter; this quantity is also referred to as the productive yield (PY).

For a while it was speculated that abortive initiation was an *in vitro* artefact. However, small abortive RNAs have been identified *in vivo* [24], demonstrating that this process occurs in living cells. Following the discovery that abortive transcripts appear *in vivo*, it was not certain whether the abortive transcripts had any cellular function or if they were merely artefacts of the transcription initiation process. However, two studies have shown two different functions of these short transcripts. In one, aborted transcripts from the $\phi10$ promoter were found to deactivate a transcriptional terminator hairpin [25]. In the other, short abortive products of 2 to 4 nucleotides in length were found to act as primers for the RNA polymerase *in vivo* [26]; previously, it was not known if RNAP, unlike the DNA polymerase, could use primers *in vivo.*

It is still not clear if abortive initiation is rate-limiting for natural promoters *in vivo*. However, *in vitro* and *in vivo* experiments in *E. coli* have shown that the transcription factors GreA and GreB greatly reduce the amount of abortive initiation from the $\mathrm{N25_{anti}}$ promoter [23], which they do by restoring backtracked RNAP to productive RNAP by stimulating RNAP's intrinsic mechanism for cleaving backtracked RNA [17, 27]. This suggests that abortive initiation has the potential to be rate limiting *in vivo*, but that this potential is countered by the expression of GreA/B. In support of an *in vivo* role for GreA in mitigating abortive initiation, it was found that GreA resolves promoter-proximal stalling of RNAP [28]. However, more work is still needed to confirm the precise role GreA/B play in promoter escape *in vivo.*

Transcription elongation

Once the σ-promoter bonds are broken, RNAP is free to undergo processive transcription elongation. However, even though σ has broken contacts with the promoter, it is still in a complex with RNAP immediately after promoter escape, although the complex is weakened as the nascent RNA has displaced parts of the RNAP-σ interactions on its way to the RNA exit channel of RNAP [29, 30]. It is thought that σ is released stochastically from this weakened complex, as σ is intermittently found retained with RNAP in far downstream sequences [31]. Several studies have found that as a consequence of σ^{70} remaining bound to RNAP after promoter escape, the still-attached σ can rebind promoter-like elements during transcription elongation, causing RNAP to pause in its track [32–34]. To escape from these pauses, it has been suggested that RNAP-σ must again undergo scrunching as if it were bound to a promoter at a transcription start site [35].

Transcription elongation (Figure 2.2) happens with great processivity: RNAP can accurately transcribe tens of thousands of nucleotides without dissociating from DNA. In spite of this processivity, elongation does not occur at a constant rate. RNAP will reproducibly pause or backtrack at certain sites [36]. Sometimes this pausing or backtracking has a regulatory function, for example to allow time for proper folding of the nascent RNA [37]. Transcription elongation is therefore not just a mandatory step for copying DNA to RNA, but also another stage of gene expression where regulation takes place.

Transcription termination

Eventually, RNAP will dissociate DNA and release its RNA product. Two distinct mechanisms have been identified for the release of RNA from RNAP. In one, the protein Rho binds an unstructured region of the nascent RNA and moves along RNA in the direction of RNAP until they meet at RNAP pause sites. At these pause sites the interaction between Rho and RNAP causes the release of both RNA and RNAP from the DNA template [38]. The Rho-independent mechanism of termination begins by the formation of a strong (GC-rich) RNA hairpin on the nascent RNA right outside the RNA exit channel. If this hairpin is followed by a downstream A-rich sequence on DNA which destabilizes the RNA-DNA hybrid, interactions between the hairpin and RNAP cause RNAP to release its hold on the RNA. However, the details of the process are not clear [39].

In both cases, once RNA has been released, the affinity of RNAP to DNA is greatly reduced and RNAP itself disengages DNA. When RNAP is released, it is again free to associate with a σ factor and begin transcription anew.

2.2.2 Computational models

A computational model of transcription can be used alongside wet-lab experiments to improve our knowledge of how RNA synthesis occurs. Such a model has two main uses: the first is to evaluate our existing conceptual models of transcription by writing that conceptual model in a mathematical language and comparing the resulting computational model with the available experimental data. If the mathematical description cannot fit the data, it is a sign that the conceptual model may be incomplete. Another use of such a model is to experiment with the computational model: if a modification of the computational model gives an improved fit to experimental data, it could be worthwhile to investigate if that modification can be validated through wet-lab experiments.

The nucleotide addition cycle and the equilibrium assumption of translocation

The core of the computational models of transcription is a mathematical description of the nucleotide addition cycle (NAC). In a two-step description, the NAC consists of a reversible translocation step, where the RNAP active site is made available for NTP binding, and a synthesis step, where an incoming bound NTP is added to the 3′ end of the growing RNA chain (Figure 2.7). While the NTP binding-and-synthesis step is reversible through pyrophosohorolysis, this is highly unfavorable under typical experimental conditions [40].

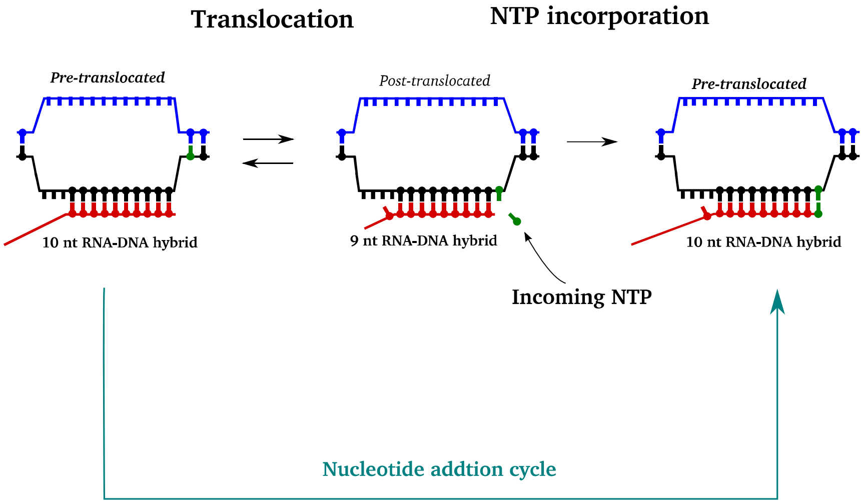


Figure 2.7 – The nucleotide incorporation cycle proceeds through translocation and NTP incorporation. Before NTP binding, RNAP is in the pre-translocated state. RNAP then translocates, thereby entering the post-translocated state. In the post-translocated state, NTP can bind. After the incoming NTP has bound and become incorporated onto the 3′ end of the RNA, RNAP is once again in the pre-translocated state.

For most published computational models of transcription, a key assumption about the NAC is that the reversible translocation step attains equilibrium before the NTP synthesis step [41–43]. While it is difficult to conclusively demonstrate that the equilibrium assumption always holds true, it does facilitate the description of translocation with an equilibrium equation:

$$Keq = e^{-\frac{\Delta G_{\text{RNA-DNA}} + \Delta G_{\text{DNA-DNA}} + \Delta G_{\text{RNAP}}}{RT}}. \tag{2.1}$$

This equation contains the terms currently believed to contribute energetically to translocation. These are the free energy change of the RNA-DNA hybrid ($\Delta G_{\text{RNA-DNA}}$), the DNA-DNA transcription bubble ($\Delta G_{\text{DNA-DNA}}$), as well as a term for non-specific interactions between RNAP, DNA, and RNA (ΔG_{RNAP}) [41]. The latter term is not known and has previously been set to a constant value [44, 45].

By calculating the equilibrium constant of translocation, one finds the equilibrium balance between the pre-translocated and post-translocated states. This constant can in the next step be used in a model of transcription to calculate the probabilities of NTP binding, pausing, or backtracking [41–43]. Thus, the calculation of translocation is central to the description of transcription.

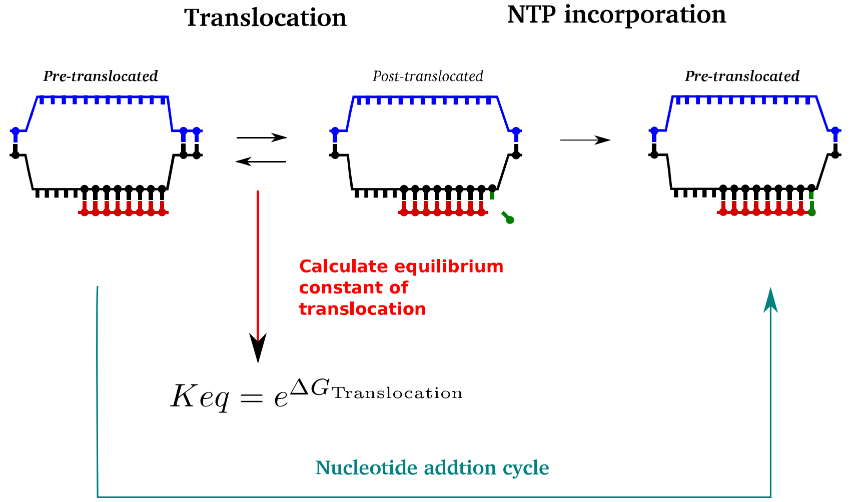


Figure 2.8 – The equilibrium constant of translocation is a simple model of the movement of RNAP along DNA during transcription.

This conceptual model of RNAP movement as an equilibrium reaction seems orderly: one needs only to calculate the change in free energy of $\Delta G_{\text{DNA-DNA}}$ and $\Delta G_{\text{RNA-DNA}}$ from available energy tables [46, 47] to find out if RNAP will move forward, backtrack, pause, or terminate at any given location on DNA. However, it is not known how large the contribution of the ΔG_{RNAP} term to equation (2.1) is, or if this term is sequence dependent. One way to investigate the veracity of the conceptual model of RNAP movement is by using the equilibrium equation for computational modelling of transcription. This enables a more rigorous testing of the conceptual model by comparing the computational model results with experimental data.

Computational models of transcription elongation

Several kinetic and thermodynamic models of transcription elongation have been published [43–45, 48]. What they have in common is that in some form they incorporate the terms from (2.1) and calculate the $\Delta G_{\text{RNA-DNA}}$

and $\Delta G_{\text{DNA-DNA}}$ terms from published tables of these values for different dinucleotide combinations. In the case of Tadigotla et al. [44] the free energy from the nascent RNA secondary structure close to the RNA exit channel is used as well; this calculation was used to model the effect RNA structures outside RNAP have in preventing backtracking of RNAP [49].

The early computational models of transcription had only partial predictive power. For example, Tadigotla et al. [44] predict pause sites during transcription elongation. While their best optimized model manages to identify 84% of pauses, only 68 % of their predicted pause sites overlapped with experimentally identified pause sites, indicating that there were a number of false positives. The model of Bai et al. [42] was published without any statistical measures, making it difficult to interpret. Instead, Tadigotla et al. implemented the algorithm from Bai et al. and found that the Bai et al. model performed only little better than random for detecting pausing [44]. The performance of these models signalled that more work was needed to reach a mature and accurate model for transcription.

Recently, Maoiléidigh et al. published a model of transcription elongation which fitted well several transcription parameters that had previously been measured by single-molecule experiments [50]. To adapt their model to the results, however, they added to the model an intermediary state between processive translocation and backtracking which has not been observed experimentally [50]. Their results suggests that translocation does not occur as an equilibrium reaction, which contradicts the this previously made assumption. While this is the most up to date model of transcription to date, this model still relies only on the free energy change of the DNA bubble and RNA-DNA hybrid to calculate translocation [50]. As will be discussed below, it is now known that more terms than these influence this process, which necessitate an additional free energy variable for the description of translocation.

Recently, Hein et al. [51] showed that the sequence of the $3'$ dinucleotide of the nascent RNA has a strong effect on translocation rates. Hein et al. found that if the $3'$ nucleotide of the RNA was U, RNAP had a much stronger preference for the pre-translocated state than if the $3'$ nucleotide was G. This finding was supported by Malinen et al. [52] who proposed that contacts between the RNA $3'$ end and the RNAP active site determine the preference for the pre-translocated or the post-translocates states. The findings by Hein et al. and Malinen et al. have the potential to greatly improve on the existing free energy calculation of translocation, since up until now most work has focused on the free energy change of the RNA-DNA hybrid and the DNA bubble.

A computational model of transcription initiation

In addition to the above described models of transcription elongation, one model bye Xue et al. [53] has previously been published for transcription initiation.

For the sake of modeling, the key differences between transcription initiation and transcription elongation is that during initiation i) the DNA-DNA bubble grows with scrunching-translocation instead of being constant as for elongation ii) the RNA-DNA hybrid lacks its full length until the active site has reached +9/10, and iii) that RNA secondary structures outside RNAP do not influence backtracking during initiation since these structures cannot form until after promoter escape when a sufficient length of RNA has emerged from the RNA exit channel. By incorporating these changes into equation (2.1), one can model transcription initiation in a similar manner as for elongation.

The model by Xue et al. takes as input the ITS of a promoter and predicts measurable output such as the probability to abortively release nascent RNA at a given position, and the ratio of abortive to successful initiation attempts. The model is able to accurately predict the maximum size of abortive transcript and the abortive to productive ratio of the N25 promoter. However, the model does not mange to predict the abortive probabilities for different lengths of the nascent RNA. Some criticism of this study is that only the data from the N25, $\mathrm{N25}_{\mathrm{anti}}$ and T7A1 promoters were investigated, even though an equivalent but more comprehensive dataset with 43 ITS variants was available [17]. Further, the T7A1 promoter has sequence variation in the core promoter sequence compared to the other two, which only vary in the ITS. Variation in core promoter sequence has been shown to affect promoter escape properties [54], yet the model does not take this into account. This makes it difficult to compare the model output for the analysis of the three different promoters. Finally, the model by Xue et al. was not used predictively, in the sense that it was not used to predict abortive initiation properties of new DNA sequences which were then tested in the lab.

In conclusion, the study by Xue et al. was the first quantitative model for transcription initiation, and the model managed to predict some experimentally measured parameters, although the lack of predictive usage of the model makes it difficult to evaluate. Therefore, in the field of transcription initiation, it remains to be published a quantitative model with predictive power that provides new insight into the process itself.

Bibliography

[1] J. D. Watson and F. H. Crick. “Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid”. *Nature* 171.4356 (Apr. 1953), pp. 737–738.

[2] Dmitry G. Vassylyev et al. “Structural basis for substrate loading in bacterial RNA polymerase”. *Nature* 448.7150 (July 2007), pp. 163–168.

[3] Thomas A. Steitz and Peter B. Moore. “RNA, the first macromolecular catalyst: the ribosome is a ribozyme”. *Trends in Biochemical Sciences* 28.8 (Aug. 2003), pp. 411–418.

[4] Peter Yakovchuk, Ekaterina Protozanova, and Maxim D. Frank-Kamenetskii. “Base-stacking and base-pairing contributions into thermal stability of the DNA double helix”. *Nucleic Acids Research* 34.2 (Jan. 2006), pp. 564–574.

[5] David H. Mathews and Douglas H. Turner. “Prediction of RNA secondary structure by free energy minimization”. *Current Opinion in Structural Biology* 16.3 (June 2006), pp. 270–278.

[6] Sergei Borukhov and Evgeny Nudler. “RNA polymerase: the vehicle of transcription”. *Trends in Microbiology* 16.3 (Mar. 2008), pp. 126–134.

[7] Wilma Ross et al. “A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase”. *Science* 262.5138 (Nov. 1993), pp. 1407–1413.

[8] Lilian M Hsu. “Promoter clearance and escape in prokaryotes”. *Biochimica et Biophysica Acta* 1577.2 (Sept. 2002), pp. 191–207.

[9] Mark SB Paget and John D. Helmann. “The ^{70} family of sigma factors”. *Genome Biology* 4.1 (Jan. 2003), p. 203.

[10] Sofia Österberg, Teresa del Peso-Santos, and Victoria Shingler. “Regulation of Alternative Sigma Factor Use”. *Annual Review of Microbiology* 65.1 (2011), pp. 37–55.

[11] Wilma Ross and Richard L. Gourse. “Analysis of RNA Polymerase-Promoter Complex Formation”. *Methods* 47.1 (Jan. 2009), pp. 13–24.

[12] Shanil P. Haugen et al. “Fine structure of the promoter-sigma region 1.2 interaction”. *Proceedings of the National Academy of Sciences* 105.9 (Mar. 2008), pp. 3292–3297.

[13] W. R. McClure, C. L. Cech, and D. E. Johnston. "A steady state assay for the RNA polymerase initiation reaction." *Journal of Biological Chemistry* 253.24 (Dec. 1978), pp. 8941–8948.

[14] Achillefs N. Kapanidis et al. "Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism". *Science* 314.5802 (Nov. 2006), pp. 1144–1147.

[15] Andrey Revyakin et al. "Abortive initiation and productive initiation by RNA polymerase involve DNA scrunching". *Science* 314.5802 (Nov. 2006), pp. 1139–1143.

[16] Agamemnon J. Carpousis and Jay D. Gralla. "Cycling of ribonucleic acid polymerase to produce oligonucleotides during initiation in vitro at the lac UV5 promoter". *Biochemistry* 19.14 (July 1980), pp. 3245–3253.

[17] Lilian M. Hsu et al. "Initial transcribed sequence mutations specifically affect promoter escape properties". *Biochemistry* 45.29 (July 2006), pp. 8841–8854.

[18] Monica Chander et al. "An alternate mechanism of abortive release marked by the formation of very long abortive transcripts". *Biochemistry* 46.44 (2007), pp. 12687–12699.

[19] Lilian M. Hsu. "Monitoring abortive initiation". *Methods* 47.1 (Jan. 2009), pp. 25–36.

[20] Lilian M. Hsu et al. "*in vitro* studies of transcript initiation by *Escherichia coli* RNA polymerase. 1. RNA chain initiation, abortive initiation, and promoter escape at three bacteriophage promoters". *Biochemistry* 42.13 (Apr. 2003), pp. 3777–3786.

[21] Lilian M. Hsu. "Quantitative parameters for promoter clearance". *Methods in Enzymology* Volume 273 (1996). Ed. by Sankar Adhya, pp. 59–71.

[22] W. Kammerer et al. "Functional dissection of *Escherichia coli* promoters: information in the transcribed region is involved in late steps of the overall process." *The EMBO Journal* 5.11 (Nov. 1986), pp. 2995–3000.

[23] Lilian M Hsu, Nam V. Vo, and Michael J Chamberlin. "*Escherichia coli* transcript cleavage factors GreA and GreB stimulate promoter escape and gene expression *in vivo* and *in vitro*." *Proceedings of the National Academy of Sciences* 92.25 (Dec. 1995), pp. 11588–11592.

[24] Seth R. Goldman, Richard H. Ebright, and Bryce E. Nickels. "Direct Detection of Abortive RNA Transcripts in Vivo". *Science* 324.5929 (May 2009), pp. 927–928.

[25] Sooncheol Lee, Huong Minh Nguyen, and Changwon Kang. "Tiny abortive initiation transcripts exert antitermination activity on an RNA hairpin-dependent intrinsic terminator". *Nucleic Acids Research* 38.18 (Oct. 2010), pp. 6045–6053.

[26] Seth R. Goldman et al. "NanoRNAs Prime Transcription Initiation In Vivo". *Molecular Cell* 42.6 (June 2011), pp. 817–825.

[27] Francine Toulme et al. “GreA and GreB proteins revive backtracked RNA polymerase *in vivo* by promoting transcript trimming”. *The EMBO Journal* 19.24 (Dec. 2000), pp. 6853–6859.

[28] Yoko Kusuya et al. “Transcription Factor GreA Contributes to Resolving Promoter-Proximal Pausing of RNA Polymerase in Bacillus subtilis Cells”. *Journal of Bacteriology* 193.12 (June 2011), pp. 3090–3099.

[29] Vladimir Mekler et al. “Structural Organization of Bacterial RNA Polymerase Holoenzyme and the RNA Polymerase-Promoter Open Complex”. *Cell* 108.5 (Mar. 2002), pp. 599–614.

[30] Bryce E. Nickels et al. “The Interaction Between ^{70} and the -Flap of *Escherichia coli* RNA Polymerase Inhibits Extension of Nascent RNA During Early Elongation”. *Proceedings of the National Academy of Sciences* 102.12 (Mar. 2005), pp. 4488–4493.

[31] Rachel Anne Mooney, Seth A. Darst, and Robert Landick. “Sigma and RNA Polymerase: An On-Again, Off-Again Relationship?” *Molecular Cell* 20.3 (Nov. 2005), pp. 335–345.

[32] Brian Z. Ring, William S. Yarnell, and Jeffrey W. Roberts. “Function of E. coli RNA Polymerase Factor- ^{70} in Promoter-Proximal Pausing”. *Cell* 86.3 (Aug. 1996), pp. 485–493.

[33] Achillefs N. Kapanidis et al. “Retention of Transcription Initiation Factor ^{70} in Transcription Elongation: Single-Molecule Analysis”. *Molecular Cell* 20.3 (Nov. 2005), pp. 347–356.

[34] Marni Raffaelle et al. “Holoenzyme Switching and Stochastic Release of Sigma Factors from RNA Polymerase In Vivo”. *Molecular Cell* 20.3 (Nov. 2005), pp. 357–366.

[35] Ekaterina Zhilina et al. “Structural transitions in the transcription elongation complexes of bacterial RNA polymerase during -dependent pausing”. *Nucleic Acids Research* 40.7 (Apr. 2012), pp. 3078–3091.

[36] Kristina M. Herbert et al. “Sequence-resolved detection of pausing by single RNA polymerase molecules”. *Cell* 125.6 (June 2006), pp. 1083–1094.

[37] R. Landick. “The regulatory roles and mechanism of transcriptional pausing”. *Biochemical Society transactions* 34.Pt 6 (Dec. 2006), pp. 1062–1066.

[38] M. Sofia Ciampi. “Rho-dependent terminators and transcription termination”. *Microbiology* 152.9 (Sept. 2006), pp. 2515–2528.

[39] Evgeny Nudler and Max E. Gottesman. “Transcription termination and anti-termination in *E. coli*”. *Genes to Cells* 7.8 (Aug. 2002), pp. 755–768.

[40] Lu Bai, Alla Shundrovsky, and Michelle D. Wang. “Chapter 9. Kinetic Modeling of Transcription Elongation”. In: *RNA polymerases as molecular motors.* Ed. by Henri Buc and Terence Strick. Cambridge: Royal Society of Chemistry, 2009, pp. 263–280.

[41] Sandra J. Greive and Peter H. von Hippel. “Thinking quantitatively about transcriptional regulation”. *Nature Reviews Molecular Cell Biology* 6.3 (Mar. 2005), pp. 221–232.

[42] Lu Bai, Robert M. Fulbright, and Michelle D. Wang. “Mechanochemical Kinetics of Transcription Elongation”. *Physical Review Letters* 98.6 (Feb. 2007), p. 068103.

[43] Richard Guajardo and Rui Sousa. “A model for the mechanism of polymerase translocation”. *Journal of Molecular Biology* 265.1 (Jan. 1997), pp. 8–19.

[44] Vasisht R. Tadigotla et al. “Thermodynamic and kinetic modeling of transcriptional pausing”. *Proceedings of the National Academy of Sciences* 103.12 (Mar. 2006), pp. 4439–4444.

[45] Lu Bai, Alla Shundrovsky, and Michelle D. Wang. “Sequence-dependent kinetic model for transcription elongation by RNA polymerase”. *Journal of Molecular Biology* 344.2 (Nov. 2004), pp. 335–349.

[46] Peng Wu, Shu Nakano, and Naoki Sugimoto. “Temperature dependence of thermodynamic properties for DNA/DNA and RNA/DNA duplex formation”. *European Journal of Biochemistry* 269.12 (June 2002), pp. 2821–2830.

[47] John SantaLucia and Donald Hicks. “The thermodynamics of DNA structural motifs”. *Annual Review of Biophysics and Biomolecular Structure* 33.1 (2004), pp. 415–440.

[48] Thomas D. Yager and Peter H. von Hippel. “A thermodynamic analysis of RNA transcript elongation and termination in *Escherichia coli*”. *Biochemistry* 30.4 (1991), pp. 1097–1118.

[49] Bradley Zamft et al. “Nascent RNA Structure Modulates the Transcriptional Dynamics of RNA Polymerases”. *Proceedings of the National Academy of Sciences* 109.23 (June 2012), pp. 8948–8953.

[50] Dáibhid O Maoileidigh et al. “A unified model of transcription elongation: what have we learned from single-molecule experiments?” *Biophysical Journal* 100.5 (Mar. 2011), pp. 1157–1166.

[51] Pyae P. Hein, Murali Palangat, and Robert Landick. “RNA transcript 3 -proximal sequence affects translocation bias of RNA polymerase”. *Biochemistry* 50.32 (2011), pp. 7002–7014.

[52] Anssi M. Malinen et al. “Active site opening and closure control translocation of multisubunit RNA polymerase”. *Nucleic Acids Research* 40.15 (Aug. 2012), pp. 7442–7451.

[53] Xiao Xue, Fei Liu, and Zhong Ou-Yang. “A kinetic model of transcription initiation by RNA polymerase”. *Journal of Molecular Biology* 378.3 (May 2008), pp. 520–529.

[54] Nam V. Vo et al. “*In vitro* studies of transcript initiation by *Escherichia coli* RNA polymerase. 3. Influences of individual DNA elements within the promoter recognition region on abortive initiation and promoter escape”. *Biochemistry* 42.13 (Apr. 2003), pp. 3798–3811.

2.3 Translation initiation in bacteria

The paper by Kucharova et al. on page 111 deals with gene expression systems in bacteria and how these systems are regulated by codon usage bias, mRNA stability, and RNA secondary structures at the ribosome binding site (RBS). Here, we review each of these topics. We begin with a brief account of translation initiation and then talk about how this process is regulated by RNA secondary structures. We then proceed to translation elongation and how this process is affected by condon usage bias.

Brief overview of translation

Translation is the conversion of the genetic code of an mRNA into an amino acid sequence. The molecule responsible for this is the ribosome, a macromolecular complex of RNA and protein. The bacterial ribosome consists of two subunits, S30 and S50, which initiate translation by binding sequentially (first S30, then S50) close to a start codon in the $5'$ region of an mRNA. Together, they read the genetic code in the form of nucleotide triplets (e.g. AAG) called codons. Each codon is matched with an anticodon in a transfer RNA (tRNA) which carries the amino acid that corresponds to this codon. The amino acids from the tRNAs are joined to each other in the ribosome to form the primary protein sequence which eventually will fold into a sequence specific conformation.

Since bacteria have no nucleus, the ribosome can bind the $5'$ -end of an mRNA at the same time as the mRNA is being synthesized. The coupling of translation with transcription is called co-transcriptional translation. The first ribosome that initiates translation on an mRNA follows the transcribing RNA polymerase closely and even pushes it, causing both of them to follow the same speed [1]. In this way, the ribosome can prevent RNAP from backtracking, which can increase the speed of transcription by reducing RNAP pausing. Fascinatingly, it was recently shown that by preventing RNAP backtracking in this way, ribosomes reduce collisions between RNAP and DNA polymerases during chromosome duplication, and that reducing these collisions leads to fewer errors during genome duplication [2]. This links the seemingly disparate topics of translation and genome integrity and is an example of the interconnectedness of the cell.

Translation initiation: binding of the 16S RNA to the Shine-Dalgarno sequence

The 16S RNA is a ribosomal RNA which is part of the S30 ribosome subunit and is important for translation initiation. It was shown by Shine and Dalgarno in 1974 that the last nine bases of the 3′ end of the 16S RNA in *E. coli* are ACCUCCUUA [3]. They suggested that this sequence could hybridize to a previously discovered conserved complementary motif close to the start codon in the 5′ untranslated region (UTR) of mRNA, and that this hybridization facilitates the initiation of translation [3]. That this actually occurs was later confirmed and it is now known to be the canonical method of translation initiation in bacteria and archaea [4]. Examples of translation initiation occurring without 16S RNA hybridization have been found, but they are exceptions rather than the rule [5, 6].

The complementary motif on mRNA that 16S RNA hybridizes to is known as the Shine-Dalgarno sequence, and the complementary bases on 16S RNA are known as the anti Shine-Dalgarno sequence. The Shine-Dalgarno (SD) sequence is located 3 to 10 bases upstream the start codon. Its sequence is often given as GGAGGA although the core motif can be reduced to GGAG. This sequence and its position relative to the start codon are conserved across prokaryotes with very little variation [4].

When the 16S RNA hybridizes with the SD sequence during translation initiation, the S30 subunit becomes physically anchored to the transcript. The binding of the S30 subunit facilitates the binding of the S50 subunit, which completes the binging of the ribosome to mRNA. The ribosome now covers an area +/- 15 nt around the start codon; this area is called the ribosome binding site (RBS) [7] (see Figure 2.9). When the 30S subunit binds the SD sequence, the start codon on mRNA (mostly AUG in *E. coli*) becomes aligned to the peptidyl site on S30. This allows the first initiating tRNA to bind and begin amino acid chain synthesis.

Since the SD sequence must basepair with 16S RNA, translation initiation can be blocked by having the SD sequence basepair with another RNA sequence. There are two main types of RNA that bind the SD sequence to block translation initiation. One type are trans-acting factors like microRNAs [8]. The other, which is the most common source of RNA that binds the SD sequence, is cis-acting RNA from the same mRNA molecule that the ribosome is binding to. Cis-acting RNA can bind to the SD if a sequence on the mRNA is complementary to the SD sequence (Figure 2.10). This type of RNA folding is a potent regulator of transcription initiation [9, 10], and will now be covered in more detail.

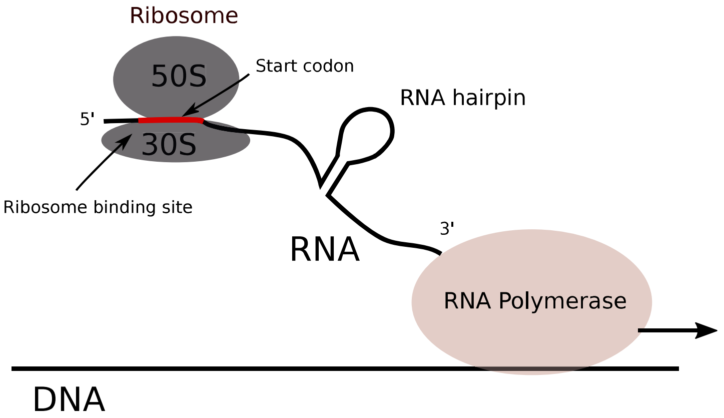


Figure 2.9 – Translation initiation. The 30S subunit binds the Shine-Dalgarno sequence, upon which the 50S subunit binds the 30S subunit to form the full ribosome. The bound ribosome occupies the ribosome binding site around the start codon.

Co-transcriptional RNA folding

To understand how the folding of RNA in the RBS affects translation initiation, it is necessary to understand how the folding of mRNA occurs in general. The basis for RNA folding is that RNA, as DNA, can basepair with itself; C pairs with G and A pairs with U. The reason why RNA folds under physiological conditions is that base-paired nucleotides are more thermodynamically favorable than free nucleotides [11]. Folded RNA is often depicted as a series of so called hairpins with base-paired stems and loops on top. Hairpins are examples of RNA secondary structures. More complex tertiary RNA structures can also form by the hybridization of secondary structures. However, tertiary structures form much more slowly than secondary structures [11], and are therefore not as relevant in obstructing ribosome binding.

As the nascent mRNA emerges from the RNA exit channel of bacterial RNAP it is immediately free to fold into an energetically favorable conformation. The folding of the mRNA at the same time as it is being transcribed is called co-transcriptional folding. A key question in co-transcriptional RNA folding has been how comparable the time scales for RNA folding and RNA synthesis are [12]. The degree of RNA co-transcriptional folding depends on the timescale for folding relative to the time scale for RNA synthesis. In bacteria, RNA synthesis happens at a rate of between 20 and 80 ms per nucleotide, while formation and dissociation of semi-stable structures like helices occur on the 10 to 100 μs timescale [13]. In other words, for each nucleotide synthesized, there is on average time for 1000 refolding events. It

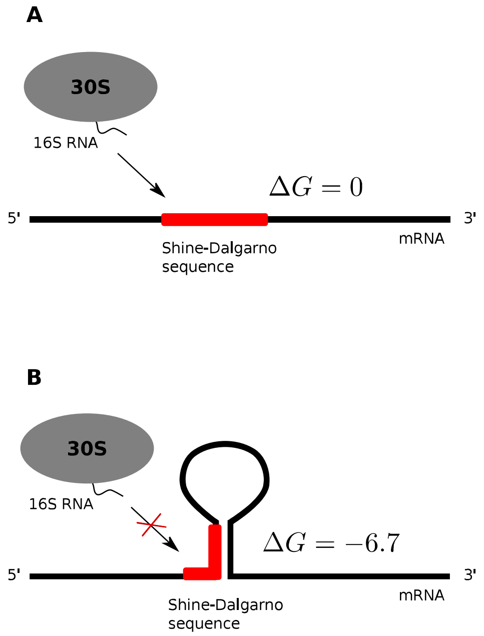


Figure 2.10 – Binding of the ribosome to mRNA can be blocked if a secondary structure is formed using the SD sequence.

should however be noted that the time needed for the spontaneous refolding of an RNA structure depends on binding strength of that structure; only relatively weak structures can be expected to change during co-transcriptional folding.

The exact order of folding steps the mRNA undergoes as it is synthesized is called the RNA's folding pathway. It is assumed that the folding pathway of an RNA is under selective pressure to either ensure or prevent that certain transient or permanent secondary structures are formed [14]. An example of such selection could be a folding pathway that avoids secondary structures in the RBS, which could be the result of a selective pressure for a high translation initiation rate at this RBS.

RNA secondary structures in the ribosome binding site

It was shown already in the early 80s that translation rates are strongly affected by secondary structures at the RBS [9]. In a seminal study, Smit and Duin modified both the location and the binding strength of secondary structures in the RBS of an mRNA to find that the level of the expressed protein could be varied over 500 fold [10]. They concluded that this reflected a variation in the efficiencies of translation initiation on the mRNA caused by these changes. In particular, they found a non-linear relationship between the folding energy of structures in the RBS and the translation initiation rate: above a certain threshold, the translation initiation rate did not change with respect to the strength of the secondary structure; but below that limit, translation initiation decreased exponentially as the secondary structures in the RBS got stronger [10].

In general, it is thought that the entire RBS sequence must be unstructured for translation initiation to occur [15]. It is therefore not surprising that the absence of strong secondary structures is a hallmark of ribosome binding sites in several species [16]. The folding energy of the RBS in the *E. coli* transcriptome can even be used to distinguish between active genes and pseudogenes [17]. This may be because pseudogenes are no longer under selective pressure to avoid strong secondary structures in the RBS as they no longer code for functional proteins and therefore do not need to be translated by ribosomes [17].

It should be noted that although the RBS is generally void of strong secondary structures, there are many examples of structured ribosome binding sites [18]. To account for how the ribosome would bind to mRNA with structures in the RBS, it was initially suggested that the ribosomes would only bind when the structures in the RBS spontaneously unfolded [19]. However, it was later appreciated that the time scale for RNA folding and unfolding are orders of magnitude faster than the time scale for ribosome binding to RNA. This led to the suggestion that ribosomes could first bind to a so called ribosome standby site close to the RBS from where they could approach the RBS when the secondary structures there unfolded [12]. This has however not been conclusively demonstrated, but was partially supported by a study that showed that that the ribosome, together with translation initiation factors, may unwind secondary structures at the RBS presumably from ribosome standby sites [18].

Modifying the RBS to increase gene expression

As previously mentioned, it has been known since the early 80s that nucleotide substitutions in the RBS can affect gene expression [20]. This is obvious for the SD sequence and the start codon, since these sequences must be bound by complementary sequences in 16S RNA and the initiating tRNA, respectively. However, mutations outside these elements may also affect gene expression, primarily by modifying the RNA secondary structure around the RBS [21, 22]. If mutations in the RBS are introduced upstream the start codon in the $5'$ UTR, secondary structures in the RBS may be altered with the advantage of not having to modifying the peptide sequence. An alternative is to make synonymous codon changes in the first codons of a gene [23], which also will leave the amino acid sequence intact.

A different approach than mutagenizing the RBS to modify expression is to introduce a new RBS in the form of a 5′ fusion partner [24]. The fusion partner is usually the 5′-end (5′ UTR and early coding region) of a gene that is already known to be well expressed in the host organism or has other useful properties (such as the His-tag for protein purification [23]). When using a 5′ fusion partner the peptide coded by the early coding region of the fusion partner will be added to the N-terminus of the protein that is being expressed. This fusion peptide may later be cleaved off by specific proteases or left in place if its presence is tolerated [25].

When one wants to optimize the expression of a given gene, one may turn to commercial providers of gene optimization, such as DNA 2.0 or GenScript. However, there are published alternatives. In addition to what has been already mentioned about fusion tags and RBS mutagenesis, the most comprehensive tool yet published for optimization of translation initiation is the RBS calculator [26]. This software takes into account several sequence dependent variables that have been shown to play a role in translation initiation: the match of the 16S RNA to the SD sequence, the distance from the SD sequence to the start codon, the folding energy of any RBS secondary structures, the type of start codon, and the folding energy of the ribosome standby site [26]. Given an input mRNA sequence centered around the start codon, the tool will suggest sequence changes that will result in a calculated optimal translation initiation rate for the gene of interest.

Codon bias affects gene expression and cellular fitness

During translation elongation the ribosome matches incoming amino acids on transfer RNAs (tRNA) to codons on the mRNA. There are 64 (4^3) codons present in the genomes of nearly all organisms studied to date. Generally, 61 of these match the anti-codons on tRNA and the remaining three are stop codons. On the other hand there are only 20 amino acids carried by the 61 tRNA. The explanation for this discrepancy is that several codons are associated with the same amino acid. These codons are synonymous: they code for the same amino acid. For example, the codons AAA and AAG both code for lycine.

Even though several codons can be synonymous, they are often found with different frequencies in genomes. The preference of one synonymous codon over another is called codon usage bias. Codon usage bias is a universal phenomenon as it is found throughout the kingdom of life [27]. In many species, codon bias is especially strong in highly expressed genes, which seems to indicate that the over-represented codons in these genes are especially suited for high rates of translation. This has led to the suggestion that some codons are more “efficient” than others during translation and that genes with efficient codons are translated more rapidly [28]. It was eventually shown that the efficient codons are those that have the highest copy numbers of the corresponding tRNA genes [29, 30]. This seems to explain that efficient codons are translated more rapidly because the corresponding tRNAs are more abundant in the cell, shortening the waiting time between each tRNA binding event.

A commonly used measure for codon bias is the codon adaptation index (CAI). It measures how similar a gene’s codon content is compared to the codons in highly expressed genes in the same organism [31]. The CAI has accordingly been shown to correlate positively with gene expression levels in several species [32, 33]. The CAI has been used for codon optimization of genes for heterologous expression; for example, codon optimization has been used to achieve high expression levels when expressing human genes in bacterial hosts [34]. The underlying reason is that codon usage bias differs between species: a codon which is rapidly translated in human may be slowly translated in *E. coli*.

Two recent publications have shed light on the effect that codon bias has on the rate of translation. In the first, Kudla et al. made random synonymous codons changes in the green fluorescent protein (GFP) gene to generate a library of 154 GFP gene variants with on average 114 different codons each [35]. The GFP variants displayed a 250 fold variation in expression in *E. coli*, and caused a marked difference in cell culture growth rates. The CAI of the GFP variants did in this study not correlate with their expression levels, but instead, surprisingly, correlated with the cell division rate of the cells. The expression level, on the other hand, was found to correlated strongly with secondary structures around the start codon, but did not show any correlation with cell growth rates. The authors concluded from this that codon bias exists in highly expressed genes not to optimize translation rates, but to optimize overall cellular fitness. The hypothesis is that if highly expressed genes contained codons for which there were few tRNA, ribosomes would often pause when translating the corresponding mRNA, causing fewer ribosomes to be available to the rest of the mRNA pool, thereby slowing the cellular growth rate [35]

In the second paper, Tuller et al. [36] examined the codon bias of 27 organisms from all three domains of life using the tRNA adaptation index (tAI), a similar measure to the CAI. (the tAI measure ranks each codon with the gene copy number of the associated tRNA in the genome [36]). They found a species-wide trend where genes tend to have inefficient codons in the early coding region (first 30-70 codons), but efficient codons in mid and late coding regions. The early inefficient codons were labeled a slowly translated "ramp". Their hypothesis is that by reducing the speed during early translation, ribosomes are more evenly spaced out in the mid and late stages of translation elongation, which reduces collisions between the ribosomes and thereby increases the overall translation efficiency in the cell.

A criticism of the Tuller et al. paper is that the early codons of mRNA are also under selective pressure to reduce mRNA folding [16]. This would reduce the degrees of freedom for selection of optimal codons in the early codon region, which could partly explain the ramp effect [37].

Finally, there are other sources of codon bias than have been mentioned so far. One is the specific order that codons appear in, which has been linked to more efficient recycling of tRNA [38]. Another is the avoidance of message-bearing motifs like the Shine-Dalgarno element in the coding sequence. It was shown in *E. coli* that when Gly-Gly amino acid pairs are coded for in a gene, the most common codon pair is GGC-GGC, which out of all possible Gly codon pairs has the lowest possible affinity for the anti-Shine Dalgarno sequence; furthermore, the rarest codon pair for Gly-Gly was GGA-GGU, which is an exact match to the Shine-Dalgarno sequence [39]. In the same study it was shown that Shine-Dalgarno sequences inside coding regions could cause rebinding of the 16S RNA during translation elongation and thereby cause translation pausing. Interestingly, this rebinding behavior is the same as found for the sigma factor during transcription elongation by RNA polymerase [40], as discussed on page 13. This shows that the strong binding affinities of sigma for the promoter and 16S RNA for the Shine-Dalgarno sequence can have the secondary effect where DNA sequence in specific regions are selected on to avoid these signals.

Bibliography

[1] Sergey Proshkin et al. "Cooperation Between Translating Ribosomes and RNA Polymerase in Transcription Elongation". *Science* 328.5977 (Apr. 2010), pp. 504–508.

[2] Dipak Dutta et al. "Linking RNA Polymerase Backtracking to Genome Instability in *E. coli*". *Cell* 146.4 (Aug. 2011), pp. 533–543.

[3] J. Shine and L. Dalgarno. "The 3 -Terminal Sequence of *Escherichia coli* 16S Ribosomal RNA: Complementarity to Nonsense Triplets and Ribosome Binding Sites". *Proceedings of the National Academy of Sciences* 71.4 (Apr. 1974), pp. 1342–1346.

[4] So Nakagawa et al. "Dynamic Evolution of Translation Initiation Mechanisms in Prokaryotes". *Proceedings of the National Academy of Sciences* 107.14 (Apr. 2010), pp. 6382–6387.

[5] Patricia Skorski et al. "The Highly Efficient Translation Initiation Region from the *Escherichia coli rpsA* Gene Lacks a Shine-Dalgarno Element". *Journal of Bacteriology* 188.17 (Sept. 2006), pp. 6277–6285.

[6] Irina V. Boni et al. "Non-canonical mechanism for translational control in bacteria: synthesis of ribosomal protein S1". *The EMBO Journal* 20.15 (2001), pp. 4222–4232.

[7] Marilyn Kozak. "Regulation of translation via mRNA structure in prokaryotes and eukaryotes". *Gene* 361 (Nov. 2005), pp. 13–37.

[8] Gisela Storz, Jason A. Opdyke, and Aixia Zhang. "Controlling mRNA stability and translation with small, noncoding RNAs". *Current Opinion in Microbiology* 7.2 (Apr. 2004), pp. 140–144.

[9] Michael N. Hall et al. "A role for mRNA secondary structure in the control of translation initiation". *Nature* 295.5850 (Feb. 1982), pp. 616–618.

[10] M. H. De Smit and J. Van Duin. "Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis". *Proceedings of the National Academy of Sciences* 87.19 (1990), p. 7668.

[11] Bibiana Onoa and Ignacio Tinoco Jr. "RNA folding and unfolding". *Current Opinion in Structural Biology* 14.3 (June 2004), pp. 374–379.

[12] Maarten H. de Smit and Jan van Duin. “Translational Standby Sites: How Ribosomes May Deal with the Rapid Folding Kinetics of mRNA”. *Journal of Molecular Biology* 331.4 (Aug. 2003), pp. 737–743.

[13] Hervé Isambert. “The jerky and knotty dynamics of RNA”. *Methods* 49.2 (Oct. 2009), pp. 189–196.

[14] Tao Pan and Tobin Sosnick. “RNA FOLDING DURING TRANSCRIPTION”. *Annual Review of Biophysics and Biomolecular Structure* 35.1 (2006), pp. 161–175.

[15] Sang Woo Seo, Jina Yang, and Gyoo Yeol Jung. “Quantitative correlation between mRNA secondary structure around the region downstream of the initiation codon and translational efficiency in *Escherichia coli*”. *Biotechnology and Bioengineering* 104.3 (2009), pp. 611–616.

[16] Wanjun Gu, Tong Zhou, and Claus O. Wilke. “A Universal Trend of Reduced mRNA Stability near the Translation-Initiation Site in Prokaryotes and Eukaryotes”. *PLoS Comput Biol* 6.2 (Feb. 2010), e1000664.

[17] Thomas E. Keller et al. “Reduced mRNA Secondary-Structure Stability Near the Start Codon Indicates Functional Genes in Prokaryotes”. *Genome Biology and Evolution* 4.2 (Jan. 2012), pp. 80–88.

[18] Sean M. Studer and Simpson Joseph. “Unfolding of mRNA Secondary Structure by the Bacterial Translation Initiation Complex”. *Molecular Cell* 22.1 (Apr. 2006), pp. 105–115.

[19] Maarten H. de Smit and Jan van Duin. “Translational initiation on structured messengers: Another role for the Shine-Dalgarno interaction”. *Journal of Molecular Biology* 235.1 (Jan. 1994), pp. 173–184.

[20] N. Warburton, P. G Boseley, and A. G Porter. “Increased Expression of a Cloned Gene by Local Mutagenesis of Its Promoter and Ribosome Binding Site”. *Nucleic Acids Research* 11.17 (Sept. 1983), pp. 5837–5854.

[21] Young Seoub Park et al. “Design of 5’-untranslated region variants for tunable expression in *Escherichia coli*”. *Biochemical and Biophysical Research Communications* 356.1 (Apr. 2007), pp. 136–141.

[22] S. Care et al. “The translation of recombinant proteins in *E. coli* can be improved by *in silico* generating and screening random libraries of a 70/+96 mRNA region with respect to the translation initiation codon”. *Nucleic Acids Research* (Dec. 2007), pp. 1–6.

[23] Régis Cèbe and Martin Geiser. “Rapid and easy thermodynamic optimization of the 5 -end of mRNA dramatically increases the level of wild type protein expression in *Escherichia coli*”. *Protein Expression and Purification* 45.2 (Feb. 2006), pp. 374–380.

[24] Edward R. LaVallie and John M. McCoy. “Gene fusion expression systems in *Escherichia coli*”. *Current Opinion in Biotechnology* 6.5 (1995), pp. 501–506.

[25] Dominic Esposito and Deb K Chatterjee. “Enhancement of soluble protein expression through the use of fusion tags”. *Current Opinion in Biotechnology* 17.4 (Aug. 2006), pp. 353–358.

[26] Howard M. Salis, Ethan A. Mirsky, and Christopher A. Voigt. “Automated design of synthetic ribosome binding sites to control protein expression”. *Nature Biotechnology* 27.10 (Oct. 2009), pp. 946–950.

[27] Paul M. Sharp et al. “Codon Usage Patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a Review of the Considerable Within-Species Diversity”. *Nucleic Acids Research* 16.17 (Sept. 1988), pp. 8207–8211.

[28] Etsuko N. Moriyama and Jeffrey R. Powell. “Gene Length and Codon Usage Bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*”. *Nucleic Acids Research* 26.13 (July 1998), pp. 3188–3193.

[29] Mario Dos Reis, Renos Savva, and Lorenz Wernisch. “Solving the Riddle of Codon Usage Preferences: A Test for Translational Selection”. *Nucleic Acids Research* 32.17 (Jan. 2004), pp. 5036–5044.

[30] Johan Elf et al. “Selective Charging of tRNA Isoacceptors Explains Patterns of Codon Usage”. *Science* 300.5626 (June 2003), pp. 1718–1722.

[31] Paul M. Sharp and Wen-Hsiung Li. “The Codon Adaptation Index-a Measure of Directional Synonymous Codon Usage Bias, and Its Potential Applications”. *Nucleic Acids Research* 15.3 (Feb. 1987), pp. 1281–1295.

[32] Laurent Duret and Dominique Mouchiroud. “Expression Pattern and, Surprisingly, Gene Length Shape Codon Usage in Caenorhabditis, Drosophila, and Arabidopsis”. *Proceedings of the National Academy of Sciences* 96.8 (Apr. 1999), pp. 4482–4487.

[33] Ronald Jansen, Harmen J Bussemaker, and Mark Gerstein. “Revisiting the Codon Adaptation Index from a Whole-genome Perspective: Analyzing the Relationship Between Gene Expression and Codon Occurrence in Yeast Using a Variety of Models”. *Nucleic Acids Research* 31.8 (Apr. 2003), pp. 2242–2251.

[34] Claes Gustafsson, Sridhar Govindarajan, and Jeremy Minshull. “Codon bias and heterologous protein expression”. *Trends in Biotechnology* 22.7 (July 2004), pp. 346–353.

[35] Grzegorz Kudla et al. “Coding-Sequence Determinants of Gene Expression in *Escherichia coli*”. *Science* 324.5924 (Apr. 2009), pp. 255–258.

[36] Tamir Tuller et al. “An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation”. *Cell* 141.2 (Apr. 2010), pp. 344–354.

[37] Joshua B. Plotkin and Grzegorz Kudla. “Synonymous but not the same: the causes and consequences of codon bias”. *Nature Reviews Genetics* 12.1 (Jan. 2011), pp. 32–42.

[38] Gina Cannarozzi et al. “A Role for Codon Order in Translation Dynamics”. *Cell* 141.2 (Apr. 2010), pp. 355–367.

[39] Gene-Wei Li, Eugene Oh, and Jonathan S. Weissman. “The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria”. *Nature* 484.7395 (Mar. 2012), pp. 538–541.

[40] Rachel Anne Mooney, Seth A. Darst, and Robert Landick. “Sigma and RNA Polymerase: An On-Again, Off-Again Relationship?” *Molecular Cell* 20.3 (Nov. 2005), pp. 335–345.

2.4 Transcription termination in eukaryotes and RNA sequencing

This section contains the background material for the analysis of cleavage and polyadenylation from RNA-seq data in chapter 6. Part of this analysis is included in the paper by Djebali et al. on page 101. In this review, we will first cover the eukaryotic transcription process in general, and then focus in particular on transcription termination, with emphasis on the process of $3'$ cleavage and polyadenylation. Thirdly we review genome-wide studies that have investigated $3'$ cleavage and polyadenylation. Finally we briefly review the RNA-seq technology used to generate the data, since some steps in the RNA-seq protocol affect to what degree polyadenylation sites may be identified from this type of sequence data.

Overview of transcription in eukaryotes

In this section we will review the process of transcription termination in mammalian cells, although we will sometimes reference studies of other eukaryote cells if there are commonalities. Since the previous sections covered transcription and translation in bacteria, we will in this section often point out the contrast between transcription in eukaryotes and in bacteria.

In eukaryotes there are several types of RNA polymerases, each one responsible for transcribing different classes of genes. The polymerase which transcribes protein coding genes is called RNAP II. With 12 subunits in mammals, RNAP II is larger than its bacterial counterpart, but the core subunits are structurally and functionally conserved compared to bacterial RNAP [1].

As in bacteria, transcription in eukaryotes begins when RNAP is recruited by transcription factors to promoter sites, however in eukaryotes there is no equivalent to the σ factor system in bacteria; eukaryotes instead rely on more diversified methods of promoter recruitment [2]. Transcription initiation in eukaryotes involves the same steps of open bubble formating and abortive cycling as in bacteria, although additional assisting transcription factors are involved in this process compared to in bacterial cells [3]. Transcription elongation by RNAP also involves pausing and backtracking, but pausing on eukaryotic DNA is more complicated due to the wrapping of DNA around histones in the chromatin [4].

3′ cleavage and polyadenylation

Transcription termination is markedly different between eukaryotes and bacteria. In bacteria, the position where RNAP terminates transcription also marks the 3′ terminal position of the mRNA. In eukaryotes, however, the 3′ end of transcripts synthesized by RNAP II is created while RNAP II is still transcribing. When RNAP II transcribes past a polyadenylation signal (PAS), often in the 3′ UTR of a gene, a cleavage and polyadenylation complex may bind to that PAS and to sequences around it [5]. This complex may then cleave the still-transcribed RNA around 10 to 35 nt downstream the PAS [6]. Right after cleavage, a specialized poly(A) polymerase will synthesize around 250 adenosine residues onto the newly formed 3′ end of the pre-mRNA (Figure 2.11), forming what is known as the poly(A) tail.

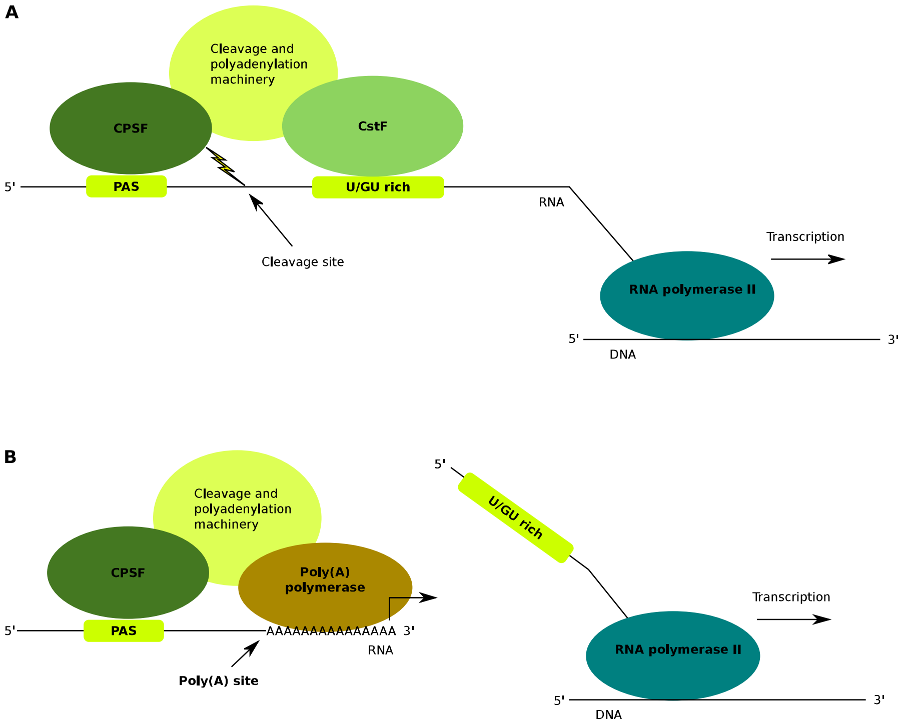


Figure 2.11 – Cleavage and polyadenylation. **A**: The cleavage and polyadenylation machinery bound to a polyadenylation signal (PAS) and U/GU rich element. **B**: After cleavage by CPSF, a poly(A) polymerase adds around 250 adenosine residues to the 3′ end of the cleaved RNA.

The PAS is a six-nucleotide sequence, most often AAUAAA or AUUAAA, but around 10 closely related variants have also been found [7]. In addition to the PAS, a U/GU-rich sequence is also sometimes found upstream polyadenylation sites. The cleavage and polyadenylation machinery that bind these sequences consist of several protein complexes. The most prominent are the cleavage and polyadenylation specificity factor (CPSF) which binds the upstream PAS and cleaves the RNA; the cleavage stimulation factor (CstF) which can bind the U/GU rich downstream sequence; and the poly(A) polymerase which performs polyadenylation [8] (Figure 2.11).

RNA $3'$ end formation of mRNA in eukaryotes is part of pre-mRNA processing, which is a necessary step between mRNA synthesis in the nucleus and mRNA translation in the cytoplasm. Other processing steps are mRNA $5'$ capping and the removal of introns from mRNA, known as splicing. If any of the processing steps are inefficiently executed or not executed at all, the pre-mRNA transcript will be targeted for degradation either in the nucleus itself or in the cytoplasm after transport [9]. This quality control step reduces the risk that errors during transcription and pre-mRNA processing will result in nonfunctional or possibly harmful protein products.

Usage of alternative polyadenylation sites

It is well known that alternative splicing can produce different alternative isoforms for a given gene: by splicing out parts of a coding sequence, a different mRNA and thereby a different protein will be produced. However, alternative isoforms can also be created through what is called alternative polyadenylation, which is the usage of alternative polyadenylation sites, all of them often in the $3'$ UTR of the same gene [8].

Depending on which polyadenylation site is used, the length of the $3'$ UTR in the final mRNA will be different. A polyadenylation site close to the beginning of the $3'$ UTR will result in a short $3'$ UTR while a polyadenylation site far from the beginning of the $3'$ UTR will result in a long $3'$ UTR in the mature mRNA. The choice of polyadenylation site can have regulatory effects, since the $3'$ UTR region often contains regulatory elements, and a short $3'$ UTR will therefore in general contain fewer regulatory sequences than a long $3'$ UTR. One of the best characterized examples of regulation in the $3'$ UTR is regulation by microRNA [10]. microRNA are short RNA of around 20 nucleotides in length that perform regulatory roles by basepairing with other RNA molecules. When microRNA binds the $3'$ UTR region of an mRNA in the cytoplasm, they generally decrease expression from the mRNA they bind to [11]. For a long time it was unclear whether microRNA binding in the $3'$ UTR decrease expression by inducing transcript degradation or simply by blocking translation. Recently, it was established that at least in mammals microRNA in the $3'$ UTR decrease expression by causing transcript degradation [12]. MicroRNA binding in the 3′ UTR are known to regulate a host of metabolic processes and human diseases [13]. Therefore, the choice of where to cleave and polyadenylate a transcript can have widespanning consequences. It is estimated that half the genes in the genome undergo alternative polyadenylation [14].

There are also examples of cleavage and polyadenylation in sites outside the 3′ UTR. The most prominent of these are sites within introns. The use of an intronic cleavage and polyadenylation site will cause any downstream exons to be left out of the transcript. In turn, this will result in a shorter peptide sequence upon translation of the mRNA. Thus, the site of cleavage and polyadenylation can also modulate the protein coding content of the mRNA. A well known example is the case of the immunoglobin protein in B cells. Depending on whether a polyadenylation site in an intron is used or not, the membrane bound or the secreted version of the immunoglobin protein is made [15].

Polyadenylation unrelated to mRNA processing

In the last decade it has become clear that there is polyadenylation of RNA in eukaryotic cells that is unrelated to mRNA 3′ processing. First, it was discovered that poly(A) tails were added to aberrant transcripts in the nucleus of yeast [16]. The protein complex responsible for this polyadenylation was given the name TRAMP, and transcripts polyadenylated in this way were found to be targeted for degradation in the nucleus [16, 17]. This was a surprising and important discovery, as previously degradation-related polyadenylation was only known from bacteria, where polyadenylation is part of RNA-degradation pathways. The discovery prompted the suggestion that degradation-associated polyadenylation by TRAMP has been conserved from bacteria in the nucleus from the origin of the eukaryotic cell [17]. Later, degradation-related polyadenylation was found in the nucleus of mammalian cells too, and eventually even in the cytoplasm of human cells [18, 19]. In summary, there is an emergent role for degradation-related polyadenylation of some RNA species in eukaryotes.

Genome-wide studies of polyadenylation

Historically, polyadenylation has been investigated using traditional molecular biology techniques, one polyadenylation site at a time. However, in the last two decades, it has become possible to perform global studies of polyadenylation due to the emergence of genome-wide assays.

The study of sites of polyadenylation across the genome has occurred in three stages in the last 15 years, with each stage resting on a different type of technology. The first wave used cDNA and EST sequence data obtained by laborious Sanger sequencing. The second wave used microarray and SAGE technologies, and the third wave used next generation RNA-sequencing (RNA-seq). We will review key results obtained in these three stages chronologically.

From the early 90s onward, more and more human expressed sequence tags (ESTs) became available. (An EST is a part of a cloned RNA which has been isolated and sequenced, typically by Sanger sequencing.) The increasing amount of sequence data facilitated for the first time large scale analysis of $3'$ UTRs and polyadenylation sites. The polyadenylation site of an EST is found by identifying a poly(A) or poly(T) sequence in the extremity of the EST which does not correspond to any genomic sequence. By trimming that extremity, and matching what remains of the EST to databases of sequenced RNA or to a genome sequence, it is possible to identify the genomic location of the polyadenylation site [7, 14]. These early genome-wide studies were successful in determining i) the frequency of occurrence of the different PAS variants [7]; ii) that over half of human genes employ alternative polyadenylation [14]; iii) that sites of alternative polyadenylation are evolutionary conserved between humans and mice [14]; and vi) that polyadenylation in intronic regions is common [20].

However, EST data limits the type of questions that can be investigated. First, EST data was in low quantity, due to the expense and time needed for Sanger sequencing. Thus, the only practical way to compare alternative polyadenylation on a genome-wide scale was to include EST data from different experiments, often resulting in a mix of data from different cell lines and tissues. Our literature review revealed no studies with *de novo* EST sequencing for the purpose of studying polyadenylation sites; all studies used EST sequences from databases. Another limitation of EST data is that it is biased toward protein coding genes that were found interesting enough to sequence individually. Thus many classes of polyadenylated RNA, such as long noncoding RNA, were possibly missed by these studies. Finally, although EST data may be used to give a quantitative profile of gene expression, the output data is often normalized so that the quantitative profile is lost [21]. It is therefore difficult to compare expression values across genes with EST data, although some approaches have been developed for this purpose [21]. A quantitative profile of polyadenylation site usage is desirable when studying alternative polyadenylation as one can identify the usage frequency of the different polyadenylation sites for a given mRNA.

As the microarray technology matured in the early 2000s and more full-length genomes became available, microarrays, often in combination with EST data, were used to study $3'$ UTR length variation by alternative polyadenylation. To use microarrays to investigate alternative polyadenylation, the probes in the microarray were designed to bind to sequences present the different $3'$ UTRs formed by alternative polyadenylation. By comparing the intensities of the probes under different conditions, one could compare the usage of different polyadenylation sites [22, 23]. Microarrays could thereby be used to directly compare $3'$ UTR length and expression levels in time-series under different experimental condition with different cell lines and tissues.

Key results obtained with a combination of microarray and EST data include the identification of tissue-specific patterns of polyadenylation in humans [24], and wide-spread shortening of $3'$ UTR length during immune cell activation [22]. A combination of EST, microarray, and SAGE (Sanger-based RNA tag sequencing) showed a progressive lengthening of mouse $3'$ UTRs during embryonic development [23]. Thus, the arrival microarrays enabled the observation of the dynamic nature of $3'$ cleavage and polyadenylation.

The microarray studies provided a wealth of insight, but were still limited in one sense when studying alternative polyadenylation: it is difficult to use microarray experiments to identify novel polyadenylation sites. Further, in terms of quantification, microarray output is typically only used to provide relative differences of gene expression for each gene between experiments. This makes it difficult to compare polyadenylation site usage for different genes in the same experiment.

With the advance of second generation sequencing technology in the form of RNA-seq in the late 2000s, many of the limitations of both EST and microarray data seemed to be resolved. RNA-seq was introduced in 2008 [25] and has been used to compare the differential expression of genes, to discover new genes, and to discover novel isoforms of known genes by finding new splice-sites and new $5'$ and $3'$ terminals [26]. RNA-seq combines the best of EST and microarray data when studying polyadenylation sites. Firstly, like ESTs, the RNA-seq data is in sequence format, allowing the direct detection of poly(A) tails and thereby the site of cleavage and polyadenylation. Secondly, like microarray data, RNA-seq is quantitative, allowing the direct comparison of expression levels of $3'$ UTRs across the genome. And thirdly, like microarrays, RNA-seq can easily be performed on RNA samples in time-series from different cell types and tissues, allowing direct hypothesis testing which cannot be done when using EST information obtained from databases.

RNA-seq was rapidly used to study the polyadenylation landscape for cell lines and tissues. These experiments first of all confirmed what had been discovered earlier by single-mRNA studies and EST analysis; that AAUAAA is the canonical polyadenylation signal, that single genes can be represented with multiple sites of polyadenylation, and that there is frequent polyadenylation of intronic sequences. New discoveries included many novel polyadenylation sites scattered across the genome [27, 28]. It was also found that intronic and intergenic polyadenylation sites are in humans associated with a novel TTTTTTTTT motif which does not occur at the normal polyadenylation sites in 3′ UTRs [27]. Further, RNA-seq was used for genome-wide annotation of polyadenylation sites for the first time in *C. elegans* and *A. thaliana* [29, 30], species for which too little EST data had been available for genome-wide polyadenylation site annotation.

RNA-seq was also used to follow up and test the conclusions that had so far been made about alternative polyadenylation in earlier publications. One groundbreaking study had previously found a shortening of the 3′ UTR of many transcripts in a cancer cell line, and it was proposed that this was a general characteristic of cancer cells [31]. As a follow up to this study, Fu et al. compared the relative change in 3′ UTR length between two cancer cell lines and a non-cancer control cell line [32]. They did not find a consistent pattern of shortening of 3′ UTRs in the cancer cell lines. Instead, the 3′ UTR lengths of one of the cancer cells was shorter and 3′ UTR lengths of the other was longer than the control. This suggests that there is no clear-cut genome-wide trend of short 3′ UTRs in cancer cells, contrary to what had previously been concluded.

One unexpected result from using RNA-seq to study polyadenylation sites was the finding of poly(A) tails for histone mRNA in both human, mice, and *C. elegans* [29, 33]. Histone mRNA were previously thought to be the only mRNA in metazoans without a poly(A) tail [34], even though several of the histone genes had been found to contain the AATAAA polyadenylation signal at their 3′ ends [35]. A possible explanation for why histone mRNA have prior to RNA-seq not been found with poly(A) tails is that the histone transcripts are first cleaved and polyadenylated, and subsequently processed to lose their poly(A) tail [29] (supplementary materials). This example shows that some discoveries can arrive unexpectedly when taking a neutral, genome-wide look at established findings.

One recent and thorough study of genome-wide polyadenylation was performed using a novel sample-preparation protocol by Derti et al. [28] They used RNA-seq to find polyadenylation sites in five mammal species, including human, in 24 tissues [28]. They found over 400.000 polyadenylation sites in human tissues, compared to 150.000 found previously. One reason why they have found so many sites compared to previous studies could be the increased resolution in this study: most novel polyadenylation sites were found in lowly expressed transcripts, which may not previously have been detected. Derti et al. also found that although many polyadenylation sites were tissue specific, 70 % of genes showed the same usage of alternative polyadenylation across all tissues [28].

Transcriptome sequencing with RNA-seq

Here we will go through the stages of the RNA-seq experiments that were performed to produce the data that was analysed as presented in chapter 6. We will then comment on the stages of the experiment where biases can be introduced that affect the final result. Finally we will discuss the matter of mapping the output of RNA-seq – short RNA sequences called reads – to the genome.

RNA-seq and mapping: the method, errors, and biases

Here follows a list of steps performed to generate the RNA-seq data that has been used in this thesis. In the second and forth step, it is necessary to use poly(T) primes to capture RNA with poly(A) tails. This ensures that poly(A) tails will be present in the final output.

1. The first step for the RNA-seq experiment is to isolate RNA from a cell sample obtained from a tissue or a cell culture

2. Second, the isolated RNA is divided into poly(A)+ and poly(A)- fractions. This is done by using a poly(T) primer that binds to the poly(A) tail of the mRNA. The RNA that binds the poly(T) primer is called the poly(A)+ fraction and the RNA that does not bind the primer is separated and called the poly(A)- fraction.

3. Next, the RNA samples are treated to remove ribosomal RNA. Since ribosomal RNA is by far the most abundant species of RNA in the cell, its removal will increase the sensitivity for detecting lowly expressed transcripts in the sample.

4. Next, the single stranded RNA is converted into double stranded complementary DNA (cDNA). This is required because Illumina sequencers sequence DNA and not RNA. To turn RNA into cDNA one

anneals primer sequences to the RNA template, which are used by the reverse transcriptase enzyme for synthesizing cDNA. At this step, it is important when studying polyadenylation that at least some of the primers are poly(T) primers, otherwise the poly(A) tail itself may not be converted into cDNA. Since the poly(T) primers can bind anywhere in the poly(A) tail, the lengths of the poly(A) tails in cDNA will be reduced compared to those in the original RNA sample. In this step, the original single stranded RNA is degraded so only the double stranded cDNA remains.

5. The next step is to fragment the cDNA into smaller pieces, usually by sound energy (sonication), to reach a desired average fragment size compatible with the sequencing machine (usually around 300 nt),
6. Finally, short oligonucleotide adapters are added to the 3′ and 5′ ends of the cDNA and the cDNA is amplified from these adapters with PCR. PCR amplification is necessary to get the volume of DNA required by the sequencing machines. Now the cDNA is ready for sequencing.

During sequencing, the double stranded cDNA is split into single strands, and both strands are sequenced. This has the effect that sequencing outputs the reverse-transcribed version of the original RNA molecule in addition to the original. For investigation polyadenylation sites, it is important to note that this means that usually only the reverse-transcribed leading poly(T) sequence, and not the poly(A) sequence of the original, will be sequenced (Figure 2.12). For example, a polyadenylated 5′ CCCGAAAA 3′ input will most often be output as 5′ TTTTCGGG 3′ from the sequencing machine. The poly(A) sequence is rarely sequenced because the read length is generally shorter than the fragment length (76 bp vs 300 bp in Figure 2.12), and sequencing happens in the 5′ to 3′ direction.

It is also important to note that when sequencing homopolymers (single-nucleotide repeat sequences) like poly(A) or poly(T) stretches, there is a slight increase in sequencing error rates with the Illumina platform [36]. This implies that when sequencing a poly(A) sequence, other nucleotides than just A may be reported with a higher probability compared to the rest of the output sequence.

Mapping reads to the genome

The sequence-snippets that are output from sequencing machines are called reads, and generally come in sizes from 30 to 500 basepairs, depending on the

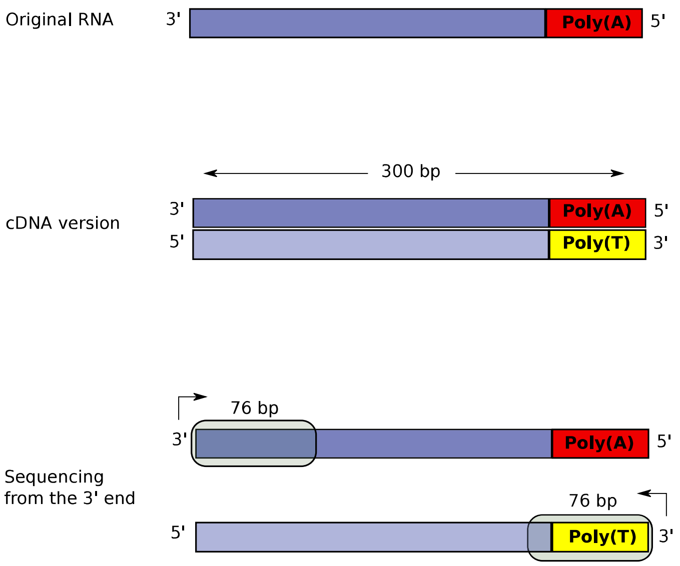


Figure 2.12 – When sequencing polyadenylation sites it is normally the reverse-transcribed version of the transcript that is actually sequenced

technology used. If a reference genome exists for the organism from which the RNA sample was taken, these reads can be mapped to that genome by various pattern-matching algorithms to find out where the RNA originated from [37]. All methods used for mapping allow for mismatches between the read and the genome to allow for errors introduced through cDNA library creation and sequencing itself.

Bibliography

[1] Richard H. Ebright. "RNA Polymerase: Structural Similarities Between Bacterial RNA Polymerase and Eukaryotic RNA Polymerase II". *Journal of Molecular Biology* 304.5 (Dec. 2000), pp. 687–698.

[2] Kevin Struhl. "Fundamentally Different Logic of Gene Regulation in Eukaryotes and Prokaryotes". *Cell* 98.1 (July 1999), pp. 1–4.

[3] J Wade and K Struhl. "The transition from transcriptional initiation to elongation". *Current Opinion in Genetics & Development* 18.2 (Apr. 2008), pp. 130–136.

[4] Robert J. Sims, Rimma Belotserkovskaya, and Danny Reinberg. "Elongation by RNA Polymerase II: The Short and Long of It". *Genes & Development* 18.20 (Oct. 2004), pp. 2437–2468.

[5] D. F. Colgan and J. L. Manley. "Mechanism and regulation of mRNA polyadenylation". *Genes & Development* 11.21 (Nov. 1997), pp. 2755–2766.

[6] Nick J. Proudfoot. "Ending the Message: poly(A) Signals Then and Now". *Genes & Development* 25.17 (Sept. 2011), pp. 1770–1782.

[7] E. Beaudoing. "Patterns of Variant Polyadenylation Signal Usage in Human Genes". *Genome Research* 10.7 (July 2000), pp. 1001–1010.

[8] Carol S. Lutz. "Alternative Polyadenylation: A Twist on mRNA 3 End Formation". *ACS Chemical Biology* 3.10 (2008), pp. 609–617.

[9] Meenakshi K. Doma and Roy Parker. "RNA Quality Control in Eukaryotes". *Cell* 131.4 (Nov. 2007), pp. 660–668.

[10] Dafne Campigli Di Giammartino, Kensei Nishida, and James L. Manley. "Mechanisms and Consequences of Alternative Polyadenylation". *Molecular Cell* 43.6 (Sept. 2011), pp. 853–866.

[11] David P. Bartel. "MicroRNAs: Genomics, Biogenesis, Mechanism, and Function". *Cell* 116.2 (Jan. 2004), pp. 281–297.

[12] Eric Huntzinger and Elisa Izaurralde. "Gene silencing by microRNAs: contributions of translational repression and mRNA decay". *Nature Reviews Genetics* 12.2 (Jan. 2011), pp. 99–110.

[13] Yong Huang et al. "Biological functions of microRNAs: a review". *Journal of Physiology and Biochemistry* 67.1 (Oct. 2010), pp. 129–139.

[14] B. Tian. “A large-scale analysis of mRNA polyadenylation of human and mouse genes”. *Nucleic Acids Research* 33.1 (Jan. 2005), pp. 201–212.

[15] M. L. Peterson and R. P. Perry. “The regulated production of mu m and mu s mRNA is dependent on the relative efficiencies of mu s poly(A) site usage and the c mu 4-to-M1 splice.” *Molecular and Cellular Biology* 9.2 (Feb. 1989), pp. 726–738.

[16] Françoise Wyers et al. “Cryptic Pol II Transcripts Are Degraded by a Nuclear Quality Control Pathway Involving a New Poly(A) Polymerase”. *Cell* 121.5 (June 2005), pp. 725–737.

[17] John LaCava et al. “RNA Degradation by the Exosome Is Promoted by a Nuclear Polyadenylation Complex”. *Cell* 121.5 (June 2005), pp. 713–724.

[18] Shimyn Slomovic et al. “Polyadenylation of ribosomal RNA in human cells”. *Nucleic Acids Research* 34.10 (2006), pp. 2966–2975.

[19] Shimyn Slomovic et al. “Addition of poly(A) and poly(A)-rich tails during RNA degradation in the cytoplasm of human cells”. *Proceedings of the National Academy of Sciences* 107.16 (Apr. 2010), pp. 7407–7412.

[20] Bin Tian, Zhenhua Pan, and Ju Youn Lee. “Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing”. *Genome Research* 17.2 (Feb. 2007), pp. 156–165.

[21] Donglin Liu and Joel H. Graber. “Quantitative comparison of EST libraries requires compensation for systematic biases in cDNA generation”. *BMC Bioinformatics* 7.1 (Feb. 2006), p. 77.

[22] Rickard Sandberg et al. “Proliferating Cells Express mRNAs with Shortened 3’ Untranslated Regions and Fewer MicroRNA Target Sites”. *Science* 320.5883 (June 2008), pp. 1643–1647.

[23] Zhe Ji et al. “Progressive lengthening of 3 untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development”. *Proceedings of the National Academy of Sciences* 106.17 (Apr. 2009), pp. 7028–7033.

[24] Haibo Zhang, Ju Youn Lee, and Bin Tian. “Biased alternative polyadenylation in human tissues”. *Genome Biology* 6.12 (2005), R100.

[25] Ugrappa Nagalakshmi et al. “The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing”. *Science* 320.5881 (June 2008), pp. 1344–1349.

[26] Zhong Wang, Mark Gerstein, and Michael Snyder. “RNA-Seq: a revolutionary tool for transcriptomics”. *Nature Reviews Genetics* 10.1 (Jan. 2009), pp. 57–63.

[27] Fatih Ozsolak et al. “Comprehensive Polyadenylation Site Maps in Yeast and Human Reveal Pervasive Alternative Polyadenylation”. *Cell* 143.6 (Dec. 2010), pp. 1018–1029.

[28] Adnan Derti et al. “A Quantitative Atlas of Polyadenylation in Five Mammals”. *Genome Research* (Mar. 2012).

[29] Marco Mangone et al. “The Landscape of C. elegans 3 UTRs”. *Science* 329.5990 (July 2010), pp. 432–435.

[30] Xiaohui Wu et al. “Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation”. *Proceedings of the National Academy of Sciences* 108.30 (July 2011), pp. 12533–12538.

[31] Christine Mayr and David P. Bartel. “Widespread Shortening of 3 UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells”. *Cell* 138.4 (Aug. 2009), pp. 673–684.

[32] Yonggui Fu et al. “Differential genome-wide profiling of tandem 3 UTRs among human breast cancer and normal cells by high-throughput sequencing”. *Genome Research* 21.5 (May 2011), pp. 741–747.

[33] Peter J. Shepard et al. “Complex and Dynamic Landscape of RNA Polyadenylation Revealed by PAS-Seq”. *RNA* 17.4 (Apr. 2011), pp. 761–772.

[34] William F. Marzluff, Eric J. Wagner, and Robert J. Duronio. “Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail”. *Nature Reviews Genetics* 9.11 (Nov. 2008), pp. 843–854.

[35] Rebecca Keall et al. “Histone gene expression and histone mRNA 3’ end structure in Caenorhabditis elegans”. *BMC Molecular Biology* 8 (June 2007), p. 51.

[36] André E. Minoche, Juliane C. Dohm, and Heinz Himmelbauer. “Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems”. *Genome Biology* 12.11 (2011), R112.

[37] Manuel Garber et al. “Computational methods for transcriptome annotation and quantification using RNA-seq”. *Nature Methods* 8.6 (2011), pp. 469–477.

Chapter 3

RNA secondary structures as barriers to protein synthesis

3.1 Summary of the paper

The primary aim of the publication by Kucharova et al., attached in the Appendix on page 111, was to express in *E. coli* a variant of the human interferon gene called *inf-α2b*. The protein product INF-α2b is of pharmaceutical interest as a drug for treating hepatitis C [1]. The expression of a foreign gene in a host organism is called heterologous expression, and carries with it many challenges [2]. Since heterologous expression of INF-α2b is necessary to produce the drug in large quantities, and *E. coli* is a much used expression host, it is important to investigate the limitations and mechanisms of *inf-α2b* expression in *E. coli*.

Kucharova et al. reports that the *inf-α2b* gene is not expressed in *E. coli* in its native form. A codon optimized version of *inf-α2b* was obtained to investigate if the non-native codon usage in the gene was behind the lack of expression. That is, the native human codons were replaced with codons that are in high usage in *E. coli*. However, in spite of of codon optimization no detectable expression of transcript or protein could be found. Protein product was only detected when a $5'$ fusion tag was added to the *inf-α2b* gene. This indicated that events occurring in the $5'$ region of the gene is involved in regulating the switch between expression and nonexpresssion of the human *inf-α2b* gene in *E. coli*.

In their paper, Kucharova et al. suggest that translation initiation of the *inf-α2b* transcript is a possible reason why the gene is not expressed. Further, they note that codon usage in first translated codons are also known to impact gene expression. That prompts them to investigate variants of *inf-*

α2b that make synonymous codon substitutions to reduce RNA secondary structures around the ribosome binding site (RBS). Some of these variants result in detectable transcript levels for *inf-α2b*. However, the increase in transcript level is not followed by an increase in protein level. This suggests that further barriers than ribosome binding lie behind the poor expression of *inf-α2b*. Having established this, Kucharova et al. move on to investigate the effect of different 5′ terminal fusion peptides. They show that that several short versions of the celB fusion peptide increase expression of the *inf-α2b* gene and several other genes. Finally, they improve the expression level with the celB leader by screening a random mutagenesis library of celB around the RBS. In conclusion, these fusion peptides probably facilitate translation initiation both by having favorable secondary structure and also by some downstream effect that could not be achieved only by alleviating RNA secondary structures at the RBS.

3.2 Bioinformatic contribution by thesis author

As described in the paper by Kucharova et al. (see the Materials and Methods section of the paper) the bioinformatic contribution to the paper was primarily the construction an *in silico* library of *inf-α2b* variants, and the screening of this library with respect to: i) RNA free energy around the RBS and ii) a score for the codon usage of the first 8 codon after the start codon. See Figure 3.1 (Figure 1 in the paper) for a graphical representation of the library in terms of the free energy and codon score of each variant, where the variants that were experimentally tested are indicated.

3.2.1 Bioinformatic implementation

Here, we will outline, in pseudo code, the implementation of the source code that was used to produce and process the *in silico inf-α2b* library. The actual implementation is available as outlined in the Appendix on page 87.

```
import Bio # import the BioPython library, which holds information about
#E coli codons

# Load the sequence of the infa2b gene
infa2b_sequence = obtain_infa2b_sequence('infa2b.txt')

# Obain the length of the sequence
inf_len = len(infa2b_sequence)

# Make a list of all the codons in infa2b
```

```
variable_codons = [infa2b_sequence[i:i+3] for i in range(0, inf_len, 3)]

# Load a list of all codons and condon usage frequencies in E coli
# Use the Python package Bio for this, which contains codon tables,
# together with the E coli genome sequence.
coli_codon_usage = Bio.obtain_codon_usage(species='E coli',
                                          genome='ecoli.fa')

# Set the number of variable codons after ATG
codon_freedom = 8

# Genrate all possible variants of infa2b with synonymous mutations
# The variants are represented as instances of a sequence class.
seqObjects = make_variants(infa2b_sequence, variable_codons,
                           coli_codon_usage, codon_freedom)

# Calculate folding energy for the variants from -32 to +30 around the
# start codon, which begins on nucleotide +32. Use the program
# hybrid-ss-min to calculate the energies. The temperature for the
# experiments was 30 degrees centigrade, so calculate RNA folds also for
# this temperature.
calculate_folding_energy(seqObjects, program='hybrid-ss-min',
                         area=(-32, 30), start_codon=32,
                         temperature=30)

# Calculate the codon usage index of the first 8 codons after start codon
calculate_codon_usage(seqObjectsm, codon_freedom, start_codon=32)

# Write lists of the synonymous variants with the least secondary
# structures and the lowest codon usage index (some of these were
# selected for wet-lab testing)
write_candidates(seqObjects, 'sequence_output_directory')

# Plot the in silico library in terms of folding energy and codon score
plot_candidates(seqObjects, 'figure_output_directory')
```

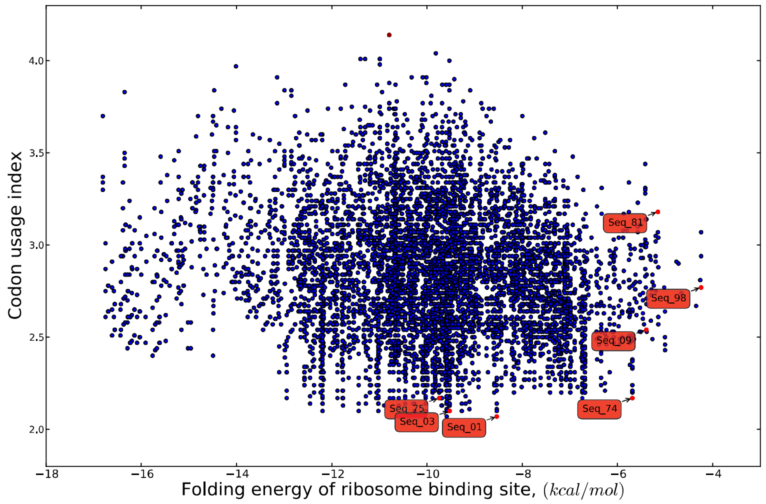


Figure 3.1 – *in silico* library with 7680 synonymous variants of *inf-α2b*. The dark red circle at the top is the wild type codon optimized *inf-α2b*. In red are shown the variants that were selected for experimental testing.

Bibliography

[1] Michael P Manns et al. “Peginterferon alfa-2b plus ribavirin compared with interferon alfa-2b plus ribavirin for initial treatment of chronic hepatitis C: a randomised trial”. *The Lancet* 358.9286 (Sept. 2001), pp. 958–965.

[2] Claes Gustafsson, Sridhar Govindarajan, and Jeremy Minshull. “Codon bias and heterologous protein expression”. *Trends in Biotechnology* 22.7 (July 2004), pp. 346–353.

Chapter 4

Thermodynamic model of initial transcription

4.1 Summary of the paper

The manuscript ”Sequence-Dependent Promoter Escape Efficiency Is Strongly Influenced by Bias for the Pre-Translocated State During Initial Transcription” can be found attached in the appendix on page 91. The manuscript details how the promoter escape efficiency of E. coli RNA polymerase varies with the sequence composition of the first 20 transcribed bases, which is referred to the initial transcribed sequence, or ITS. The authors have used an improved thermodynamic model to show that variability in promoter escape efficiency can be largely attributed to how translocation equilibria vary with respect to the ITS during initial transcription. It was found that promoter ITS variants associated with low promoter escape efficiencies tend to bias RNAP for the pre-translocated state during the process of scrunching. This finding led to inference of additional mechanistic insight into the process of scrunching itself. The results indicate that translocation by scrunching during initial transcription and processive translocation during elongation are governed by the same regulatory parameter, namely the 3' dinucleotide sequence of the nascent RNA product. Evidence regarding this insight would be difficult to obtain experimentally, due to the unstable nature of the initial transcribing complexes. The thermodynamic modeling approach used by the authors compensates for this experimental intractability. As proof of the robustness of the model, new promoter-ITS constructs were designed which gave experimental confirmation of the predicted promoter escape efficiencies.

This is the first paper that uses a thermodynamic model of translocation to show a clear correlation between template DNA/nascent RNA sequence and promoter escape efficiency during initial transcription. The model is distinguishable from previous ones in that it includes the effect of the nascent RNA 3' dinucleotide on translocation as a key component for model accuracy. It is also distinguished in that it is used predictively and not just descriptively.

Chapter 5

Kinetic model of initial transcription

5.1 Summary of the paper

The manuscript ”The Speed of Transcription Is the Same for Promoter Bound and Elongating RNA Polymerase”, submitted to BMC Bioinformatics, can be found attached in the appendix on page 123.

The results reported in this manuscript concern the kinetics initial transcription. We approach parameter estimation for this process with a new method that combines bulk and single molecule transcription data. With this method we identify rate constants of the initial transcription process that have so far not been measured. We find that the speed of transcription is highly similar to the speed that has been found previously for transcription elongation. This could not be expected beforehand, since RNAP transcribes DNA in quite different conformations depending on if it is promoter-bound or promoter-free. Our findings allow us to conclude that neither the promoter-bond through the sigma factor nor the accommodation of a scrunched DNA bubble (the special characteristics of promoter-bound transcription) have a big affect on the reactions involved in translocation during initial transcription. This information is of interest to those studying the exact method by which DNA becomes scrunched with RNAP and the mechanism of promoter escape.

Chapter 6

RNA cleavage and polyadenylation in human cell lines

This chapter is about the investigation of polyadenylation sites using RNA-seq data. We developed a pipeline, called Utail, for finding polyadenylation sites from RNAs-seq, and ran it on RNA-seq data to characterize polyadenylation sites in 12 human cell lines. Parts of the analysis was published as part of the ENCODE project in the paper “Transcription landscape of human cells”, which is attached on page 101 in the Appendix, by Djabeli et al. We begin this chapter by briefly explaining what the ENCODE project is about. Then we summarize the findings in the Djabeli et al. paper and the contribution from this thesis work to the paper. Finally we cover the findings from the investigation that were not included in the paper.

6.1 The ENCODE pilot project

The complete human genome was published in 2003 (augmenting the working draft that was published in 2001). To find out what could be learned from the newly available genome, the ENCODE (**Enc**yclopedia **O**f **D**NA **E**lements) pilot project was launched to investigate in detail 1% of the human genome through a collaboration of research labs all over the world. With the then prevalent genome-wide technologies, which included among others microarrays, CAGE, and ChIP-Chip, combined with bioinformatic analysis, the outcome of the ENCODE pilot was a wealth of data that confirmed previous tentative genome-wide findings, added new knowledge, and was used to improve the annotation the human genome. The outcome was published in a summary paper in Nature [1] as well as in 28 companion papers in a special edition of Genome Research. The data generated by ENCODE is freely available from http://genome.ucsc.edu/ENCODE/ including important metadata about how the experiments were performed.

6.2 The ENCODE follow-up project

The ENCODE pilot was deemed a success and led to a follow-up project that would cover 100% of the genome. This was made possible by the then-emerging next-generation DNA sequencing technology which had considerably lowered the cost of sequencing while also increasing the throughput. The technologies from the ENCODE pilot were mostly replaced with the next-generation version: microarrays were generally substituted for RNA-seq, CAGE could use next-gen sequencers instead of Sanger sequencing, and ChIP-Chip was replaced by ChIP-seq.

6.3 Summary of the paper

The paper by Djabeli et al. focuses on using RNA-seq data from ENCODE in order to characterize broadly the transcriptome of human cell lines. They found that the cumulative RNA coverage for 15 human cell lines is 62% for processed transcripts and 75% for primary transcripts (e.g. pre-RNA). In other words, three quarters of the human genome is transcribed. However, the average transcript coverage for each cell line was 22% and 39% for processed and primary transcripts. This tells us that while most of the genome is capable of being transcribed, each cell line will express little more than a quarter of it. Using the Cufflinks software, which assembles a transcriptome from RNA-seq data [2], they discovered 45% more transcripts than are currently annotated, with most of the newly identified transcripts coming from intergenic regions. As a result, the median length of intergenic regions decreased from around 14.000 bp to 4.000 bp, showing that the human genome is not as ”barren” as was once thought. The study identified around 22.000 new RNA splice sites, demonstrating the flexibility of the human genome whereby each gene may express multiple RNA isoforms. Many genes were found to express up to 12 alternative isoforms, although more than 30 % of gene expression can be attributed to one dominant isoform. It was found that three quarters of protein coding genes have at least two dominant isoforms. This shows that while alternative isoform usage is ubiquitous, in general a few dominating isoforms exist for each gene. Additionally, they discovered many new long noncoding RNAs (lncRNA). These RNAs appear similar to mRNA and undergo processing, but do not code for protein. An over-all conclusion about expression levels is that protein coding transcripts are highly expressed, while non-coding transcripts like lncRNA are generally lowly expressed, down to less than one transcript estimated per cell. Using PET-data, they discovered roughly 128.000 transcription start sites, of which around 30.000 were novel. They also discovered around 129.000 cleavage and polyadenylation sites inside annotated transcripts, 80% of which were not previously annotated. The last finding suggest that transcription start sites are generally better annotated than transcript end sites.

6.4 Contribution

Below are presented the results from the investigation of polyadenylation in human cell lines using ENCODE RNA-seq data. The parts of the results from this investigation that were included in the paper by Djabeli et al. was that are that around 129.000 polyadenylation sites were identified in annotated regions from the ENCODE RNA-seq data, and that 80% of these sites, although landing in annotated regions, were not previously annotated as polyadenylation sites. The inclusion of the analysis of polyadenylation sites make up an important contrast to the investigation of transcription start sites which also features in the paper.

6.5 Results

6.5.1 Total number of polyadenylation sites

By merging all polyadenylation sites from all datasets (see Table 6.1), we identified a total of 163.537 polyadenylation sites in the genome for the poly(A)+ fraction, around 129.000 of which were in annotated regions. 94040 or 58% of these sites were found with a downstream PAS (within 40 nt) and 33054 or 20% were previously annotated in GENCODE [3] or in polyAdb [4]. In addition, we found 57.031 polyadenylation sites for the poly(A)- fraction, many of which overlapped the polyadenylation sites in the poly(A)+ fraction. Figure 6.2 shows how the number of polyadenylation sites saturates as the number of included datasets increases. The saturation is sharper both for polyadenylation sites which we found to have a downstream PAS and sites supported by PET data. The saturation of polyadenylation sites is sharper for the poly(A)- fraction (Figure 6.2B), and fewer of the poly(A)- sites are associated with PAS or PET.

6.5.2 Distribution of polyadenylation sites across the genome

As expected, we found that most of the polyadenylation sites were in the $3'$ UTR exonic regions, but we located many polyadenylation sites in the intergenic and intronic regions as well (Figure 6.4). In the poly(A)+ fraction, there were more intronic polyadenylation sites in the nucleus than in the cytoplasm. This was to a certain degree expected since there is much more intronic RNA in the nucleus compared to the cytoplasm. However, introns do naturally exist in the cytoplasm. Some can be included as part of the final mRNA (intron inclusion), and some can be included if an intron contains an active polyadenylation site [5]. The number of polyadenylation sites in the other genomic regions, such as the $5'$ UTR and $3'$ UTR intronic regions, was low and is not shown in Figure 6.4.

The evidence for polyadenylated RNA in the poly(A)- fraction (Figure 6.4B) was unexpected, since the poly(A)- fraction is by design not supposed to contain RNA with poly(A) tails. We reasoned that the polyadenylated RNA in the poly(A)- fraction could have three possible sources (shown in Figure 6.1).

The first source (S1) is purely technical: normal poly(A)+ mRNA with full length poly(A) tails that were not adsorbed during the poly(A)+ filtration step and therefore ended up in the poly(A)- fraction.

Source two (S2) and source three (S3) are on the other hand of biological origin. S2 is RNA with poly(A) tails that are shorter than 20nt, which is the threshold length of poly(A) tails captured by the poly(A)+ filtration step. One type of S2 RNA are mRNA which have had their poly(A) tails degraded to below 20 nucleotides at the time of sampling: poly(A) tail degradation is part of the mRNA degradation pathway, and some mRNA are bound to be undergoing degradation at the time of sampling. Another type of S2 RNA are mRNA that were actively undergoing polyadenylation at the time of sampling and did not reach a poly(A) tail length of more than 20 nucleotides.

Source three (S3) on the other hand is distinct from S1 and S2. We propose that S3 consists of RNA with the short, degradation-related transient poly(A) tails that have been identified in the nucleus of mammalian cells [6]. These RNA may be aberrant rRNA [7], aberrant pre-mRNA [8], or possibly intronic RNA undergoing degradation [9].

6.5.3 Isolating polyadenylation sites unique to the poly(A)- and poly(A)+ fractions

We wanted to separate the S3 from the S1 and S2 polyadenylation sites in the dataset. We did this by assuming that many of the S1 and S2 sites are likely to be present in both the poly(A)- and the poly(A)+ fractions. Therefore, we filtered out all the polyadenylation sites that were common to the poly(A)+ and poly(A)- fraction and removed these from both fractions. After removing the polyadenylation sites common to poly(A)- and poly(A)+, a difference emerged in the now “pure” fractions. As can be seen in in Figure 6.3, almost no polyadenylation sites were found in the pure poly(A)- fraction of the cytoplasm. The pure nuclear poly(A)- fraction has fewer polyadenylation sites in the 3′ UTR exonic region compared to the original poly(A)- fraction, while the number of sites in the intergenic and intronic regions did not change much.

6.6 Discussion

Polyadenylation is a post-transcriptional modification of RNA with two opposite regulatory functions. The poly(A) tail confers stability when it is is added as part of mRNA processing. Conversely, the poly(A) tail signals for degradation when used in bacteria and as recently discovered for some RNA species in eukaryotes [7].

Here we have investigated sites of polyadenylation in the transcriptome in an RNA-seq library that contains data from RNA both in the whole cell and in the nuclear and cytoplasmic compartments. Further, the RNA in the library was sequenced separately for the poly(A)+ and poly(A)- fractions.

The discovery of polyadenylation sites saturated slowly for all sites as the number of datasets increased, but less slowly for sites associated with a PAS or supported by PET (Figure 6.2). We consider polyadenylation sites with a downstream PAS to be more likely to be true positives. Thus, the different saturation curve for non-PAS and PAS-associated polyadenylation sites indicates that the number of false positives increases when a large number of datasets are included.

We found many polyadenylation sites in intergenic regions (Figure 6.4), which are regions of the genome that do not contain annotated genes. These sites may represent either unannotated 3′ UTR ends from known genes or cleavage and polyadenylation sites of novel transcripts.

In general, we see that the pattern of polyadenylation in the whole cell extracts looks like the sum of the polyadenylation sites in the nuclear and cytoplasmic extracts (Figure 6.4). This is as expected; the nucleus and the cytoplasm make up the whole cell. However, it shows the power of studying the individual compartments instead of just the whole cell. To the best of our knowledge, all studies of polyadenylation to date have used RNA samples from whole cell extracts. Information gathered in these studies would therefore correspond to the view in the top row in Figure 6.4. That the intronic polyadenylation sites come mainly from the nucleus would be difficult to infer without compartmentalized RNA-seq data.

Before removing the polyadenylation sites common to the poly(A)+ and poly(A)- RNAs, there were few polyadenylation sites in the cytoplasmic poly(A)- RNA, and most of then in $3'$ UTR exonic regions (Figure 6.4D). However, after removing common sites, it became clear that the few polyadenylation sites in the cytoplasmic poly(A)- RNA were in common with the poly(A)+ RNA (Figure 6.3E and 6.3F). The pure cytoplasmic poly(A)- RNA is therefore practically void of polyadenylation sites. This indicates that there are very few RNA with short poly(A) tails in the cytoplasm. We propose that the signals common to the poly(A)+ and poly(A)- RNAs in the cytoplasm originate from normal polyadenylated mRNA that had their poly(A) tails reduced to less than 20 nt because they were undergoing degradation (see the lower S2 RNA in Figure 6.1).

The low number of unique polyadenylation sites in the pure poly(A)- RNA of the cytoplasm (Figure 6.3F) reflect partially that there is less poly(A)- RNA than poly(A)+ RNA in the cytoplasm. However, it may also be an indication that the number of false positives in the study is not high. If false positives are not biased toward any compartment or any of the poly(A)- or poly(A)+ fractions, the low number of polyadenylation sites in the poly(A)- RNA in the cytoplasm may be an indication of the number of false positives in the study.

The nuclear poly(A)- RNA was found to be enriched in polyadenylation sites compared to the cytoplasmic poly(A)- RNA (Figures 6.4F and 6.4D). Removing the sites common to nuclear poly(A)- and poly(A)+ RNA revealed that most polyadenylation sites in the $3'$ UTR of the nuclear poly(A)- RNA were common with the nuclear poly(A)+ RNA (Figure 6.3H). We propose that the polyadenylation sites in the $3'$ UTR common to the nuclear poly(A)+ and poly(A)- RNA can represent mRNA undergoing polyadenylation (see the upper S2 RNA in Figure 6.1).

However, unlike for the cytoplasmic RNA, there was after separation still an enrichment of polyadenylation sites in the intergenic and intronic regions of the poly(A)- RNA. This enrichment could be partially expected since since introns comprise a much larger part of nuclear RNA than cytoplasmic

RNA, simply because human pre-mRNA sequences contain over 20 times more intronic than exonic sequence [10]. However, all intronic sites are not likely to be false positives; and judging from the cytoplasmic pure poly(A)- fraction (Figure 6.3F) the number of false positives in the study may be low. We propose that the general enrichment in the poly(A)- sites in the nuclear fraction compared to the cytoplasm may be due to degradation-related transient polyadenylation that occurs in the nucleus of mammalian cells [6] (see S3 RNA in Figure 6.1). The polyadenylated RNA in the nuclear intergenic region may stem from degradation of spurious transcripts as found in yeast [11]. The polyadenylation sites in the intronic region may stem from a yet-to-be discovered polyadenylation-mediated degradation mechanism for introns, as has been postulated [9].

We end the discussion with a quote from the review by N. Proudfoot “Ending the message: poly(A) signals then and now” [12]. Using early Sanger-like sequencing techniques in 1976, Proudfoot was the original discoverer of the AAUAAA hexamer, and deduced both that this was the signal for polyadenylation and that the signal was conserved across mammals [13]. In this regard, Proudfoot is in a good position to comment on the recent surge in genome-wide polyadenylation studies. He writes [12]:

“It is abundantly clear that bioinformatic analysis of genomic data has provided invaluable generality to our understanding of PAS [polyadenylation signal] function in gene expression. However, current genome-wide analyses often only provide bioinformatic correlations and lack direct functional experimentation. Genomic analysis will only achieve its full potential when bioinformatics can be matched by hypothesis-driven experimental approaches.”

6.6.1 Summary

In summary, by studying polyadenylation in six dimensions (whole cell, cytoplasm, and nucleus in the poly(A)+ and poly(A)- fractions) compared to normally just one (whole cell, poly(A)+), we have made observations that would otherwise not be possible to make. First of all, the polyadenylation sites in whole cell RNA were shown to be the sum of polyadenylation sites the nuclear and the cytoplasmic RNA, which verifies the integrity of the datasets. Second, by separating the polyadenylation sites common to poly(A)+ and poly(A)- RNA, we were able to potentially identify between polyadenylation signals from mRNA undergoing degradation in the cytoplasm and mRNA undergoing polyadenylation in the nucleus. Further, we identified an enrichment in intronic and intergenic polyadenylation sites in the poly(A)- RNA in the nucleus which is not present in the cytoplasm. These polyadenylation sites may be traces of the newly discovered polyadenylation-mediated degradation mechanism in the nucleus of eukaryotic cells [6].

6.7 Materials and Methods

6.7.1 Methods

Since no tool for obtaining information about sites of polyadenylation for an RNA-seq experiment had been published, we developed a tool called *Utail* (from Untranslated poly(A) tail) for this purpose. Utail takes as input mapped RNA-seq reads and outputs a tab-delimited list of all the polyadenylation sites in the RNA-seq data along with other statistics and relevant information for the analysis of those polyadenylation sites.

The Utail pipeline involves identifying poly(A) reads by trimming poly(A/T) stretches of unmappable reads, remapping the poly(A) reads after trimming, and clustering the 3′ ends of the mapped reads are within 20 nucleotides of each other, since this is close to the maximal range of the stochastic effect of choice of 3′ cleavage site [14]. After clustering the genomic sequence 40 nucleotides downstream the polyadenylation site is searched for one of the polyadenylation signals (PAS). As well, the polyadenylation sites are intersected with the 3′ PET sites (see below) to look for overlap between poly(A) clusters and PET. To avoid false positives, Utail checks the genomic sequence for poly(A) tracts and discarded putative poly(A) reads that land at genomic locations that are poly(A) rich. Pseudo code for Utails follows later in the Materials and Methods section.

Presently, the input reads to Utail must be provided either in the output format of the GEM mapper [15] or as Bed-files. If a genome annotation is provided, the Utail output will include information about the proximity of annotated polyadenylation sites to the polyadenylation sites Utail detects. If a reference annotation is provided, it is important that the annotation is based on the same version of the reference genome that the RNA-seq reads were mapped to.

6.7.2 The dataset

The datasets used in this study are available from
http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/
encodeDCC/wgEncodeCshlLongRnaSeq/

The data was generated from RNA-seq experiments from 12 human cell lines. Six of the cell lines have RNA-seq data from whole cell extracts and the remaining six cell lines have RNA-seq data from cytoplasmic and nuclear extracts in addition to whole cell extracts. RNA-seq data is further available for both the poly(A)+ and poly(A)- RNA pools for each cell line. Table 6.1 shows the cell lines and compartments used and the number of replicates. In total, 23 datasets were from whole cell extracts, 11 from cytoplasmic extracts, and 12 from nuclear extracts. This brings the total to 92 RNA-seq datasets. The datasets contains only RNA-seq from long RNA, defined as RNA over 200 nucleotides in length. Each dataset has been generated with Illumina paired-ended sequencing with a read-length of 76 basepairs and contains between 150 and 250 million reads.

Cell line	Whole Cell	Cytoplasm	Nucleus
GM12878	2	2	2
K562	2	2	2
HeLa-S3	2	2	2
HUVEC	2	2	2
HEPG2	2	2	2
H1Hesc	1	1	1
Nhek	2	0	1
MCF7	2	0	0
AG04450	2	0	0
HSMM	2	0	0
NHLF	2	0	0
A549	2	0	0

Table 6.1 – Number of replicates of the compartmentalzied RNA-seq data from 12 different cell lines from the ENCODE consortium.

6.7.3 The short RNA mapper

The short read mapper used in this work is the GEM mapper [15].

6.7.4 The PET data

Paired End diTag (PET) sequencing is a technique that is specific for locating the $3'$ end of transcripts. This technique adds two tags to the $3'$ end and the $5'$ end of an RNA. Those tags then join with each other, forming a circular RNA. Subsequently, the sequences close to the two tags (less than 30 nt) are sequenced. Thereby, one obtains the sequence information about both the transcription start site and the transcript termination site of the RNA. We have used PET data which is available from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeGisRnaPet/ to compare with the polyadenylation sites we discovered. We used a threshold of 10 PET reads to accept a site as transcript termination site.

6.7.5 Bioinformatic implementation

Here follows a pseudo code for the implementation of Utail. The pseudo code gives an accurate overview of the algorithms of the pipeline but does change the order of some of the execution steps. This is partially because Utail was designed to perform more functions that what it has been used for in this study, and we omit the description of those functions to give a relevant presentation of the code. The full source code is available as outlined in the Appendix on page 87.

```
# Load the settings object from the settings file. The settings object
# has information about i) the location of the mapped RNA-seq datasets,
# ii) the genome, and iii) optionally a genome annotation. Other
# settings, such as the number of CPU cores to be used by the pipeline,
# are also found in the settings
# file.
settings = Settings(read_settings(settings_file))

#### Obtaining poly(A) reads ####
#
# The first part of the pipeline reads through an already-mapped RNA-seq
# library and collects the unmapped reads. Those reads are inspected if
# they have a poly(A) end or a poly(T) start. If so, they are stored in
# a fasta file.
#

# Loop though each dataset in the settings; represented by ID and file
# paths.  Poly(A) reads are identified, trimmed, remapped, and clustered
# for each dataset_id separately. A dataset_id could be for example on
# the type CellLine_Compartment_RNA_fraction (e.g.
# HeLa_Nucleus_Poly(A)+), and the dataset_paths refer to the individual
# RNA-seq datasets associated with this dataset_id.
for (dataset_id, dataset_paths) in settings.datasets.items():

    # Create a unique location for storage of the filtered poly(A) reads
    polyA_reads_path = 'polyA_storage/polyA_reads'\
                             + dataset_id + '.fasta'

    # Go through each dataset-file to read from it
    for file_path in dataset_paths:
```

```
    # Go through each line in the input file
    # Each line corresponds to a read from a mapping of an RNA-seq
    # experiment. The read was either mapped or unmapped
    for read in file_path:

        # Only investigate unmapped reads. They are potential
        # poly(A) reads
        if was_not_mapped(read):

            # Check if the read begins with poly(T). This uses
            # regular expressions to allow for nucleotide mismatches
            # from either misincorporation by the poly(A) polymerase
            # or sequencing noise
            if begins_with_T(read):

                # Remove the poly(T) part of the read using a
                # recursive algorithm
                stripped_read = strip_read(read, remove='T')

                # Check if the stripped read is long enough to be
                # reliably re-mapped to the genome (> 25 nt)
                if len(stripped_read) > 25:

                    # Write read to file
                    write(stripped_read, location=polyA_reads_path)

            # Do the same for reads that end with poly(A)
            if ends_with_A(read):

                stripped_read = strip_read(read, remove='A')

                if len(stripped_read) > 25:

                    write(stripped_read, location=polyA_reads_path)

##### Re-mapping the poly(A) reads ####
#
# Now that some putative poly(A) reads have been obtained, we
# attempt to remap them to the genome to identify sites of RNA
# cleavage and polyadenylation.  If a read is successfully remapped,
# we check the genomic sequence for poly(A/T) stretches, and look
# for polyadenylation signal (PAS) upstream the cleavage and
# polyadenylation site.
#
```

```
# Define an output path for the successfully remapped poly(A) reads
read_file = open("remapped_polyA_reads")

# Create the command for the gem-mapper which remaps the reads to
# the human genome (hg19). Accept up to two nucleotide mismatches
# when mapping.
mapping_command = "gem-mapper input_reads="+polyA_reads_path +"\
                                -mismatches=2"

# Run the mapping program and obtain the path for the output of the
# mapping
mapping_output = execute(mapping_command)

# Go through all the putative polyA reads obtained in the previous
# step
for remapped_read in open(mapping_output):

    # Only process putative poly(A) reads that were uniquely
    # remapped
    if was_uniquely_mapped(remapped_read):

        # Obtain the genomic location of the remapped reads
        genomic_location = get_loacation(remapped_read)

        # Check if the trimmed poly(A/T) tail of the poly(A) read
        # overlaps with a poly(A/T) rich region on the genome
        if not_noise(remapped_read):

            # Write to file the remapped read and its genomic
            # location
            read_file.write(remapped_read, genomic_location)

##### Cluster the poly(A) sites of the re-mapped poly(A) reads ####
#
# Now that we have obtained the poly(A) reads, we cluster together
# the poly(A) sites which are the 3' ends of the poly(A) reads.
# Cleavage and polyadenylation is a stochastic process in terms of
# the choice of cleavage site. A given poly(A) site can vary with as
# much as 40 nt. The finally reported poly(A) site is the center of
# the cluster.

# The clustering happens iteratively. First, the genome is broken
# down into the different genomic regions (3UTR, 5UTR, exonic,
# intronic), and then those regions are further divided into unique
# segments on each chromosome.  These segments are then intersected
# with the polyA reads and each segment is clustered individually.

# Break down the genome into genomic segments and intersect the
# regions with the poly(A) reads
segment_with_polyA_reads = intersect_genome_polyA(settings.genome,
                                                  "remapped_polyA_reads")

# Keep track of the polyA sites in each segment individually in a
# hash-table (= dictionary in python)
polyA_reads = {}

# Loop over the segment ID (e.g. chr1_13_78_exonic) and the list of
# polyA sites in that segment
for (segment_id, polyAs_in_segment) in segment_with_polyA_reads.items():

    # Cluster separately the poly(A) sites that have landed on the
    # positive and the negative strand
    plus_sites = []
    minus_sites = []

    # Loop over the polyA_reads in each segment
    for polyA_read in polyAs_in_segment:

        # To be able to cluster the poly(A) sites, it is necessary
        # to identify the exact location of the poly(A) read. Even
        # though the sequencing protocol is not strand specific,
        # poly(A) reads can be identified in a strand specific way.
        # To find out, you need to know if the read was identified
        # with a poly(A) tail or a poly(T) header, and which strand
        # the polyA read mapped to. (This theory is covered in the
        # introduction of the thesis.)

        # Get the beginning and the end locations of the mapped
        # read, the strand (+ or -), and if the read was trimmed for
        # poly(A) or a poly(T)
        (beg, end, strand, A_or_T) = get_read_info(polyA_read)

        # The poly(A) site originated from the - strand. Therefore,
        # the poly(A) site corresponds to the beginning of the
        # mapped read.
        if (A_or_T == 'T' and strand == '+'):
            minus_sites.append(beg)

        if (A_or_T == 'A' and strand == '-'):
            minus_sites.append(beg)
        # The poly(A) site originated from the + strand. Therefore,
        # the poly(A) site corresponds to the end of the mapped
        # read.
        if (A_or_T == 'T' and strand == '-'):
            plus_sites.append(end)

        if (A_or_T == 'A' and strand == '+'):
            plus_sites.append(end)

    # Cluster the poly(A) sites on the minus and plus strand
    # individually.  The clustering algorithm iteratively adds
    # poly(A) sites to clusters and updates the cluster-center after
    # each poly(A) site is added. New poly(A) sites are added to the
    # cluster if they are within +/- 20 nt of the cluster center.
    # The exact clustering implementation is omitted here.
    polyA_reads[segment_id] = (cluster(plus_sites),
                               cluster(minus_sites))


##### Finding upstream polyadenylation signal and annotated poly(A) sites ####
#
# Now that the poly(A) sites have been clustered, we search for a
# polyadenylation signal (PAS) upstream of these sites.
# Additionally, we see if the poly(A) sites are located close to
# already annotated poly(A) sites (annotated sites are obtained from
# the genome annotation specified in the settings-file). After PAS
# and annotated poly(A) site have or have not been found, we write
# to the final output file the following information about each
# poly(A) site:
#
# chromosome; start; end; strand; PAS; PAS_distance;
# annotated_distance; nr_of_reads_in_cluster; A_or_T;
#
# Here, PAS_distance is how far the PAS signal is from the poly(A)
# site, annotated_distance is the distance of an annotated poly(A)
# site from the cluster center, nr_of_reads_in_cluster is the number
# of poly(A) reads that were clustered together to form this poly(A)
# site, A_or_T signifies if the poyl(A) read had leading Ts or
# trailing As.
#

# the final output file
final_output = open('final_output_file')
```

```
    # Go through each segment
    for segment_id, clusters in polyA_reads.items():

        # Go through each poly(A) site cluster in the segment
        if cluster in clusters:

            # Get any PAS and distance to PAS site from cluster center
            PAS, PAS_distance = get_PAS(cluster)

            # Get the distance to annotated poly(A) site
            annotated_distance = get_annotated_distance(cluster,
                                                        settings.annotation)

            # get genomic location of cluster
            chrom, beg, end, strand = get_location(cluster)

            # get number of reads in cluster
            nr_of_reads_in_cluster = get_cluster_size(cluster)

            # Write to file
            final_output.write(chrom, beg, end, strand, PAS,
                               PAS_distance, annotated_distance,
                               nr_of_reads_in_cluster, A_or_T)


# Plots can now be generated from these output files
# generate plots.
```

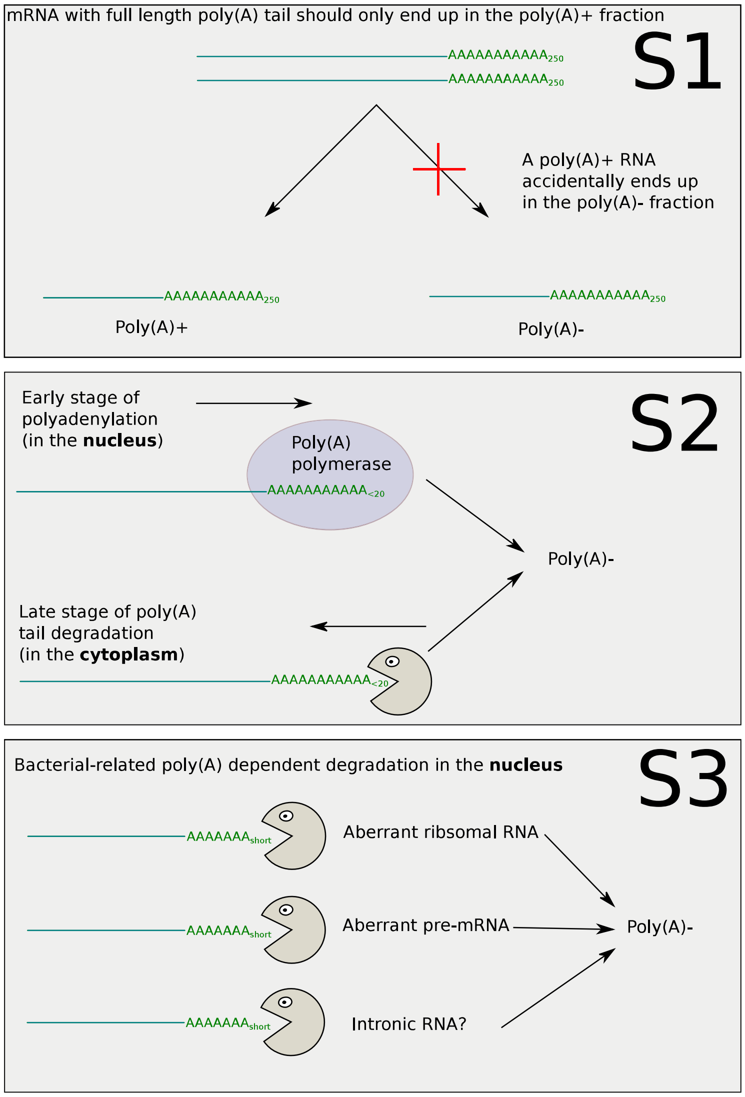
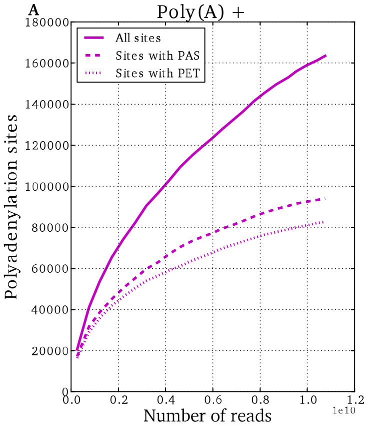

mRNA with full length poly(A) tail should only end up in the poly(A)+ fraction
S1
AAAAAAAAAAA250
AAAAAAAAAAA250
A poly(A)+ RNA accidentally ends up in the poly(A)- fraction
AAAAAAAAAAA250
Poly(A)+
AAAAAAAAAAA250
Poly(A)-
Early stage of polyadenylation (in the nucleus)
S2
Poly(A) polymerase
AAAAAAAAAAA<20
Poly(A)-
Late stage of poly(A) tail degradation (in the cytoplasm)
AAAAAAAAAAA<20
Bacterial-related poly(A) dependent degradation in the nucleus
S3
AAAAAAAshort
Aberrant ribsomal RNA
AAAAAAAshort
Aberrant pre-mRNA
Poly(A)-
AAAAAAAshort
Intronic RNA?

Figure 6.1 – Different proposed origins of polyadenylation signals in poly(A)-. **S1** RNA is a result of random error in the technical separation of poly(A)+ and poly(A)- RNA. **S2** RNA is the result of polyadenylated mRNA which have a short poly(A) tail at the moment of sampling, either due to early polyadenylation in the nucleus (upper) or degradation in the cytoplasm (lower). **S3** RNA are proposed to be RNA undergoing poly(A)-dependent degradation in the nucleus.



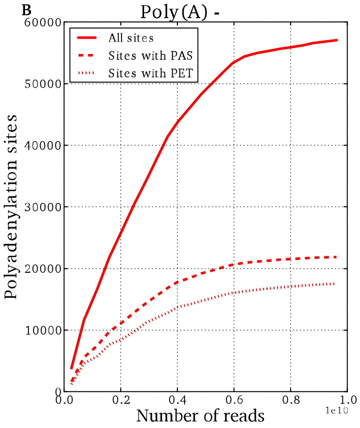


Figure 6.2 – The number of identified polyadenylation sites increases and slowly saturates as more datasets are included. **A**: poly(A)+ RNA. **B**: poly(A)- RNA.

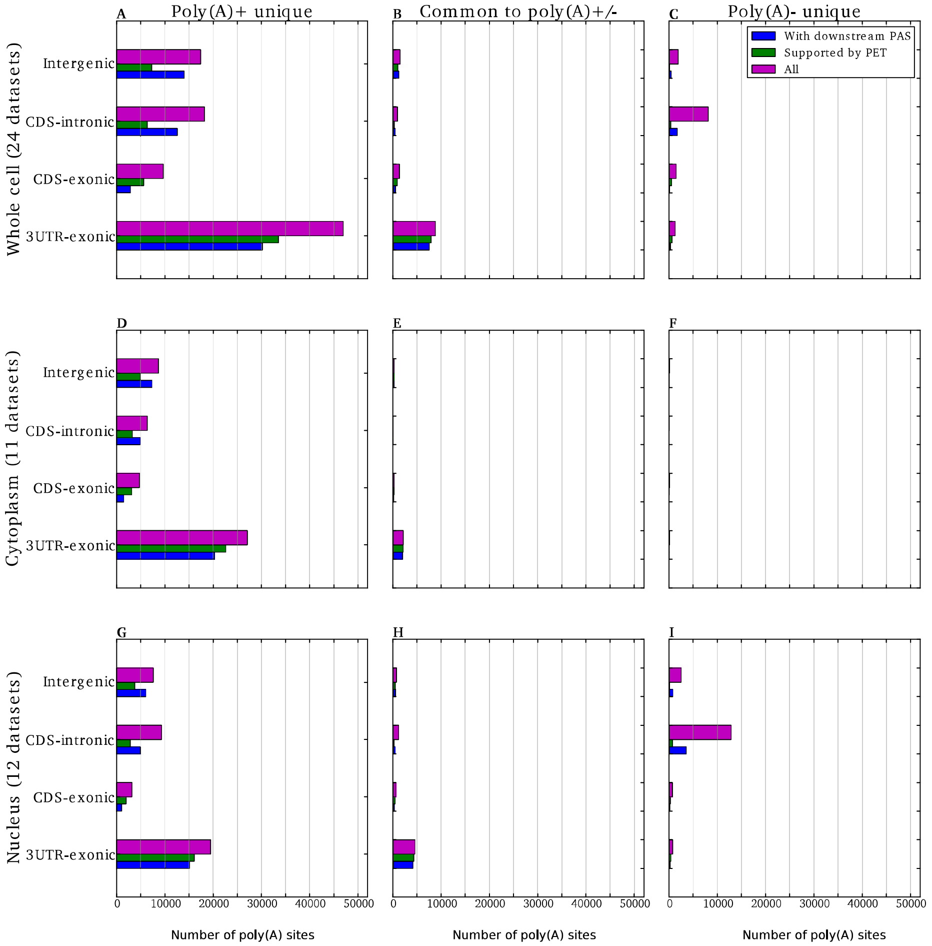


Figure 6.3 – Separation of unique and common polyadenylation sites across genomic regions poly(A)+ and poly(A)- RNA. The left and right columns show the number of polyadenylation sites after removing the sites common to the poly(A)+ and poly(A)- fractions. The middle column shows the sites that were common. Row one, two, and three show polyadenylation sites in the whole cell, cytoplasm, and nucleus, respectively. The blue line indicates the fraction of sites which had a PAS within 40 basepairs downstream. The green line shows the sites found within 50 nucleotides of a PET signal.

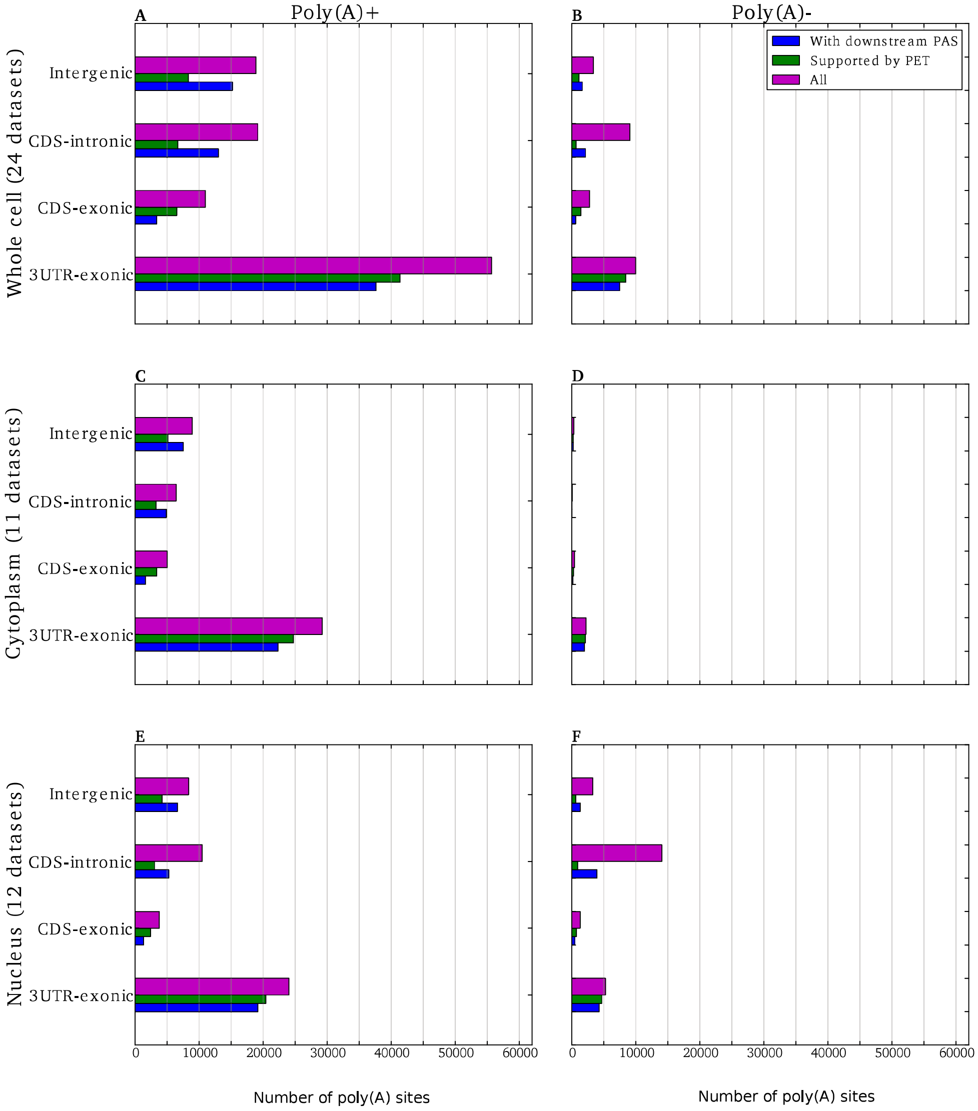


Figure 6.4 – Polyadenylation across genomic regions for the poly(A)+ and poly(A)- fractions. See Figure 6.3 for further description.

Bibliography

[1] Ewan Birney et al. “Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project”. *Nature* 447.7146 (June 2007), pp. 799–816.

[2] Cole Trapnell et al. “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation”. *Nature Biotechnology* 28.5 (2010), pp. 511–515.

[3] Jennifer Harrow et al. “GENCODE: The reference human genome annotation for The ENCODE Project”. *Genome Research* 22.9 (Sept. 2012), pp. 1760–1774.

[4] J. Y. Lee et al. “PolyADB 2: mRNA polyadenylation sites in vertebrate genes”. *Nucleic Acids Research* 35.Database (Jan. 2007), pp. D165–D168.

[5] Bin Tian, Zhenhua Pan, and Ju Youn Lee. “Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing”. *Genome Research* 17.2 (Feb. 2007), pp. 156–165.

[6] Jean-François Lemay et al. “The Nuclear Poly(A)-Binding Protein Interacts with the Exosome to Promote Synthesis of Noncoding Small Nucleolar RNAs”. *Molecular Cell* 37.1 (Jan. 2010), pp. 34–45.

[7] Natalia Shcherbik et al. “Polyadenylation and degradation of incomplete RNA polymerase I transcripts in mammalian cells”. *EMBO Reports* 11.2 (Feb. 2010), pp. 106–111.

[8] Steven West et al. “Adenylation and Exosome-Mediated Degradation of Cotranscriptionally Cleaved Pre-Messenger RNA in Human Cells”. *Molecular Cell* 21.3 (Feb. 2006), pp. 437–443.

[9] Marie-Joëlle Schmidt and Chris J. Norbury. “Polyadenylation and beyond: emerging roles for noncanonical poly(A) polymerases”. *Wiley Interdisciplinary Reviews: RNA* 1.1 (July 2010), pp. 142–151.

[10] J. Craig Venter et al. “The Sequence of the Human Genome”. *Science* 291.5507 (Feb. 2001), pp. 1304–1351.

[11] Françoise Wyers et al. “Cryptic Pol II Transcripts Are Degraded by a Nuclear Quality Control Pathway Involving a New Poly(A) Polymerase”. *Cell* 121.5 (June 2005), pp. 725–737.

[12] Nick J. Proudfoot. “Ending the Message: poly(A) Signals Then and Now”. *Genes & Development* 25.17 (Sept. 2011), pp. 1770–1782.

[13] N. J. Proudfoot and G. G. Brownlee. “3’ non-coding region sequences in eukaryotic messenger RNA”. *Nature* 263.5574 (Sept. 1976), pp. 211–214.

[14] B. Tian. “A large-scale analysis of mRNA polyadenylation of human and mouse genes”. *Nucleic Acids Research* 33.1 (Jan. 2005), pp. 201–212.

[15] Santiago Marco-Sola et al. “The GEM mapper: fast, accurate and versatile alignment by filtration”. *Nature Methods* (2012).

Chapter 7

Discussion and Conclusion

7.1 Discussion

The three topics that have been investigated in this thesis – sequence dependent abortive initiation, sequence dependent translation initiation, and sequence dependent cleavage and polyadenylation – all center around how the genetic code regulates gene expression at these different levels.

When working with these projects we have made use of two distinct bioinformatic approaches. For transcription and translation initiation we worked with kinetics and thermodynamic calculations in a gene-centric manner, while for the 3′ cleavage and polyadenylation study we analyzed next generation sequencing data to perform a genome-wide investigation. Having both these approaches in this thesis illustrates well the methodological contrast that exists in biology today, where "-omic" studies are increasingly finding applications in areas that were previously only studied with traditional molecular biology techniques.

What follows is a discussion of the way these two approaches have been used in this thesis to provide new knowledge about gene expression. Additionally, the two approaches will be compared in terms of how they facilitate different research strategies and how the lead to different kinds of research challenges. A summary of this discussion is given in Table 7.1.

7.1.1 Data, research strategies, and challenges

What fundamentally separates genome-wide and gene-centric studies in general are the underlying data. The data from the transcription initiation studies were in terms of radioactive intensity of gel-migrating RNA oligonucleotides and was generated *in vitro*. This kind of data is straightforward to interpret (abortive and full length RNA), and in the controlled, *in vitro* experimental set-up it is easy to perturb the input DNA sequence to achieve variation in the output for hypothesis testing. The data from the work on translation initiation were from western-blots and RT-PCR. Again the interpretation of the data is relatively straightforward (relative protein and mRNA concentrations).

Since the data from these gene-centric experiments were easy to interpret and low in volume, they did not require extensive filtering or heavy computational analysis. Importantly, this means that less time was devoted to data management and data analysis, and more time was spent on interpretation of the biological meaning of the experiments. In other words, for the gene-centric studies the challenge was to understand the biological mechanism that was studied. For translation initiation we needed to understand which steps during gene expression are affected by mutations around the 5′ ribosome binding site. For transcription initiation we had to combine how RNA polymerase moves relative to DNA with abortive RNA synthesis; the challenge was to relate how the equilibrium constant of translocation was linked with the abortive release of short RNA.

On the other hand, the RNA-seq data underlying the study of 3′ cleavage and polyadenylation imposed a different strategy for this project. Due to the size and complexity of the data, the main challenge for the work was handling, filtering, and processing the data. Much time was devoted to constructing the analysis pipeline named Utail that was used for data filtering and analysis. The datasets from all the different cell lines and compartments were over 1 terabyte in uncompressed form and took two days to process with a powerful, multi-core workstation. As a direct consequence of the data-challenge, there was less time devoted to the study of the biological mechanism. Perhaps anticipating that this would be the case, the initial

	Genome-wide	Gene-centric
Experiment type	RNA-seq	RT-PCR, western blot, *in vitro* transcription assay
Data amount	1 terabyte	1 megabyte
Data processing	Significant	Minimal
Data interpretation	Challenging	Straight forward
Experimental follow-up	No	Yes
Study objective	General	Specific
Results	Overview, hypothesis	Mechanistic insight

Table 7.1 – A summary of the discussion. The left column holds different characteristics that differ between the genome-wide and gene-centric studies.

goal for the project was made general: to characterize polyadenylation in different cell compartments for different human cell lines. This is in contrast to the more detailed, mechanism-oriented goals for the gene-centric experiments, namely the movement of RNAP relative to DNA and the binding of the ribosome to mRNA. Since less time was spent on the computational aspect, the gene-centric approach facilitated collaboration with wet-lab biologists about the biological problem itself. The direct involvement of wet-lab biologists with the core scientific problem is a likely reason why these studies resulted in follow up experiments. Therefore, by having less computational focus, gene-centric approaches may increase the chance of having follow-up experimental work, as they are more likely to involve the collaboration of wet-lab biologists.

7.1.2 Contribution to knowledge about gene expression

Due to the differences in underlying data and study approach, the genome-wide and gene-centric studies are bound to give rise to different types of biological research questions that can be asked and answered. In chapter 3 we were able to point to ribosome binding as a limiting factor for the heterologous expression of *inf-α2b*. This is a specific type of information for a specific gene, but it does verify previous reports that translation initiation can be a bottleneck for heterologous expression, thereby adding to an existing body of evidence. The study also reinforced the message that using RNA-RNA folding energy from secondary structures is a useful approach to augments standard techniques in molecular biology like RT-PCR and western blot when working with optimization of gene expression.

In chapter 4 we propose an answer to the question of why abortive initiation varies with the initial transcribed sequence. We found that sequences which were predicted to bias RNAP toward the pre-translocated state were also those sequences associated with the lowest productive yields. As an explanation, we linked the bias toward the pre-translocated state toward RNAP having a higher probability of backtracking, which leads to abortive RNA release. Even though this study used only the N25 promoter, these results are likely to hold for all strong promoters that undergo abortive initiation. Crucially, this study linked several topics that had not been linked together before. First, it approximated equilibrium constants of translocation from those of pyrophosphorolysis; thereafter it linked translocation to the efficiency of promoter escape, before finally completing the logical circuit by reasoning that backtracking from the pre-translocated state is the rate limiting step during transcription initiation. When we began the transcription initiation study we used the free energies of the RNA-DNA and

DNA-DNA duplexes to study transcription initiation. These were at that time the known components that contribute energetically to translocation by RNAP. However, it turned out that the free energy of the interaction between the RNA 3′ dinucleotide and the RNAP active site was the most important variable in the study. This is a good example of how one sets out by building on the existing knowledge (the RNA-DNA hybrid and DNA-DNA bubble contribute to RNAP translocation), and ends up adding a new component (free energy associated with the 3′ dinucleotide matters more).

Chapter 4 focused on modelling translocation during initial transcription, but the equilibrium model used there is not capable of describing the kinetics of the process. To therefore investigate the kinetics of initial transcription, we introduced in chapter 5 a model that covers the nucleotide addition cycle (which includes translocation), backtracking, and unscrunching and abortive RNA release. This model has a sequence-dependence in that it uses sequence-dependent abortive probability profiles to infer the rate of backtracking. By fitting this model to published kinetic data of abortive cycling, we were able to pin down key rate constants for the process. We showed that that backtracking and abortive RNA release are slow steps compared to forward transcription, confirming previous results that had indicated this. Having identified the key rate constants, we then showed how the median time to reach promoter escape increases strongly when initial transcription is performed in the absence of GreB. Most interestingly, we showed that the speed of promoter-bound transcription, indicated by the rate constant of the nucleotide addition cycle, is highly similar to that of promoter-free transcription. This result is important, as it shows that being bound to a promoter does not slow down transcription. Taken as a whole, the kinetic evidence points to that the only difference between promoter-bound and promoter-free transcription is that promoter-bound transcription has a comparatively higher rate of backtracking. To date, there is no published data of experiments being able to follow the rapid kinetics of initial transcription. Our work shows how a computational model, aided with the right experimental findings, is able to fill this gap in the literature.

The main message from the study in chapter 6 is that polyadenylation sites inferred from RNA-seq vary between different genomic regions for different cellular compartments. In particular, we found that there was a substantial increase in polyadenylation sites from poly(A)- RNA in the intergenic regions of nuclear extract that was not present in the poly(A)+ RNA. This may be an indication of regulation by gene expression through degradation-related polyadenylation of intronic RNA. This contribution to the understanding of gene expression is of a different kind than for the gene-centric studies. Instead of analyzing a particular aspect of polyadenylation, we looked broadly were able to identify general patterns. In turn, this result should be used to investigate how intronic RNA is degraded in mammalian cells. As such, the outcome of the genome-wide study becomes a roadmap from which hypotheses can be formed. Our results show the importance of studying the transcriptome separately for different cell compartments.

7.2 Conclusion

The target of this thesis was to use computational tools in combination with wet-lab experiments to study regulation of gene expression.

In chapter 3, we found that decreasing the RNA secondary structure around the ribosome binding site led to increased transcript expression of the *inf-α2b* gene in *E. coli*, likely as a result of increased RNA stability through ribosome binding. We did this by *in silico* screening a library of variants of *inf-α2b* that had synonymous mutations in the first 9 codons, and testing those variants whose synonymous mutations led to decreased secondary structures in the ribosome binding site. We conclude that heterologous gene expression of *inf-α2b* in *E. coli* may be limited by secondary structures around the translation start site.

In chapter 4, we found that the sequence of the ITS modulates the promoter escape efficiency through its effect on translocation. This study proposes a solution for the 25-year case of why the initial transcribed sequence modulates promoter strength and promoter escape efficiency. In addition, contrary to what was previously assumed, we found no role for the free energy of the scrunched DNA bubble in modulating promoter escape efficiency. We conclude that Gibbs free energy of the DNA bubble is unrelated to promoter escape efficiency, and that sequence dependent productive yield from initial transcription can be explained by variation in translocation caused by interactions between the $3'$ dinucleotide and internal sites in RNAP.

In chapter 5 we investigated the kinetics of initial transcription on the N25 promoter using sequence-dependent abortive probabilities. We conclude that the nucleotide addition cycle proceeds at the same average speed for promoter-bound and promoter-free RNAP, and that the rate limiting steps for promoter escape are unscrunching and abortive RNA release.

In chapter 6, we classified evidence of polyadenylation in different cell compartments for 12 human cell lines. We found evidence of polyadenylation of poly(A)- RNA in intergenic and intronic regions. We conclude that our findings indicate evidence of the recently discovered nuclear degradation-related polyadenylation in human cells, and that RNA with short poly(A) tails may be abundant in the nucleus.

Overall, the findings in this thesis have shed additional light on some of the myriad ways in which gene sequences influence regulation of gene expression, and shows how the potential of bioinformatics and computational biology is best achieved when used predictively to inform experimental biology.

Appendix A

Obtaining the source code

In this section we describe how the computational results in this thesis can be reproduced. The source code for the different projects is too large to be physically included in the thesis, and as such is stored on the source code hosting site Github (`www.github.com`). The source code is being made publicly available on Github as the articles from each project are published. At the time of writing, the source codes for chapters 4 and 3 are not yet publicly available.

To obtain the source code from Github, the program *git* must be installed (`www.git-scm.com`). Use git from the command line to obtain the code like this:

```
git clone git@github.com:jorgsk/exampleProject.git
```

The Github locations for the different projects are given below:

- Translation initiation project (chapter 3):

  `git@github.com:jorgsk/translation_initiation.git`

- Transcription initiation project (chapter 4):

  `git@github.com:jorgsk/transcription_initiation.git`

- $3'$ cleavage and polyadenylation (chapter 6):

  `git@github.com:jorgsk/Utail.git`

After downloading the code for each project, see the readme-file for each project for further instructions. The readme-file contains instructions for running the source code as well as a list of software dependencies for each project. However, all projects have the following common requirements:

- Modern Linux distribution (i.e. Ubuntu Linux)
- Python version 2.7+
- The Python libraries

  Numpy (version 1.5+)
  Scipy (version 0.9+)
  Matplotlib (version 1.0+)