Speed and accuracy in transcription and translation

*Modelling of transcript and polypeptide elongation*

HARRIET MELLENIUS
Abstract

The information encoded in DNA is transcribed to mRNA by the RNA polymerase, and then translated to proteins by the ribosome. The processes of transcription and translation have evolved to be very fast and very accurate, despite the fact that enzymatic accuracy usually comes at a cost of reaction speed. Here, we make mathematical models of prokaryotic transcription and translation, to quantitatively describe the speed and accuracy under various conditions. The models are based on experimental measurements of accuracy, reaction rates, and interaction energies of base pairs of DNA and RNA.

We find that there is a large variation in the transcriptional accuracy due to the template DNA. The variation arises from the sequences of DNA and RNA in the transcription bubble that have a very large variation in free energy between the bases. Due to the large accuracy variation, transcription errors will primarily occur in error hot spots. The error rate in these positions can be larger than the average error rate of translation, making transcription a putative source of amino acid substitution errors for some sequence contexts. Furthermore, the model predicts that transcriptional accuracy is maintained by two steps of proofreading.

The speed of translation in the presence of the antibiotic fusidic acid is described in another model that includes the effect of interactions between ribosomes on the same mRNA. When fusidic acid binds to a ribosome-bound elongation factor G, the ribosome is stalled and other ribosomes can catch up and form queues. The queues hence increase the effect of fusidic acid on the elongation rate, but the queuing also decreases the number of ribosomes on the mRNA, reducing the queue formation.

Finally, the variance in a dataset of translational accuracy measurements was analyzed, and a complete model of the translational accuracy for any codon was designed. The model shows that the accuracy variation in translation is sensitive to the tRNA concentrations, which together with the codon usage bias counters the speed–accuracy trade-off by the cells investing in the translation of already accurate codons to make them faster.

Keywords: RNA polymerase, Transcription, Ribosome, Translation, Modelling, Accuracy, Proofreading

Harriet Mellenius, Department of Cell and Molecular Biology, Structure and Molecular Biology, 596, Uppsala University, SE-751 24 Uppsala, Sweden.

© Harriet Mellenius 2015

ISSN 1651-6214
urn:nbn:se:uu:diva-262903 (http://urn.kb.se/resolve?urn=nbn:se:uu:diva-262903)
“Fast is fine, but accuracy is final. You have to learn to be slow in a hurry.”

Wyatt Earp
List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


Reprints were made with permission from the respective publishers.

* Authors contributed equally to this work.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>9</td>
</tr>
<tr>
<td>The central dogma of life</td>
<td>9</td>
</tr>
<tr>
<td>The problematic accuracy</td>
<td>10</td>
</tr>
<tr>
<td>Dissertation overview</td>
<td>11</td>
</tr>
<tr>
<td>Methods</td>
<td>13</td>
</tr>
<tr>
<td>Theoretical biology</td>
<td>13</td>
</tr>
<tr>
<td>Kinetics and reaction rates</td>
<td>13</td>
</tr>
<tr>
<td>Michaelis-Menten kinetics and the master equation – an example</td>
<td>15</td>
</tr>
<tr>
<td>Mean time models</td>
<td>17</td>
</tr>
<tr>
<td>Stochastic simulations</td>
<td>18</td>
</tr>
<tr>
<td>Statistical models</td>
<td>19</td>
</tr>
<tr>
<td>Accuracy in theory</td>
<td>20</td>
</tr>
<tr>
<td>Enzymatic accuracy</td>
<td>20</td>
</tr>
<tr>
<td>Initial selection</td>
<td>22</td>
</tr>
<tr>
<td>Proofreading selection</td>
<td>23</td>
</tr>
<tr>
<td>Transcription</td>
<td>26</td>
</tr>
<tr>
<td>Models of transcription</td>
<td>28</td>
</tr>
<tr>
<td>The modelled transcription speed</td>
<td>30</td>
</tr>
<tr>
<td>The modelled transcription accuracy</td>
<td>31</td>
</tr>
<tr>
<td>Translation</td>
<td>35</td>
</tr>
<tr>
<td>The modelled translation speed in the presence of fusidic acid</td>
<td>37</td>
</tr>
<tr>
<td>Statistical model of the translation accuracy</td>
<td>38</td>
</tr>
<tr>
<td>Discussion</td>
<td>40</td>
</tr>
<tr>
<td>Summary in Swedish: Hastighet och noggrannhet i transkription och translation</td>
<td>43</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>46</td>
</tr>
<tr>
<td>References</td>
<td>48</td>
</tr>
</tbody>
</table>
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>tRNA</td>
<td>Transfer RNA</td>
</tr>
<tr>
<td>aa-tRNA</td>
<td>Aminoacyl-tRNA</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>rRNA</td>
<td>Ribosomal RNA</td>
</tr>
<tr>
<td>bp</td>
<td>Base pair(s)</td>
</tr>
<tr>
<td>NTP</td>
<td>Nucleoside triphosphate</td>
</tr>
<tr>
<td>A</td>
<td>Adenine or Adenosine</td>
</tr>
<tr>
<td>C</td>
<td>Cytosine or Cytidine</td>
</tr>
<tr>
<td>G</td>
<td>Guanine or Guanosine</td>
</tr>
<tr>
<td>T</td>
<td>Thymine or Thymidine</td>
</tr>
<tr>
<td>U</td>
<td>Uracil or Uridine</td>
</tr>
<tr>
<td>c</td>
<td>Cognate (to the template)</td>
</tr>
<tr>
<td>nc</td>
<td>Non-cognate or near-cognate</td>
</tr>
<tr>
<td>FA</td>
<td>Fusidic acid</td>
</tr>
</tbody>
</table>
Introduction

The central dogma of life

All living organisms that we know of store the blueprint of their design in DNA. DNA consists of two entwined nucleotide chains, and the information is encoded in the sequence of the nucleobases inside the double helix, which face each other and connect the two strands. The double strands make the DNA a suitable molecule for information storage, since it is stable over time and because the two strands hold the same information, should one be damaged. However, this also makes access of information slow and cumbersome, since the two strands must be separated for access to the nucleobases.

For quick access to the genetic information, the DNA sequence is copied to a single strand messenger RNA, through a process called transcription, which is performed by the enzyme RNA polymerase. The mRNA is also a chain of nucleotides that contains the genetic information, but is short-lived compared to DNA. Instead, mRNA offers easy access of the genetic information to the ribosomes that read the genetic code and translate it into proteins. One mRNA can be translated many times, so that each transcription of a gene in the DNA quickly can result in many proteins.

In the translation of mRNA, triplets of nucleosides, called codons, are matched with tRNA molecules that each carries an amino acid. The amino acids are linked to form a polypeptide, with a specific amino acid sequence that is encoded in the DNA. The polypeptides are folded into proteins that carry out or catalyse almost all processes in the living cell, as dictated by the instructions in the DNA.

This flow of information, from DNA to mRNA to proteins that execute the orders and also replicate the DNA at cell division, is called the central dogma of molecular biology [1] (Fig. 1). Essential to the central dogma is accuracy in the information transmissions; transcription by RNA polymerase and translation by the ribosome. If the transmission was not accurate, the high fidelity information storage would be meaningless since the functions would not be reliably performed, and even evolution would be impossible since traits would no longer be heritable, even if the genes were inherited.
Fortunately, these information transmission systems have evolved to be very accurate, with an average error frequency of around $10^{-5}$-$10^{-4}$ in transcription [2; 3; 4; 5] and $10^{-4}$ in translation [3; 6; 7]. A typical gene of around 1,000 nucleotides thus has a 90-99% probability of being transcribed without errors, and a probability of 97% to escape translation errors. However, it is not obvious how this high accuracy is achieved.

**The problematic accuracy**

In reaction kinetics of chemical reactions, there is a trade-off between speed and accuracy. Just like you will discover if you try to sort the laundry in a hurry, it is difficult for an enzyme to be fast and precise at the same time. Like matching the socks in the laundry analogy, an enzyme must cull all the substrates that it could possibly interact with, of which only one is correct. In transcription and translation, this is particularly difficult, since the different substrates (nucleotides or aa-tRNAs, respectively) the polymerase or ribosome must accept are correct or incorrect only depending on the template DNA or mRNA. This is analogous to matching black socks that are almost, but not quite, identical, so that they must be juxtaposed before you can tell if they match.

At the same time, the need for speed is also very high for transcription and translation. Bacterial cells must be able to respond quickly to changing conditions, or under constant conditions grow and multiply as quickly as possi-
ble to compete for resources, so the gene expression has evolved to be fast. Since transcription and translation produce chains of macromolecules, hundreds or thousands of elongation reactions are needed to give the final product, and therefore, each elongation reaction must be very fast.

So how do the cellular machineries accomplish the high accuracy and the high speed simultaneously? The initial selection of substrates, corresponding to the first juxtaposition of socks, is not accurate enough at very high speeds, so the two systems also have methods to check again whether the accepted substrate is correct. This second step of selection is called proofreading. Even though we know of the mechanisms of initial and proofreading selection, the quantitative details of their speed and accuracy are still unclear.

Dissertation overview

This thesis will attempt to give a quantitative description of the high accuracy and speed in transcription and translation, regarding how the speed is affected by queue formation on the template and for translation also by an antibiotic, the distribution of the transcriptional and translational accuracy, how much initial and proofreading selection contribute to the transcriptional accuracy, and the mechanistic preconditions for a second proofreading step in transcription. The thesis is based on the six scientific publications or manuscripts listed above, with the following content:

I. The general approach to describe accuracy theoretically is presented, and the DNA template dependent accuracy variation in transcription is calculated using a mean-time model based on thermodynamic properties of DNA and RNA.

II. The speed of transcription is described in stochastic simulations that include interactions between polymerases on the DNA. My contributions include running some of the simulations and working on the manuscript.

III. The accuracy variation in transcription is again calculated using a mean-time model of transcript elongation, but the model is here expanded and refined, and compared to experimental accuracy measurements.

IV. It is here shown how the transcript elongation model combined with experimental observations predicts a second proofreading step in transcription.

V. Translation speed in the presence of the antibiotic fusidic acid is described by stochastic simulations, with focus on the effect caused by queuing of ribosomes.
VI. The total accuracy of translation is measured experimentally. My contribution is a statistical analysis of the results, constructing a statistic model to predict the accuracy of a codon and thereby the translational accuracy variation.

Where I am the first author (I, III-VI), every part of the project and manuscript can be attributed to me (and in V, also to E. Taberman).
Methods

Theoretical biology

Theoretical or mathematical biology is the interdisciplinary field where biological systems are described by various kinds of mathematical models. The necessity of these quantitative models arises from the limitations of the cartoonish models sometimes used in medicine or biology to represent a molecular process, where, for instance, “molecule X meets molecule Y and Z happens”. To make this description useful, we need to quantify some of the aspects of this encounter. What are the concentrations of molecules X and Y? What is the probability that they meet? What is the strength of the interaction between them? What is the probability that they are in right configuration for reaction Z? What is the rate of reaction Z? Given these numbers, how often will Z happen – if it happens at all?

These quantifications make sense of the cartoonish model, and can be used to make quantitative predictions of how the molecules will behave in the test tube or in the living cell. Besides predictions of experimental results, the model can evaluate and develop the apparent conclusions from experiments, combine the results from many different experiments, and make suggestions for future investigations.

There are many kinds of mathematical models. The methods used in this thesis are models based on mean-time calculations, stochastic simulations and statistical modelling.

Kinetics and reaction rates

Reaction kinetics is the study of chemical reaction rates, and how they can be expressed in rate equations and reaction rate constants, which is necessary to predict the reaction rates under different conditions. The kinetics of a chemical reaction can be determined through experimental measurements of how fast the reaction product is formed at different reactant concentrations, or predicted by theory if the reaction mechanism is known.
The reaction rate equation is the mathematical relation between the average rate of the reaction and the average concentrations of reactants and a rate constant, which accounts for constants such as temperature and diffusivity. The order of the reaction is the sum of the exponents of the reactant concentrations in the rate equation, and determines the unit of the reaction rate constant. For a zeroth order reaction, the rate of the reaction is independent of the concentration of the reactants, and the unit of the rate constant is Ms$^{-1}$. For a first order reaction, the reaction rate has a linear dependence on the concentration of one reactant. The unit of the reaction rate constant is s$^{-1}$. Reactions have higher orders than one if they contain more than one reactant with linear dependences, or if the dependence of the reactant is not linear with an exponent of the reactant larger than one. The rate constant of a reaction of order $n$ is M$^{(n-1)}$s$^{-1}$. The reactions in this thesis are all modelled as single step reactions, with a single transition state, for which the stoichiometry coefficients of the reaction determine the reaction order with respect to the reactants.

An example of a second order single step reaction is the bimolecular reaction where molecules A and B react to form a molecule C.

$$A + B \xrightarrow{k[A][B]} C$$

While the reaction order with respect to substance A or substance B is one, while the overall reaction order is two. The reaction equation of this equation describes the rate of change of the reactants A and B in terms of the reaction rate constant $k$ of the reaction.

$$\frac{d[C]}{dt} = -\frac{d[A]}{dt} = -\frac{d[B]}{dt} = k[A][B]$$

One way to calculate the reaction rate constant is through the Eyring equation (a progeny of the Arrhenius equation) [8]. The Eyring equation is based on transition state theory, which states that a chemical reaction from state 1 to state 2 must overpass a transition state, where the reactants are in a high-energy complex called the activated complex. The reaction rate constant depends on Gibbs free energy of activation, $\Delta G^i$, of the activated complex. Figure 2 shows an energy potential diagram, with Gibbs standard free energy of the two stationary states, $\Delta G_1^0$ and $\Delta G_2^0$, and of the transition state, $\Delta G^t_{1,2}$. In the reaction going from state 1 to state 2, the activation energy is only the energy barrier of the transition state, but when going from state 2 to state 1, the free energy of activation also includes the free energy difference of the two states. The state energy notation includes “$\Delta$” to signify that the
free energy of the state is not the total free energy of the complex, but the free energy change relative to a common ground state.

Figure 2. Potential energy diagram showing Gibbs standard free energy of the two ground states 1 and 2 and the transition state 1,2.

The Eyring equation states that the rate of a reaction is the product of the pre-exponential factor $k_{pre}$ and the exponential function of the free energy of activation, divided by Boltzmann’s constant $k_B$ of $1.38 \cdot 10^{-23}$ JK$^{-1}$ and the temperature T. Boltzmann’s constant can also be expressed per mole, as in Fig. 2 above and Eq. 3 below, and then becomes the gas constant of 8.31 JMJ$^{-1}$K$^{-1}$.

$$k = k_{pre} e^{\frac{-\Delta G^\ddagger}{RT}}$$

(3)

The pre-exponential factor $k_{pre}$ must hence have same unit as the reaction rate constant, meaning that it also depends on the order of the reaction. It generally accounts for the frequency of collisions between the reactants.

Michaelis-Menten kinetics and the master equation – an example

The biochemical reaction systems described in this thesis typically follow Michaelis-Menten kinetics, the most commonly used model for single-substrate enzyme kinetics [9]. Such a system can in its simplest form be described by the scheme

$$E + S \xrightleftharpoons[k_d]{k_c} ES \xrightarrow{k_c} E + P$$

(4)
where $E$ is the enzyme, $S$ its substrate, $k_a$ the association reaction rate constant of forming the complex $ES$ and $k_d$ its dissociation rate constant, and $P$ the product formed and dissociated with a rate constant $k_c$. For Michaelis-Menten kinetics, the steady state rate of product formation can be expressed by the Michaelis-Menten equation [10]:

$$\frac{d[P]}{dt} = j = [E]_0 [S] k_a \frac{k_c}{k_c + k_d} = [E]_0 [S] \frac{k_{cat}}{[S] + K_m}$$

where $[S]$ is the concentration of free substrate and $[E]_0$ the total enzyme concentration. At low substrate concentrations, product formation is governed by the specificity constant $k_{cat}/K_m$ that is a measure of the catalytic efficiency of the system; how efficiently the substrate is turned into product by the enzyme [10]. Different $k_{cat}/K_m$ values for different substrates are at the heart of enzymatic accuracy, as will be developed below.

Master equations are a mathematical representation of how the probability of a system being in a specific state is evolving over time, when the system can inhabit a finite number of states and where the rate of change for each state can be described by a first-order differential equation. We now imagine that in our system, resembling the systems of transcript and peptide elongation investigated later in the thesis, the reactions take place on a large macromolecular complex such as a ribosome or polymerase. Hence, the enzyme concentration is one molecule where the reaction occurs, and the concentration of substrate in solution determines the reaction rate. In order to use a master equation to describe the system in Eq. 5, we can define the reaction scheme as moving between three states; the first state (state 1), where the substrate is free; the second (state 2), where complex has formed; and the third (state 3), where the product has formed. In our example of transcription or translation, this reaction scheme constitutes one elongation cycle, and the next elongation cycle pursue state 3. The system can only occupy one state at a time.

The reaction rate of going from state 1 to state 2 is the reaction rate of the substrate association with a rate constant $k_a$. Since the enzyme concentration is neglected, this reaction becomes a first-order reaction that depends only on the concentration of the substrate, and the reaction rate is $k_a [S] \text{Ms}^{-1}$, and $k_a$ has the unit $s^{-1}$. The other reactions, substrate dissociation with a rate constant $q_1$ and product formation with a rate constant $k_c$, are zeroth-order reactions that do not depend on the concentration of any of the reactants, only on being in the state where the reaction occurs. The reaction rates then becomes $q_1 \text{Ms}^{-1}$ and $k_c \text{Ms}^{-1}$, respectively.

The time-dependent probability $P$ of being in each state in a transcriptional or translational elongation cycle is described by its master equation:
In our example, the reaction is one of many reaction cycles, meaning that the next reaction cycle follows after product formation and the system is not stalled in state 3. The master equation of this type of system is the basis of both the mean-time models and the stochastic simulations.

Mean time models

If we integrate the master equation above over the total time, we will get a system of linear equations that describe the mean time $\tau$ that each state is inhabited. In our example system, resembling an elongation cycle on a ribosome or RNA polymerase, we have the boundary conditions that the system starts in state 1 and ends in state 3, after which the next elongation cycle follows. Applying these boundary conditions, the system can be solved for the mean times expressed in terms of the reaction rates:

$$\frac{dP_1}{dt} = -k_1 \cdot [S] \cdot P_1 + q_1 \cdot P_2$$

$$\frac{dP_2}{dt} = -(q_1 + k_c) \cdot P_2 + k_1 \cdot [S] \cdot P_1$$

$$\frac{dP_3}{dt} = k_c \cdot P_2$$

(6)

In our example, the reaction is one of many reaction cycles, meaning that the next reaction cycle follows after product formation and the system is not stalled in state 3. The master equation of this type of system is the basis of both the mean-time models and the stochastic simulations.

The transcription models in papers I-IV are based on mean time calculations of this kind. In the papers, one set of mean-time calculations are used to calculate the mean times of the states within the transcript elongation cycle, and another is used to calculate the mean times that the polymerase spends in each position on the DNA template.
Stochastic simulations

The mean time calculations only describe the arithmetic mean of the system, and cannot describe the noise in the data, or the effect of perturbations or bistability. This type of model is called deterministic, as it always produces the same result from given initial conditions. However, the steady state of the system is not always distributed about the deterministic expected value since the fluctuations in the system may have a net effect on the expected value; instead, many systems with complex behaviours are modelled using stochastic simulations.

Stochastic simulations attempt to imitate the system and all the events that occur during some period of time. Each event has a probability density function given by its reaction rate, and the time of the event or the number of times it has occurred is sampled using a random number. This way, there are an infinite number of outcomes of the stochastic model from the same initial conditions. In order to generalize the results from the stochastic model, large numbers of simulations are usually needed, to create a dataset for statistical analysis. Thus, stochastic simulations are in general considerably more time-consuming and compute-intensive than deterministic models.

Papers II and V are based on stochastic simulations. In both papers, a model is used to describe nascent chain elongation events when there are many RNA polymerases on the same template DNA or many ribosomes on the same mRNA. Traffic situations arise when the polymerases or ribosomes catch up with each other, on one hand forming queues that hinder the propagation of the elongation, or on the other hand reducing the stochastic movement to make the forward movement faster. The traffic situations of polymerases and ribosomes are hence the type of complex systems, where certain chance events have a big effect on the average elongation rates, that are difficult to describe by an average elongation time.

Stochastic simulations are performed using Monte Carlo methods, which is a comprehensive term for different methods to obtain a numerical solution to a problem, for example a differential equation that is difficult to solve analytically, by generating random numbers and observing the trajectories of the variables, and approximating the expected values by the empirical mean of the samples. The main tool used in stochastic modelling of biological reaction systems is the Gillespie algorithm [11].

In the Gillespie algorithm, the time until the next reaction is sampled from the probability density that any reaction occurs, given that the probability of each reaction \( i \) is  \( P(t_i) = e^{-\lambda t_i} \), where \( k_i \) is the reaction rate. The identity of
the reaction that occurred is then sampled from the relative probabilities of all reactions, \( \frac{k_i}{\sum k_i} \), by a second random number. The benefit of using the Gillespie algorithm instead of updating the system in fixed time-steps is that also the interaction between reactions is modelled, when the reactions are not independent, such as the collisions of ribosomes on mRNA.

**Statistical models**

Statistical modelling is entirely different from the models described above. Where mean-time analysis and Monte Carlo simulations are used to predict the time evolution of a system, assuming knowledge of the rules of its behaviour, statistical modelling applies inferential data analysis to identify the rules of the systems that have generated an existing data set. A statistical model is then a set of assumptions about the underlying mechanisms that created the empirical dataset, and by testing different models against the observed data, the model with the best fit is said to describe those mechanisms.

In paper VI, empirical data on the translational accuracy is investigated by an ANOVA (analysis of variance) [12]. This method aims to explain the variation of the observed variables, the deviation from their means, by the distribution of factors affecting the variables. If the variation of the variable is statistically related to a factor, this factor is said to have an effect on the variable. This statistical relation is determined by some statistical test, in ANOVA typically a t-test that calculates the probability that the observed effect is true given the size of the dataset and the degrees of freedom in the model. The test is “passed” if the null hypothesis, that the factor does not affect the variable, is rejected with a probability of being wrong that is smaller than the parameter \( \alpha \). The parameter \( \alpha \) is arbitrarily chosen, but a common standard is 0.05. The effect can be large or small, positive or negative, making the observed outcome smaller or larger than the expected value, represented by the mean, regardless of its statistical significance. Sometimes a factor is included in the model even if the factor itself is not statistically significant, if it still improves the fit of the model.

Here, the response variable is the observed \( \frac{k_{cat}}{K_m} \) for near-cognate dipeptide formation, and the effecting factors are the variation in tRNA, mismatch positions and type of mismatch for each accuracy measurement. Different models including different numbers of factors were tested to find the model with the highest degree of explained variance; that is, the molecular properties that affect the accuracy of translation.
Accuracy in theory

Enzymatic accuracy

The total enzymatic accuracy $A_{tot}$ is generally defined as the flow ($j$) of correct product ($P^c$) formation over the flow of incorrect product ($P^{nc}$) formation, from correct and incorrect substrates ($S^c$ and $S^{nc}$, respectively). For a system that is described by Michaelis-Menten kinetics like in Eq. 4, the flow of product formation can also be expressed in terms of the Michaelis-Menten parameter $k_{cat}/K_m$, and the free enzyme and substrate concentrations $[E]$ and $[S]$, respectively.

$$A_{tot} = \frac{j^c}{j^{nc}} = \left[ \frac{[E][S^c]}{[E][S^{nc}]} \right] \left( \frac{k_{cat}}{K_m} \right)^c$$

(8)

In a simple case with only one correct substrate and one incorrect substrate, the accuracy $A$ normalized with respect to concentrations (assuming that the correct and incorrect substrate concentrations are the same) is hence a measure of how much more efficiently the enzyme processes the correct substrate instead of the incorrect substrate. Another way to express the parameter $k_{cat}/K_m$ is as the product of the rate of product association $k_s$ and the probability of product formation [10]. The probability of product formation, in turn, can be expressed by the ratio of the rate constant of the forward reaction $k_c$ and the sum of the reaction rate constants, $k_d + k_c$, from the state 2 (where the complex ES has formed).

$$A = \left( \frac{k_{cat}^c}{K_m^c} \right) = \frac{k_c^c}{k_d^c + k_c^c} = \frac{k_c^c}{k_d^c + k_c^c} \left( \frac{k_d^{nc} + k_c^{nc}}{k_d^{nc} + k_c^{nc}} \right) = \frac{k_c^c}{k_d^{nc} + k_c^{nc}} \left( \frac{1 + \frac{k_d^{nc}}{k_c^{nc}}}{1 + \frac{k_d^{nc}}{k_c^{nc}}} \right)$$

(9)

This expression of the normalized accuracy elucidates the role of the ratio $k_d/k_c$. The accuracy stems from the differences in the balance between the forward reaction with the rate $k_c$, leading to product formation, and the
backward reaction with rate $k_d$, where the substrate is discarded, and accuracy is achieved only if this balance favours the correct substrate.

The equation can conveniently be re-written in terms of the maximum discrimination $d$ and the discard parameter $a$ [13]. We assume that the rate of association is the same for the two substrates, so that $k_a^c/k_a^{nc} = 1$ (as is the case with the systems in this thesis).

$$A = \frac{k_c^c}{(k_c^c + k_c^d)} = \frac{1 + k_d^{nc}}{k_c^c} \frac{k_d^{nc}}{1 + a \cdot d} \frac{1}{1 + a}$$

where

$$a = \frac{k_d^c}{k_c^c} \text{ and } d = \frac{k_d^{nc}}{k_c^{nc}} \frac{k_c^c}{k_c^d}$$

The benefit of this formulation is that it separates the part of the kinetics, the discard parameter $a$, that is common for all substrates, from the maximum discrimination $d$ that quantifies their differences. Furthermore, in some systems, the maximum discrimination $d$ is constant while $a$ can be tuned [14]. While $a$ varies between zero and infinity, the accuracy goes from 1 to $d$, so that $a$ determines how much of the available discrimination $d$ that is utilized. However, this relation also illuminates the kinetic trade-off between speed and accuracy; since $a$ is the ratio of the backward reaction rate constant to the forward reaction rate constant (if forward is towards product formation) it means that as $a$ increases, the flow of product formation decreases [13].

The total accuracy, when there are a number of substrates with different concentrations, is the sum of all flows leading to correct product formation over all flows producing an incorrect product.

$$A_{tot} = \frac{\sum_{i=1}^{N_c} [S_i^c] \left(\frac{k_{cat}^c}{K_m^c}\right)^c}{\sum_{j=1}^{N_{nc}} [S_j^{nc}] \left(\frac{k_{cat}^{nc}}{K_m^{nc}}\right)^c}$$

The total accuracy is approximately the inverse of the error frequency. According to the definition of the accuracy as a ratio of the flows of correct and incorrect product formation, the accuracy is the number of correct per incorrect reaction products, while the error frequency is the number of erroneous products per all products. Hence, the error frequency is $1/(1 + A_{tot})$. 
This is the general formulation of the accuracy in one step of product formation. However, when product formation is a process in many steps, the end product can also be monitored in several steps of fidelity control. The probability of end product formation is thus the compound probability of surviving all steps of substrate rejection, the product of the probabilities of product formation of all steps of selection. In transcription and translation, a substrate must survive both initial selection and at least two steps of proofreading selection (for transcription, see paper IV) before it is released as product.

**Initial selection**

Initial selection $I$ of substrates basically follows the general formulation of accuracy outlined above. A substrate forms a complex with the enzyme, and is either discarded through complex dissociation or proceeds to product formation. Accuracy arises in the system when the probabilities of the two outcomes are different for correct and incorrect substrates.

If the expression of the normalized accuracy (Eq. 10) is combined with the equations of reaction rate constants (Eq. 3), the accuracy can be expressed in terms of the free energy of formation of the states that the reactions connect. The $\Delta \Delta G$ of the reaction rate constant consists of the difference in free energy of the states and a reaction rate barrier, so that $k_c = e^{-\frac{(\Delta G_{I}^{0} - \Delta G_{I}^{0} + \Delta \Delta G_{I})}{k_B T}}$ and $k_d = e^{-\frac{(\Delta G_{I}^{0} - \Delta G_{I}^{0} + \Delta \Delta G_{I})}{k_B T}}$.

\[
A = \frac{1 + a \cdot d}{1 + a} = \frac{1 + k_d^{c} \cdot k_c^{nc}}{k_d^{c} \cdot k_c^{nc}} = \frac{k_d^{c} \cdot k_c^{nc}}{k_c^{c}} = \frac{1}{1 + \frac{k_d^{c}}{k_c^{c}}}
\]

where

\[
a = \frac{k_d^{c}}{k_c^{c}} = e^{-\frac{(\Delta G_{I}^{0} - \Delta G_{I}^{0} + \Delta \Delta G_{I}) - (\Delta G_{I}^{0} - \Delta G_{I}^{0} + \Delta \Delta G_{I})}{k_B T}} = e^{-\frac{(\Delta G_{I}^{0} - \Delta G_{I}^{0} + \Delta \Delta G_{I} - \Delta \Delta G_{I})}{k_B T}} = e^{-\frac{(\Delta G_{I}^{nc} - \Delta G_{I}^{nc} + \Delta \Delta G_{I})}{k_B T}}
\]

and

\[
d = \frac{k_d^{nc}}{k_c^{c}} = e^{-\frac{(\Delta G_{I}^{0} - \Delta G_{I}^{0} + \Delta \Delta G_{I}) - (\Delta G_{I}^{nc} - \Delta G_{I}^{nc} + \Delta \Delta G_{I})}{k_B T}} = e^{-\frac{(\Delta G_{I}^{nc} - \Delta G_{I}^{nc} + \Delta \Delta G_{I})}{k_B T}}
\]
In the calculation of the maximum discrimination \( d \) (Eq. 12), the transition state barriers cancel out if they are substrate independent, so that the only discriminating free energy is the difference in free energy between the states after product formation, \( \Delta G_3^0 \), and the initial state with free substrate, \( \Delta G_1^0 \), with correct \((c)\) and incorrect \((nc)\) substrate. The discard parameter \( a \), on the other hand, depends on the same difference in energy for the correct substrate and the difference in reaction rate barrier for the reactions of dissociation and the product formation. Here, one must remember that the difference in free energy between two adjacent states can never be negative in the calculation of the reaction rate constant that connects them; in that case, the difference is zero. This means, that if the free energy of the ground state \( 2 \) (with substrate in complex with the enzyme) was higher than the free energy of states \( 1 \) and \( 3 \), these free energy differences would be zero and their maximum discrimination would be \( 1 \). Hence, a prerequisite for enzymatic accuracy is a stable enzyme-substrate complex.

In 1957, Linus Pauling famously used a similar definition of accuracy (Eq. 12) when he concluded that with the known difference of the substrates, the maximum discrimination of isoleucyl-tRNA in the selection of isoleucine over valine in the aminoacylation would be about a factor 5 [15]. Compared to the error frequency of translation mentioned above of about \( 10^{-4} \), meaning an accuracy of about \( 10^4 \), this is a very low number, particularly as the discard parameter \( a \) would tune the accuracy even lower.

Unknown to Pauling, the selectivity of the reaction is amplified by the requirement of a correct substrate conformation of the valyl-tRNA synthetases at the amino acid activation step, as shown by Fersht and Dingwall 1979 [16], so that the selectivity against isoleucine is larger than \( 6 \cdot 10^4 \). This method where the enzyme increases the accuracy in initial selection by introducing selectivity in the reaction rate barriers, such as steric hindrance of a malformed enzyme-substrate interaction, is seen in both the transcript cleavage reaction and the phosphodiester bond formation in transcription [17; 18] (see papers III and IV).

The total accuracy can also be increased by additional steps of fidelity control that corrects errors that might have survived initial selection, called proofreading selection.

Proofreading selection

Kinetic proofreading was first proposed by John Hopfield in 1974 [19] and independently by Jacques Ninio in 1975 [20]. It has two pillars: an alterna-
tive exit allowing discarding of the substrate after product formation in initial selection; and a coupled reaction of co-product formation from a co-substrate, which offers a free energy driving force towards the alternative exit. The requirement is that the ratio of the concentrations of co-substrate to co-product is shifted far above equilibrium, making the alternative exit virtually irreversible [21]. The inflow substrates, \([\text{CoS}][\text{S}]\), is thus distinguished from the outflow \([\text{CoP}][\text{S}]\) from the alternative exit. A general scheme of kinetic proofreading is outlined in Eq. 13. The alternative exit is the dissociation of substrate and co-product with the rate constant \(q_d\). The formation of the co-product \(\text{CoP}\) from the co-substrate \(\text{CoS}\) marks the end of initial selection.

\[
E + \text{CoS} + S^{\text{c/nc}} \xrightarrow{k_d^{\text{c/nc}}} E \cdot \text{CoS} \cdot S^{\text{c/nc}} \xrightarrow{k_e^{\text{c/nc}}} E \cdot \text{CoP} \cdot S^{\text{c/nc}} \xrightarrow{k_d^{\text{c/nc}}} q_d^{\text{c/nc}} E + \text{CoP} + S^{\text{c/nc}}
\]

Discrimination between correct and incorrect substrates in kinetic proofreading appears when the reaction rate constants are different for correct and incorrect substrates, just like in initial selection, making the probability of product formation larger for correct substrates than for incorrect substrates. In Eq. 13, the reaction rate constants are labelled with the superscript \(c/nc\) to show that they can differ for different substrates. Not all of them need to discriminate between the substrates, but at least one reaction step must be discriminating.

The molecular mechanisms of proofreading can be more complicated than initial selection, and obviously, more reactions are involved. We can calculate the maximum discrimination \(d\) and the discard parameter \(a\) also for proofreading selection. Again, they are represented by a forward rate and a backward rate, which were \(k_c\) and \(k_d\), respectively, for initial selection. However, in proofreading, these rates are not the rate of a single reaction, but rather the total rates of the whole scheme to go forward by product formation or backward through the alternative exit. It is hence difficult to make the same kind of generalization about proofreading as was done for initial selection in Eq. 12. For instance, the parameters \(d\) and \(a\) might not be completely separate, as is the case in transcriptional proofreading and the simple scheme in Eq. 13.

The total accuracy \(A_{\text{tot}}\), including both initial selection \(I\) and proofreading selection \(F\), is expressed by the compound probability of product formation
through both steps of selection for different substrates, expressed in terms of \( k_{\text{cat}}/K_m \).

\[
A_{\text{tot}} = \frac{\sum_{i=1}^{N^c} [S_i^c] \left( \frac{k_{\text{cat}}}{K_m} \right)^c_{I,i} \left( \frac{k_{\text{cat}}}{K_m} \right)^c_{F,i}}{\sum_{j=1}^{N^{nc}} [S_j^{nc}] \left( \frac{k_{\text{cat}}}{K_m} \right)^{nc}_{I,j} \left( \frac{k_{\text{cat}}}{K_m} \right)^{nc}_{F,j}}
\]

(14)

The general theory of enzymatic accuracy, initial selection and kinetic proof-reading is described in further detail in paper I.
Transcription

Transcription is the process where the information stored in the double-stranded DNA is transcribed to a single-stranded RNA molecule by the enzyme RNA polymerase. The information in the DNA is stored in a sequence of nucleotides with four types of nucleobases: A (adenine), C (cytosine), G (guanine) and T (thymine). In the process of transcription, a few bases of the strands of the DNA are separated, forming a transcription bubble inside the RNA polymerase, making the bases of the strands exposed for base-pairing.

Only one of the DNA strands, called the leading strand, is copied. During copying, the nucleobases are paired with complementary nucleotides, that are linked together to form the RNA transcript. As the transcription bubble travels downstream along the DNA, for every elongation cycle a new DNA base becomes exposed and forms a base-pair interaction with an incoming nucleoside triphosphate (NTP), which is incorporated at the end of the growing transcript. When the RNA polymerase moves, the nascent transcript is ejected at the opposite side of the polymerase, and the DNA strands are re-annealed at the exit from the polymerase, keeping the transcription bubble size constant.

![Figure 3. The transcription bubble in the nucleotide addition cycle, including backtracking to state BACK where transcript cleavage occurs. The base pairs that mark the boundaries of the transcription bubble in state PRE are coloured pink in order to illustrate the movement of the polymerase.](image)
Figure 3 above shows the transcription bubble during one transcript elongation cycle. The active site marked by the red circle is where free nucleotides enter the RNA polymerase, form base-pairs with the template DNA, and then become phosphodiester bonded to the adjacent nucleotide in the transcript. In Escherichia coli (E. coli), our model organism, the transcription bubble consists of 12 denatured base pairs of DNA and a double-stranded RNA/DNA hybrid of 8-9 base pairs [22; 23]. Initial selection takes place in the state POST·NTP in the active site, where the associated nucleotide may dissociate with the rate constant $q_3$ or proceed to product formation by phosphodiester bond formation to the adjacent nucleotide with the rate constant $k_c$.

In the standard Watson-Crick base-pairing [24], T is paired with A, C is paired with G, G is paired with C and A is paired with a nucleotide called uracile (U), which in RNA replaces the T of DNA. In this way, the DNA base sequence is transmitted through transcription. However, this also means that the polymerase must intermittently accept all four substrates, of which only one at a time is correct, and the only way to distinguish the correct nucleotide is through the base interaction with the DNA template. At the same time, the difference in interaction free energy between correct and incorrect base-pairing is only around -0.9 kcal/mol [25; 26]. Following Pauling’s accuracy calculation, where $d = e^{-(\Delta \Delta G)/(k_B T)}$, this would give a maximum discrimination of only a factor 100, far below the observed accuracy.

Fortunately, the transcriptional accuracy is saved both by discriminating reaction rates in initial selection [18] and by transcriptional proofreading [27], also with discriminating reaction rates [17].

At phosphodiester bond formation, the RNA polymerase closes around the active site by folding of a flexible domain called the trigger loop [18; 28]. The trigger loop not only increases the catalytic efficiency of the active site, but also provides an induced-fit mechanism that discriminates against mismatched base pairs, enhancing the rate of product formation to greater benefit for correct substrates.

Transcriptional proofreading was first suggested in 1991, when it was shown that RNA polymerase is capable of endonucleolytic cleavage of the transcript [27]. It was later shown that this cleavage occurs in the backtracked state of the nucleotide elongation cycle, and that predominately mismatched nucleotides were cleaved off from the transcript [29]. A reaction scheme of the full elongation cycle is shown in Fig. 4.
Figure 4. The transcriptional proofreading reaction scheme. The co-product formation of transcriptional kinetic proofreading is indicated.

In the pre-translocated state PRE, the polymerase may translocate backwards to the state BACK. In BACK, transcript cleavage may occur, where at least a di-nucleotide is cleaved off from the transcript. This constitutes an alternative exit route for substrates that survived initial selection and became ester bonded with the nascent transcript (the last incorporated nucleotide in the transcript). The co-substrate of transcriptional proofreading is actually part of the substrate itself, the nucleoside triphosphate. At phosphodiester bond formation, a pyrophosphate is released as the co-product, leaving a nucleoside monophosphate in the transcript. In the cell, tri-phosphates are shifted far above their equilibrium with pyrophosphates and nucleoside monophosphates, providing the necessary shift in equilibrium for kinetic proofreading.

Transcript cleavage takes place in the active site, just like phosphodiester bond formation, and also requires folding of the trigger loop [18]. However, the catalysis is much more efficient with any of the two cleavage factors GreA and GreB present. When a Gre factor is associated, it substitutes the trigger loop in the active site, making it an efficient nuclease [30]. Similar to the phosphodiester bond formation, the catalysed cleavage also discriminates between correctly matched and mismatched base pairs, but this reaction is more efficient for mismatches, to a different degree for different mismatches, thereby contributing to the proofreading selection [17].

Models of transcription

Modelling of transcription has been a valuable complement to experimental studies. Yager and von Hippel published a pioneering model of transcription in 1991 [31] that described the free energy of the ground states of the transcript elongation cycle by the free energy of formation of the nucleic acids in
the transcription bubble. The free energy of the states in the transcript elongation cycle can be calculated as the sum of three parts; \( \Delta G_{DNA/DNA}^0 \), the free energy cost of opening up the double-stranded DNA; \( \Delta G_{RNA/DNA}^0 \), the free energy gained from forming the hydrogen bonds of the RNA/DNA hybrid; and \( \Delta G_{pol}^0 \), the free energy of the polymerase stabilizing the transcription bubble. This equation was combined with the Eyring equation (Eq. 3) to calculate the reaction rates of the transcript elongation cycle.

\[
\Delta G_{state}^0 = \Delta G_{DNA/DNA}^0 + \Delta G_{RNA/DNA}^0 + \Delta G_{pol}^0
\]  

(15)

The free energies of the states, where each state is a position in the DNA, were calculated from the nucleic acids terms, \( \Delta G_{DNA/DNA}^0 \) and \( \Delta G_{RNA/DNA}^0 \), estimated using nearest neighbour parameters. Since the reaction rate calculations always use the difference in free energy between two states, all terms that are constant will cancel out. The term \( \Delta G_{pol}^0 \) that describes the interactions between the RNA polymerase and the sugar backbones of the DNA and the RNA, was assumed to be the same for all transcription bubbles and could hence be neglected in the calculations. Bai et al. [32] combined the free energy method of Yager and von Hippel with a model of the internal states of the nucleotide addition cycle (backward and forward translocation, nucleotide association and transcript elongation) using experimental measurements of the reaction rates to estimate the free energy of the reaction barriers.

The free energy needed to break the hydrogen bonds between the bases in a base pair depends on the identity of the base pair, since C:G pairs are coupled by three hydrogen bonds and A:T/U pairs only by two, but also on the identity of the adjacent base pairs. Adjacent base pairs in double stranded nucleic acids stabilize or de-stabilize each other by base stacking, where the aromatic rings in adjacent nucleobases interact when they are in a favourable position with the rings lined up in parallel. The nearest neighbour model accurately predicts the free energy cost of separating double stranded DNA and RNA/DNA hybrids by calculating the total energy as a sum of the energy needed to melt the constituent pairs of base pairs. The free energy of melting the base pairs varies considerably between different pairs of base pairs [25; 33]. In the transcription bubble, the nearest neighbour model can hence be used to calculate \( \Delta G_{DNA/DNA}^0 \) and \( \Delta G_{RNA/DNA}^0 \). Figure 5 shows the nearest neighbour pairs of base pairs of the transcription bubble.
The modelled transcription speed

The speed of transcription has been measured many times for different genes and under varying conditions, but it has been difficult to capture the variation in speed with a high resolution. The variation in the transcription speed has been modelled by Bai *et al.* [32; 34] using the above mentioned model, and also by others with focus on transcriptional pausing [35; 36; 37]. These studies (especially by Bai *et al.* [32; 34]) can under some conditions predict the speed of transcription and how it varies with the template sequence reasonably well, but the resolution in speed measurements from the experiments is in the range of \( \sim 100 \) base pairs [32]. While Bai *et al.* model only the transcript elongation cycle, Klumpp & Hwa [36] and Klumpp [37] also simulate the interactions between polymerases on the same DNA under high transcription initiation conditions, where they might interfere with each other.

In paper II, we modelled the transcription speed with the transcript elongation cycle as described above and stochastic simulations of the polymerase traffic on the DNA of the ribosomal RNA operon (*rrn*) in *E. coli*. The transcription rate over this gene was measured by Condon *et al.* [38] using electron micrograph images of transcribing polymerases on the operon, giving an RNA polymerase density profile of the gene. Furthermore, in the same study they also measured the total time of transcription for a strain with four out of seven *rrn* operons deleted. Surprisingly, the average chain elongation rate increased by a factor 1.5, even though the rate of initiation per promoter
increased by a factor 1.6 [39]. The increasing polymerase traffic caused by the increased initiation rate would presumably also decrease the elongation rate, due to more polymerase collisions. Therefore, we wanted to create a model that could reproduce both the polymerase density profile and the elongation rate results.

The model in paper II was not able to reproduce the polymerase density profile very well, even though the same level of template dependent variation in transcription speed was observed. The best fit was obtained when the polymerases were allowed to unlimited backward and forward tracking, meaning that they could translocate away from the position where the end of the transcript would be in the active site. The density profile of the 5500 nt operon consisted of only 20 data points, so the resolution was low.

Notwithstanding, the elongation rate increased with increasing initiation rates for the model with unlimited backward and forward tracking. With this model, there would be a higher frequency of long dwell times at low initiation rates that would disappear at higher initiation rates where the tightly packed polymerases prevented each other form long backward and forward tracking.

The modelled transcription accuracy

The transcriptional accuracy is relatively little investigated, compared to, for example, the accuracy in translation or replication, as it has proven difficult to measure due to methodological problems. The transcriptional error frequency was estimated in the 80’s and 90’s to be around $10^{-5}$, using assays measuring the rate of incorporation of radiolabelled nucleotides [40; 3]. However, in those studies only the *in vitro* incorporation was measured, and only in a specific sequence context. Nevertheless, since the average transcriptional error frequency was estimated to be lower than that of translation, it was assumed that the translational error would mask the transcriptional error *in vivo*. This assumption, however, did not take into consideration all the sequence dependent variation in both the transcriptional and translation accuracy.

Later attempts to deduce the error frequencies from the very large collections of RNA-seq data in databases have been futile, since the frequency of method errors from reverse transcriptase and RNA sequencing is generally higher than the transcriptional error frequency. Only recently has the sequencing error frequency been reduced to the level of the presumed transcriptional error [4; 41].
Given the meagre experimental data, it is of particular interest to model the transcriptional accuracy. In the model by Bai et al. [32], mismatches are included, but only for nucleotide competition as part of the association reaction, not in accuracy calculations. Sahoo & Klumpp [42] have modelled the effect of interference of polymerases on the transcriptional proofreading, and also the role of backtracking for the proofreading [43], but do not quantitatively describe the total template dependent transcriptional accuracy.

We have modelled the template dependent transcriptional accuracy in papers I, III and IV in order to describe the accuracy variation in transcription, describe how it depends on factors such as nucleotide concentration, make predictions of error and accuracy hotspots and also suggest a mechanism for two-step proofreading (paper IV). All papers calculate the accuracy using the formalism described above and mean-time calculations together with the presented model of the transcript elongation cycle with reaction rates from the transcription bubble free energy estimated using nearest neighbour parameters.

Paper I is primarily an in-depth presentation of enzymatic accuracy calculations, with the transcriptional accuracy, including both initial selection and proofreading, as an example. In this paper, the basis of discrimination against errors is a tuned maximum discrimination parameter and the ratio of the discard parameters for the associated nucleotide when it is correct and the correct nucleotide. The results suggest that there is a very large accuracy variation in transcription.

In paper III, the model from paper I is refined, and most notably complemented with nearest neighbour parameters for mismatches, that are compared to the parameters for correct base pairing as the basis of discrimination. The available dataset of nearest neighbour parameters for mismatches in an RNA/DNA hybrid unfortunately only contains the mismatches A:A, C:C, G:C and T:U, albeit with all combinations of neighbours [26]. Therefore, the other mismatches were approximated with the available mismatch parameter for that template base, that is, the mismatches A:C and A:G were described by the A:A parameters, and so on.

We find further support in paper III to the result from paper I about the very large variation in transcriptional accuracy due to the surrounding template sequence, resulting from the large variation in free energy of the transcription bubbles, as calculated by the nearest neighbour parameters. Furthermore, it is also investigated in paper III how the accuracy co-varies with the GC-content, and it is found that the correlation is strong but that GC-content alone does not explain the variation.
The model in paper III is tested against an experimental dataset of 83 transcription errors from the transcriptome of Caenorhabditis elegans [41]. In the comparison, the template sequences around the observed errors were used to calculate the modelled accuracy for the error positions. A subset of the errors (approximately half of them) occurred at positions with the level of accuracy that was expected from the model, but half of the occurred in positions with a high predicted accuracy that is not expected to give rise to that number of errors. A smaller subset of the errors (approximately one fourth) occurred in positions with extremely low or very low accuracy that the model predicted to be very error-prone. The results indicate that the model based on the template sequence correctly describes part the accuracy variation, but that there are also other sources of accuracy variation that might not be template dependent.

Finally, in paper IV it is shown that when combining the standard model of transcription, used by us and others, the experimental observations that transcript cleavage always cleaves at least a di-nucleotide and that the nucleotide elongation following a mismatch is slow (which is also expected due to impaired base stacking stabilization from the mismatch), the logical conclusion is that there is a second step of proofreading in transcription (Fig. 6). This means that even when the polymerase backtracks only by one step (which is the maximal backtracking allowed in our transcriptional accuracy models), it cleaves two nucleotides in the transcript. The first round of proofreading catches an error in the last position of the transcript and the second round an error in the penultimate position, but the two nucleotides are cleaved off using the same cleavage mechanism.

Figure 6. The transcription bubbles of two-step proofreading, with $F_1$ denoting the first step and $F_2$ the second. The scrutinized nucleotide is coloured orange, and is at the last position of the RNA transcript in the first step of proofreading and at the penultimate position in the second step. Note that the second step of proofreading of the orange nucleotide is the first step of proofreading for the nucleotide that was incorporated in proofreading step one, coloured green.

Furthermore, the effect of returning to previous positions by consecutive cleavage, where they undergo a second round of proofreading, is in this paper quantified for the first time. However, the calculations show that both the
effect of the second step of proofreading and the effect of returning to positions are quite small. Nevertheless, the argument for the second step of proofreading is not based on the calculations but on experimental observations, and should hold even if the calculations would prove not to be robust. There is also the possibility that this mechanism is important, at least for bacteria, under other conditions than those in typical E. coli growth.
Translation

The product of transcription described above is a single-stranded RNA, containing the sequence information from a gene or operon in the DNA. There are several classes of RNA. Unlike DNA, RNA is not only a carrier of information, but can itself also have catalytic functions in the cell. The operon chosen as our example sequence above, \textit{rrn}, codes for one such RNA called ribosomal RNA (rRNA), but the largest family of RNAs is messenger RNA, mRNA. The mRNA only carries the information from the DNA to the ribosomes, where it is translated to proteins that perform the functions programmed in the DNA sequence.

The ribosome is a large molecular machine and consists of both proteins and rRNA. All self-sustaining cellular life must contain ribosomes, and the sequence of rRNA is one of the most conserved genetic sequences over the three domains of life, why it has been used to outline the phylogeny of the tree of life. Here, we will focus on the prokaryotic ribosome and more specifically on the ribosome in \textit{E. coli}, but it is worth noting that many of the features of the \textit{E. coli} ribosomes are shared also with eukaryotic ribosomes.

In \textit{E. coli}, the ribosome consists of the 16S, 23S and 5S rRNAs and 52 ribosomal proteins, separated in the large 50S subunit and the small 30S subunit. Since ribosomes maintain the protein turnover in the cell they are crucial for cellular growth, but since they are also costly to produce, and so their efficiency is also extremely important to obtain a high growth rate at a low cost. In fact, the average intercellular concentration of ribosomes correlate very well with the average growth rate of the cell and the growth rate can be defined in terms of the ribosome concentration and translation elongation rate [44; 45].

Translation is the process where an mRNA is de-coded into a sequence of amino acids, which later folds into a functional protein. Translation has four phases: first, there is initiation, where the 30S subunit recognizes and binds to the ribosome binding site on the mRNA, after which initiator tRNA and the 50S subunit binds to form the ternary initiation complex; second is the peptide chain elongation, where the ribosome is translocated along the mRNA and for each position adds a new amino acid to the growing peptide chain; third is the termination of translation, prompted by the reading of a
stop codon by a class 1 release factor instead of a tRNA, triggering the release of the synthesized amino acid chain, and fourth, the ribosome recycling, where the ribosome splits into its two subunits. Peptide elongation is the main focus of paper V and VI and requires a more detailed description.

In the interface between the two subunits in the ribosome, there are three binding sites for tRNA that also the mRNA passes through during peptide elongation: the A site, where the mRNA codon is recognized by an incoming tRNA; the P site, where the peptidyl-tRNA, the tRNA with the growing peptide, sits; and the E site, from which the tRNA and mRNA exits the ribosome. Peptide elongation transpires in an elongation cycle of four main steps. Starting with a complex with an empty A site and peptidyl-tRNA in the P site, an aminoacyl-tRNA in complex with the translation factor EF-Tu with GTP bound enters the A site. The tRNA contains an anticodon that base-pairs with the mRNA codon in the A site if it is complementary or similar to a complementary anticodon. Following GTP hydrolysis by EF-Tu and the EF-Tu·GDP dissociation, the tRNA undergoes a conformational change (tRNA accommodation) that directs its acceptor stem towards the peptidyl-tRNA. This leads to peptidyl transfer, where the peptide is transferred from the P-site tRNA and joined to the amino acid on the A-site tRNA by peptide bond formation. Finally, translocation occurs, where the mRNA and the tRNAs shift position so that the recently deacylated tRNA enters the E-site, the tRNA carrying the peptide enters the P site, and the A site is vacated. This reaction is catalysed by GTP hydrolysis by the factor EF-G that associates with GTP bound prior to translocation. With the release of EF-G·GDP, the next round can begin [46]. The steps of peptide elongation are presented in Fig. 7.
Figure 7. An overview of ribosomal structure and mRNA translation [46]. Image used with permission from Nature Publishing Group.

The codon is a triplet of nucleotides that forms base pair interactions with the anticodon in the tRNA. The codon-anticodon interaction allows for less specific base pairing with the third base of the codon, and the anticodon base in this position is not necessarily complementary to the codon base. The accuracy of translation ensures that the peptide is elongated with the amino acid coded for by the codon. This is achieved both through initial selection of tRNAs, where a non-cognate (or near-cognate, if there is some complementarity in the codon-anticodon interaction) tRNA in complex with EF-Tu can dissociate before GTP hydrolysis, and proofreading, where the tRNA can be rejected before peptidyl transfer.

The modelled translation speed in the presence of fusidic acid

The average rate of peptide elongation in *E. coli* during exponential growth in rich medium at 37°C is around 22 amino acids per second [47; 48], and seems to be limited by the peptidyl transfer reaction [49]. However, this rate is not only affected by the reactions in the elongation cycle, but also by the interactions between ribosomes on the same mRNA, where they may collide or form queues. The translation rate also varies between codons and contexts [50], contributing to ribosome queuing.
These ribosome interactions and their effect on the average translation rate have been studied by modelling of the peptide elongation of ribosomes on the same mRNA (similarly to the polymerase traffic models discussed above) [51]. We extended the stochastic translation model to also describe translation when inhibited by an antibiotic, using the recent model of translation with fusidic acid inhibition from *in vitro* experiments by Borg *et al.* [52].

In paper V, the rate of peptide elongation in the presence of the antibiotic fusidic acid (FA) and ribosome queue formation is investigated in stochastic simulations. FA is a slow inhibitor that binds to the EF-G·GTP complex on the ribosome and prevents translocation while it is bound, for around 8 s on average. This causes a translational pause corresponding to 100 elongation cycles, meaning that there is an obvious risk of queue formation when trailing ribosomes catch up with the stalled ribosome. Therefore, we wanted to establish how and under which conditions these queues form and dissolve, and how they affect the average rate of protein synthesis.

Queue formation depends strongly on the rate of translation initiation, since queues cannot form if there is only one ribosome on the mRNA. The simulations show that increasing the FA concentration lowers the initiation frequency, since FA inhibition prevents initiation by delaying the evacuation of the ribosome binding site, a prerequisite of the initiation of the next ribosome. This creates a limited range of FA concentrations where the queues contribute to the FA inhibition of the average elongation rate, with a maximum around the FA concentration of 2 μM. At the same time, the number of ribosomes on the mRNA is at its minimum. At higher FA concentrations, the queue effect decreases with the translation initiation frequency, and the number of ribosomes increase again when queues decrease. The total rate of ribosome synthesis that depends on both the translation elongation rate and the number of ribosomes on the mRNA will however always decrease when the FA concentration increases, due to the decreasing elongation rate.

**Statistical model of the translation accuracy**

The total accuracy of translation, including both initial selection and proof-reading, was measured for three *E. coli* tRNAs in a yet unpublished study [53]. In the experiments, the anticodons of the tRNAs were tested against all near-cognate mRNA codons, with only one mismatch in the codon-anticodon interaction, and the $k_{cat}/K_{m}$ of di-peptide formation was measured. Using Eq. 8, the $k_{cat}/K_{m}$ values of the correct codon-anticodon combinations were divided by these near-cognate $k_{cat}/K_{m}$ values to give the selection by
the tRNA against each specific mismatch. The results showed that there was very large accuracy variation between the mismatches.

In order to extrapolate this data set to the mRNA codons and the total translational accuracy of the transcriptome, a statistical model was constructed in paper VI. The data set was analysed by ANOVA, evaluating different models to explain the accuracy variation in the logarithm transformed data by the factors mismatch position (first, second or third in the codon), tRNA (out of the three), and type of mismatch (G:G, G:A, and so on). According to the analysis, the model that had the highest degree of explained variance, also when adjusted for the degrees of freedom, was a model where position 3, tRNA identity, and five out of the seven tested mismatches affected the response variable, the measured $k_{cat}/K_m$.

This model was then used to compute the selection against every mRNA codon for all tRNAs with near-cognate anticodons. The total accuracy per codon was calculated according to Eq. 15, summarizing the tRNAs leading to the correct peptide elongation in the numerator, and the incorrect elongation in the denominator, with all tRNAs weighted by their relative abundance in the cell [54]. With this estimated value of the translational accuracy, the error frequency was calculated. The total error frequency, with the codons weighted by their relative frequency in the transcriptome, was found to be $1.08 \times 10^{-4}$, very close to the experimental measurements.

The dataset was further used to analyze how the translational error frequency correlates with previous measurements of codon usage bias at different growth rates. Since the accuracy calculations are based on the intracellular aminoacyl-tRNA concentrations, which are known to correlate with the codon usage in the proteome, the error frequency was negatively correlated with the codon usage [54; 55]. However, this correlation was only significant for codons with a high error rate, indicating that the codon usage bias can also be a cellular strategy to simultaneously increase the translational accuracy and the translation speed.
Discussion

This thesis studies the speed and accuracy of transcription and translation, using a variety of models. The most important conclusion regards the trade-off between speed and accuracy, and also between speed and other qualities. Despite the intuition that accuracy would be the most important aspect, considering the costs and risks of erroneous products, the very large predicted accuracy span of transcription shows that accuracy can, in fact, be unnecessarily large at a big time-cost, implying that the most favourable part of the accuracy spectrum is somewhere in the middle. The papers that examine the traffic of polymerases on the same DNA or ribosomes on the same RNA show another risk with fast reactions; that the stochasticity in the systems increases, leading to an increased probability of queue formation, which could reduce the average rate of product synthesis. On the other hand, queue formation can also increase the average rate, as it limits not only forward but also backward movement, as is the case with polymerases. In conclusion, the considerations are complex, but there seem to be leeway in the systems for the cells to adjust to an optimal strategy for the given conditions.

It is a truism, however, that the output data of a model can never be better than the input data. The parameters used are always in line with the most reliable empirical data, but these will most likely be adjusted with future experiments, possibly using improved experimental methods. Therefore, one must consider the robustness of the models; if they would deteriorate with different input parameters.

The values of the parameters in the transcription models (papers I-IV) are sometimes quite arbitrarily chosen, based only on experimental observations of which reactions are “fast” and which are “slow”. The parameters are then tuned to make the total behaviour of the model fit other experimental observations, which can be achieved with many different parameter combinations. The transcriptional accuracy models based on nearest neighbour melting energies (papers I, III and IV) are actually quite sensitive to changes in the nearest neighbour parameters, which are very likely considering the sometimes large error span in the measurements. Furthermore, the model would definitely advance if the nearest neighbour parameters for the now missing mismatches were to be included. These improvements would mean that the specific accuracy of a sequence motif could change. For this reason, the con-
Conclusions presented are mostly based on properties of the distribution of the accuracy of all sequences, which would presumably not change very much.

Translation is better studied than transcription, and the measured reaction rate parameters in the translation model (paper V) are most likely more correct than the reported transcription reaction rates. The statistical model in paper VI is based on a limited set of data, with only three tRNAs, and would certainly improve with more measurements. The predictions of the model approximate the selectivity in the absence of experimental measurements, which will hopefully be available soon, but the effects of the parameters on the selectivity are nevertheless interesting to study. One interesting question for the future is how the effects change if the observed variable is not the total accuracy, but initial selection or proofreading selection.

There is also reason to believe that the knowledge of fusidic acid will increase, judging by the recent study of Borg et al. [52]. However, the limitation of that study, and the motivation of the modelling, is that only short mRNAs can be used in vitro, preventing the study of queue formation. Future investigations of in vivo queue formation with FA inhibition are also possible, for example by ribosome profiling experiments.

Transcription, however, is a more complicated subject. Compared to translation, transcription is relatively unexplored, although the knowledge of the reaction mechanisms in the transcript elongation cycle has increased remarkably from the studies of the trigger loop by the groups of Zenkin and Yuzenkova. This research has greatly contributed to the knowledge of the mechanisms of selectivity in the polymerase, and we will hopefully see a continued advance in this field.

The transcriptional accuracy has historically been difficult to study, since the error frequency form the method has been larger than the error frequency of transcription. With new and improved methods for reverse transcription and sequencing, there are good hopes for this field. In 2013, two studies on transcriptional accuracy with two new and very different approaches were published. The first, by Gout et al. [41], searched for true transcription errors in the C. elegans transcriptome using a technique that reduced the impact of the method errors with bar-coded RNA fragments that were reverse transcribed and sequenced multiple times, and only errors that were present in multiple reads were considered true transcription errors. The results were compared to the model predictions in paper III. The second paper, by Imashimizu et al. [4], aimed to measure the accuracy in each position in a gene using a high accuracy reverse transcriptase and sequencing. The results, however, showed that the transcription errors could not be separated from the method errors, but the approach was admirable, and can hopefully work better in the future.
Another new method with great promise that could be used to detect transcription errors is the NET-seq technique [56].

Regardless of experimental outlooks, the theoretical models presented in this thesis can contribute with new insights into the studied systems. Particularly, the models investigate the consequences of commonly held assumptions of how the systems work, and if these predictions would prove to be false, the modelling can point to the assumptions that need to be revised. Theoretical predictions have leverage on empirical evidence when the complexity of the experimental setup makes it difficult to know what is actually measured and to freely test all combinations of parameters. The inherent complexity of the biological systems studied here will always promote theoretical modelling for a deeper understanding.


I translationen läses mRNA av i enheter om tre nukleotider, ett så kallat kodon. Varje tRNA har ett antikodon med tre nukleotider som basparar med kodonet enligt samma regler som i transkriptionen, men med U istället för T. En typisk gen om ca 1000 nukleotider translateras alltså till en peptid om ca 333 aminosyror. Även translationen måste alltså vara både snabb och noggrann för att ge en komplett och korrekt peptid. Noggrannhetsmekanismerna i translationen påminner om de i transkriptionen genom att det finns både initialselektion och korrekturläsning, men de molekylära mekanismerna är annorlunda.

I denna avhandling används matematiska modeller för att ge en kvantitativ beskrivning av hastigheten och noggrannheten i transkription och translation under olika förutsättningar. Modellerna är av tre typer; numerisk integrering av masterekvationer för medeltidsberäkning, stokastisk modellering och statistisk modellering.

I artikel I förklaras den teoretiska bakgrunden till noggrannhetsberäkningar, och noggrannheten i transkriptionen används som ett exempel. Samma transkriptionsmodell, men med avsevärdare förbättringar, används i artikel III. Modellen inkluderar både initialselektion av nukleotider innan de inkorporeras i transkriptet, och möjlig rättning av felinkorporeringar i den påföljande elonergingencykeln. Beräkningarna baseras på uppskattningar av den fria energin i transkriptionsbubblan med hjälp av den tidigare uppmätta bindningsenergin mellan baserna i alla olika baspar, och hur de påverkas av intilliggande baspar. Transkriptionsbubblan har många möjliga bassekvenser vilket gör att variationen i transkriptionsbubblans fria energi för olika sekvenser är väldigt stor. Denna variation leder till att även noggrannheten har väldigt stor sekvensberoende variation. Den stora variationen gör att transkriptionsfel ofta kommer att uppstå i en liten andel av alla möjliga sekvenser som har hög sannolikhet för fel. Man har förut trott att
translationsfel alltid är vanligare än transkriptionsfel, men i positioner med dessa sekvenser är det möjligt att fel från transkriptionen är vanligare än fel från translationen.

Även artikel IV behandlar samma modell av transkriptionsnoggrannheten, men här argumenterar vi för att det finns rum för två steg av korrekturläsning, inte bara ett, i RNA-polymerasets korrekturläsningsmekanism. Eftersom klyvningen av transkriptet alltid klyver minst två baser förkastas också en felinkorporering i den näst sista positionen på transkriptet.

I artikel II modelleras transkriptionens elongeringshastighet och hur den påverkas av trängsel av RNA-polymeraser på samma DNA. Trängseln av RNA-polymeraser gör oftast att transkriptionen går långsammare, eftersom de hindrar varandra från att gå framåt i transkriptionscykeln, men under vissa förutsättningar gör trängselleffekten transkriptionen snabbare, eftersom det också hindrar polymeraset från att gå bakåt.

En liknande modell av trängsel används i artikel V, men nu för ribosomer som translaterar samma mRNA. I denna artikeln undersöks framför allt hur translationshastigheten påverkas antibiotikummet fusidinsyra, som binder till ribosomen och hindrar den från att gå framåt. Dessa långa pauser gör att efterföljande ribosomer hinner ifatt den fusidinsyrebunda ribosomen så att köer av ribosomer uppstår. Köerna gör att den genomsnittliga translationshastigheten minskar ännu mer än vad den gör bara på grund av fusidinsyran. Effekten uppstår dock bara i ett visst koncentrationsspann av fusidinsyra, eftersom höga fusidinsyrekoncentrationer gör att ribosomerna binder till mRNA långsammare, så att trängseln blir mindre.

Artikel VI bygger på variansanalys i ett dataset av uppmätt noggrannhet i läsningen av korrekta och inkorrekta kodon för tre tRNA. Dessa resultat används för att konstruera en modell som förutsäger noggrannheten i varje kodon för alla tRNA. Resultaten samstämer med den experimentellt uppmätta genomsnittliga felfrekvensen i translation. Modellen visar att den totala noggrannheten i translation till stor del beror på koncentrationen av det tRNA som vanligtvis läser ett kodon. Denna koncentration och användandet av olika kodon är i E. coli optimerade mot varandra för att ge maximal translationseffektivitet, men våra resultat visar att optimalingen också gäller noggrannheten.
Acknowledgements

Måns Ehrenberg – for wisdom, integrity, intuition, and patience

David Fange – for introducing me to everything new in the beginning, giving me a head start, and the support when I needed it

Anneli Borg, Onur Ercan, Doreen Heinrich, Ka-weng Ieong, Gabriele Indrisiunaite, Magnus Johansson, Michael Pavlov, Emmeli Taberman, Jingji Zhang, and other members in the Ehrenberg group – for interesting discussions, valuable comments, pure scientific talent and fun times in our group meetings. Anneli, for always being reliable. Ka-weng, for being ~1.3 m away whenever I had a stupid question. Emmeli, for being a fighter. Jingji, for making me laugh

Everyone in the program of Structure and Molecular Biology – for interesting discussions in the big group meetings. Suparna, for the always clever comments, for being supportive to everyone, and for sharing your personal experiences of science

Johan Elf, my co-supervisor – for always being welcoming to me

Prune, Petter, Anel, Irmeli, Cia and everyone else in the Elf group – for adopting me in your fellowship, and for all the pleasant lunch breaks.

Otto Berg – for brilliant comments on my licentiate thesis

Everyone in Harmony Heights – for lessons about barbershop, music and friendship, and for incredible times. Lena, Frida, Rebecca, Mari, and so many others!

Beatrice – for your warmth, intelligence and kindness, and a fantastic intuition about human nature

My big loving family – for love, support and a different perspective

Branka – for pain, pleasure, and purpose
Mats – for everything.
References


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1293

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-262903