University of Oslo
Department of Informatics

Detection of non-coding RNA genes by searching for transcription signals in intergenic regions

Gard Thomassen

27th April 2004

Preface

First of all I would like to thank my supervisor, Torbjørn Rognes, at the Bioinformatics group at the Rikshospitalet University Hospital (Oslo), where this study has been conducted. He has been a great supervisor and has supported me throughout the study.

I would like to thank my second supervisor, Knut Liestøl, at the Department of Informatics, University of Oslo. He has been helpful and guided me well through the writing of this master thesis.

A big "thank you" should also be given to everyone at the Bioinformatics group and the people at the Department of Molecular Biology at Rikshospitalet. Especially Knut Ivan Kristiansen for giving expert advice about ncRNAs, Einar A. Rødland for brilliant help with statistics and mathematics and Karin Lagesen for helping me with creating the multiple alignments. Josef Thingnes, my fellow master-student at the Bioinformatics group should also be thanked for good fellowship and discussions during this work.

Finally I would like to thank those of my family and friends who have helped me by reading and giving feedback on this report.

Summary

Background

Non-coding RNA (ncRNA) genes produce transcripts that exert their function without ever being a recipe for proteins. ncRNA gene sequences, unlike protein coding genes, do not have strong transcription signals. This study was conducted to investigate a special version of a previously tested and suggested method of detecting RNAs. This study is a part of a larger project where many such methods are to be combined to create a general purpose ncRNA finding program.

There are many possible ways to locate ncRNA. ncRNA genes have to be transcribed to produce ncRNA, and must therefore be surrounded by sequence regions that regulate transcription. Good candidates for new ncRNA genes would therefore be parts of intergenic sequences where transcription signals are present. Searching for transcription signals has previously been applied with success to find ncRNA genes in the bacteria *Escherichia coli* (*E.coli*) (Argaman et al., 2001) and yeast (Olivas et al., 1997). This strategy has later been applied once more to the *E.coli* genome with some success by Chen et al. (2002).

Methods

The method chosen in this study is a version of the above mentioned search for transcription signals. During this study 8 promoter consensus sequences have been suggested using data from earlier studies, the consensus sequences cover the promoter sequence of five of the seven known socalled σ (sigma) factors in *E.coli*. A novel promoter sequence score function has been created resulting in the implementation of a new promoter search algorithm. This promoter search has been combined with an implementation of a previously developed terminator search and scoring algorithm (Ermolaeva et al., 2000).

The output data has been analyzed by comparing the candidates to 52 verified and 1056 suggested ncRNAs. The number of located promoters has been compared with the estimated number of promoter hits

that would occur in a random sequence which maintains the basic features of the original inputstring. Some output data have also been multiple aligned with intergenic regions of genomes from bacteria closely related to *E.coli*.

Results

During this study at least three novel promoter consensus sequences for the *E.coli* polymerase have been suggested. A novel promoter sequence scoring algorithm has been implemented together with a previously used method (Ermolaeva et al., 2000) to locate ρ-independent (rho) terminators in *E.coli*. The implemented program has eight different promoter sequences it may search for by using user-defined thresholds.

A comparison has been made on the program's candidates against the suggested and verified ncRNAs. This comparison shows a very low hit ratio. Analysis has also been made to check the program's hit ratio towards the random case to verify the significance of the search criteria.

Using about 850 ncRNA candidates from the program, multiple alignments have been made to intergenic regions in related bacteria. This has resulted in a suggestion of 20 novel ncRNAs having a high level of conservation and high scores on promoter and terminator regions. Of the 20 suggested ncRNA candidates two were inside already known ncRNA genes, this leaves 18 novel ncRNA candidates.

At http://folk.uio.no/gardt/Hovedfag/index.html the search program developed in this study can be downloaded along with the BioJava packages needed. At this site one can also download the Java code, JavaDoc for the program and also the file containing the intergenic regions of *E.coli* that were used in this study.

Conclusion

This study concludes with a suggestion of 18 novel ncRNA candidates (see table 7.12 on page 81). The search algorithm and criteria used in this study represent a slightly new approach to the problem of detecting ncRNAs, specially by including searches for promoters recognized by other σ factor than the widely used σ^{70}. Analyses have shown that the program has a low hit ratio on already known or suggested ncRNAs, however other analyses have shown that the promoter consensus sequences used in this search are significant in promoter sequences to protein coding genes. The problems of detecting ncRNAs are rather connected to their weak transcription signals.

Of the 18 candidates, none have structural similarities with known ncRNA families. This is not very remarkable since if they had shown such similarities they would have been known already, consequently the 18 candidates represent novel families of ncRNAs or they are false. The answer to whether they are real ncRNA genes will be given when the 18 novel ncRNA candidates are tested in the laboratory.

As an independent program for ncRNA detection this program is not very suited as of today, but, as indicated above, when combined with other analyses it might represent a useful tool.

Contents

1 Aims of this study **1**
- 1.1 Detecting non-coding RNAs (ncRNAs) . . . 1
- 1.2 Why detect ncRNA ? . . . 1
- 1.3 Detecting ncRNAs . . . 2
  - 1.3.1 How to detect ncRNAs in this study . . . 2
- 1.4 Presenting the work . . . 3

2 Background **5**
- 2.1 Introduction to molecular biology . . . 5
  - 2.1.1 Historical overview . . . 5
  - 2.1.2 DNA . . . 6
  - 2.1.3 RNA . . . 8
  - 2.1.4 The central dogma of molecular biology . . . 11
- 2.2 Discovery of ncRNAs . . . 12
- 2.3 Transcription . . . 13
  - 2.3.1 Promoter regions . . . 13
  - 2.3.2 Terminators . . . 17
  - 2.3.3 Intrinsic termination . . . 17
  - 2.3.4 Rho-dependent termination . . . 17
- 2.4 Earlier studies on ncRNA . . . 20
  - 2.4.1 Rivas and Eddy, 2000 . . . 20
  - 2.4.2 Wassarman *et al.*, 2001 . . . 20
  - 2.4.3 Argaman *et al.*, 2001 . . . 20
  - 2.4.4 Carter *et al.*, 2001 . . . 21
  - 2.4.5 Rivas and Eddy, 2001 . . . 21
  - 2.4.6 Chen *et al.*, 2002 . . . 21
  - 2.4.7 Tjaden et al, 2002 . . . 22
  - 2.4.8 Hershberg et al, 2003 . . . 22
  - 2.4.9 Vogel et al, 2003 . . . 22
- 2.5 ncRNAs today . . . 24
- 2.6 Known functions of ncRNAs . . . 24
- 2.7 Estimates on the number of ncRNAs in genomes . . . 25
- 2.8 Verified ncRNAs today . . . 27

3 Search algorithm **29**
3.1 Why novel ncRNAs are hard to detect compared to protein coding genes . . . 29
3.2 Where to search for ncRNA in the *E.coli* genome 29
3.3 How many nucleotides make up an ncRNA 31
3.4 Structure of the search algorithm 31
3.4.1 Input 31
3.4.2 Preprocessing 32
3.4.3 Performing the search 32
3.5 Computing the final candidates 32
3.6 Output 32
3.6.1 Example of output files 33
3.7 Searching for promoters 35
3.7.1 Different genes are recognized by different σ factors 35
3.7.2 Definition of a promoter candidate 35
3.7.3 Searching for promoter candidates 36
3.8 Searching for terminators 37
3.8.1 Secondary structure folding 38
3.8.2 Definition of a terminator candidate 39
3.8.3 Searching for terminator candidates 39

4 Scoring system **41**
4.1 Scoring the promoter candidates 41
4.1.1 Creating consensus sequences 41
4.1.2 Consensus sequences are evolving 42
4.1.3 Logo plots 43
4.1.4 Scoring the possible promoter sequences 45
4.2 Scoring terminator candidates 47

5 Program complexity and runtime **49**
5.1 Running the program 49
5.2 Introduction to runtime and complexity calculations . 50
5.3 Complexity and runtime of the promoter search 50
5.4 Complexity and runtime of the terminator search . . . 51
5.5 Complexity and runtime of the final candidate computation 51
5.6 Total complexity and program runtime 52

6 Analysing the program output **55**
6.1 Setting the threshold values 55
6.2 Comparing to the random case 60
6.2.1 Estimating number of promoter hits in a random DNA sequence 60
6.3 Aligning the ncRNA candidates 63

7 Results **65**
7.1 A new promoter score function 66
7.2 Comparing results with the random case 75
7.3 Aligning candidates to intergenic regions in related bacterias 75
7.4 Comparing new and previous candidates 78
7.4.1 Testing program on verified ncRNA sequences . . 79
7.5 Suggested ncRNA candidates from this study 79

8 Discussion and conclusion **83**
8.1 Discussion 83
8.2 Conclusion 84

9 Improvements and further work **87**
9.1 Refining the promoter search 87
9.2 Refining the terminator search 87
9.3 Speeding up the program 88
9.4 Further work 89

A Definitions **91**

B Bacteria used in alignments **93**

Chapter 1

Aims of this study

1.1 Detecting non-coding RNAs (ncRNAs)

The overall purpose of the ncRNA project at the National Hospital is to investigate possible strategies to localize ncRNA genes in a genome sequence. When referring to ncRNAs in this study it means all types of RNA that are not coding for proteins, this means that suggested candidates might as well include novel rRNAs and tRNAs. The focus will at first be on finding ncRNA genes in bacterial genomes. Such genomes are generally well annotated, which will make the development and testing of the different strategies easier. The aim is, however, to develop these methods in ways that make them useful to other genomes as well. Several methods will be investigated and tested, the results will hopefully be used for further development and also used in a larger publicly available program combining several of the methods. One of the goals is that it should be possible for other scientists to use the developed programs to analyse their own sequences.

1.2 Why detect ncRNA ?

During the last few years the number of sequencing projects has increased dramatically. The data from these projects show that there are significantly fewer protein-coding genes in higher level organisms than expected. At the same time the number of known ncRNA genes has increased. The existence of such genes will probably give us a deeper understanding of the seemingly propotionally low genomical complexity in higher-level organisms compared to low-level organisms.

As Storz (2002) put it: “There may be ncRNAs lurking behind many an unexplained phenomenon”, there are lots of questions that remain to be answered, and a possible solution lies with the ncRNAs.

1.3 Detecting ncRNAs

For detection of ncRNA in bacteria several methods have been suggested and some have been applied previously:

- *Primary structure alignment* - novel ncRNAs could be detected by searching for known ncRNA sequences in the genome of bacteria related to the bacteria where the known ncRNA sequence was extracted from.
- *Secondary structure alignment* - novel ncRNAs could be detected by locating sequences with a similar secondary structure as already known ncRNAs.
- *Transcription signals* - every gene has transcription signals that signals to the transcription mechanism in the cell that this actual part of the DNA is a gene. By locating such signals one might detect the existence of novel ncRNAs.
- *Comparative genomics* - the basic idea behind using this methods is that in closely related bacteria a novel ncRNA should be possible to detect by looking for short sequences with a high level of conservation in the bacterias intergenic DNA.
- *cDNA cloning and microarrays* - the idea here is to reverse transcript small parts (oligonucleotides) from known ncRNAs into cDNA. This is followed by hybridizing the cDNA with single stranded DNA. Points of hybridation might be areas containing ncRNAs.
- *Neural networks and machine learning programs* - the idea behind this approach is to implement a program that looks at known ncRNAs and computes what novel ncRNAs might look like, and tries to find them.

The first three methods look at the actual DNA string of nucelotides, these approaches are typically computational. Method number four and five are laboratory based methods, while number five is a newer and more experimental computational approach than the first three.

1.3.1 How to detect ncRNAs in this study

This study covers one part of the entire ncRNA project. The main idea behind this part of the project is to use transcription signals as a search criteria for novel ncRNAs. This study will define search criteria for transcription signals in *Escherichia coli (E.coli)* and develop and implement

search and score algorithms. The result of this study should be a program able to function alone and in conjunction with other programs developed during the ncRNA project.
Detecting transcription signals could be divided into three main parts.

- *Defining search criterias* - define the criterias of transcription signals, that is, what does a promoter and a terminator candidate look like.
- *Search for transcription signals* - independently search the input data for promoters and termiantors.
- *Compute final candidates* - try to match a promoter and a terminator lying in the same intergenic region, having a distance between them lying in the range of the length of typical ncRNAs.

Besides implementing and creating the actual program, an important aim of this study is to investigate this approach according to efficiency and accuracy. This is important as the program is meant to become part of a larger program, and knowing how one part works at an early stage is important for further work and development of the larger program.

1.4 Presenting the work

One part of this study has been to present the ongoing work. Because of the close relationship between my fellow master student Josef Thingnes' work and mine, these presentations have been done together. We have had a short presentation of our respective work at the Bioinformatics Forum for Young Scientists at Vatnahalsen in March 2003. Our work was also presented on a poster at the annual Norwegian conference on Biochemistry at Hafjell in January 2004, the poster can be found at: (http://www.cmbn.no/rognes/vm2004_gard_jo.pdf). In addition to this we held a short presentation of our work at the lecture-session on Bioinformatics held at The Institute of Informatics at the University of Oslo. This presentation can be found at : (http://www.ifi.uio.no/forskning/grupper/bioinf/Teaching/gardogjosef_files/frame.htm).

Chapter 2

Background

This chapter covers a short introduction to molecular biology. A reader familiar with molecular biology can jump to section 2.4 on page 20 without loosing essential information.

2.1 Introduction to molecular biology

All organisms, except viruses, are made up of cells. To construct and maintain a living organism some kind of recipe is needed. This recipe is located in something we call a genome. In humans it consists of the nuclear genome and the mithocondrial genome, which both lie in the cells. The human nuclear genome consists, in most cases, of 46 chromosomes. Chromosomes contain DNA molecules, and parts of the DNA molecules make up what we call genes. The parts of the chromosome that are genes are those parts which in fact make up the recipe of how the organism is supposed to be. The chromosome contains many genes, but nevertheless, the coding part of the genes of a human do not occupy more than about 1.5% of the basepairs in the human DNA (Mattick, 2003).

2.1.1 Historical overview

An excellent review can be found in Klug and Cummings (1996), on which the following is based.

The corner stones of biology were laid down by early researchers in the years from 1600 to 1850. During these years scientists made huge progress in the field of understanding the biological building blocks of living organisms. Many of these discoveries made the revolutionary discoveries by Darwin and Mendel possible. Their theories about heritage and natural selection opened up paths for further research in the years to

come. During the early part of the twentieth century the chromosomes were discovered, and the scientists understood that the chromosomes in some way kept hereditary information.

Until 1944 it was not known which of the chemical components in the chromosome that made up the genetical material. It was known that the chromosomes contained both nucleic acid and proteins, thus both were possible candidates. Finally in 1944 scientists could state that it was the nucleic acid (called DNA), that was the information database of heritage. The question that now emerged was : "How can DNA be an information database for the complete process of life"?

The general idea was that it must have something to do with the molecular structure of DNA, since DNA has a very systematical, but also complicated structure. A big leap forward towards an answer was made in 1953 when Watson and Crick published their hypothesis about the double-helix structure of DNA. The assumption that the function of the DNA molecule would be a lot easier to understand after the general structure of the molecule had been discovered, turned out to be correct.

In the late nineteen fifties several scientists (Meselson, Stahl, Taylor, Woods and more) published evidence on how the molecular structure of DNA is in detail, and also on how replication works. These discoveries stated that the Watson and Crick hypothesis of the double helix were correct, and by revealing how DNA replication works, the function of DNA was more or less understood. A quick overview of the history of this research can be found in table 2.1.

Finally today, approximately 50 years later, the sequencing of the human DNA structure has been completed. This sequencing project has been named "The Human Genome Project".

The Human Genome Project was initiated more than 10 years ago, with the purpose of mapping the human genome. The project more or less turned out to become a contest between two research groups. One lead by The National Institute of Health in Bethesda, USA, while the other group was the privately held US-based company Celera. The initial sequencing of the human genome was published in Nature in February 2001 (Lander et al., 2001).

The number of human genes was previous to the project set to be approximately 100 000. During The Human Genome Project it has been discovered that this number lies surprisingly lower, somewhere around 30 to 35.000 (Lander et al., 2001).

2.1.2 DNA

DNA is an abbreviation for Deoxyribonucleic Acid. DNA is the molecular storage for genetic information, and is in eukaryotic organisms localized in the nucleus of the cell. The molecular structure of DNA is a so-called

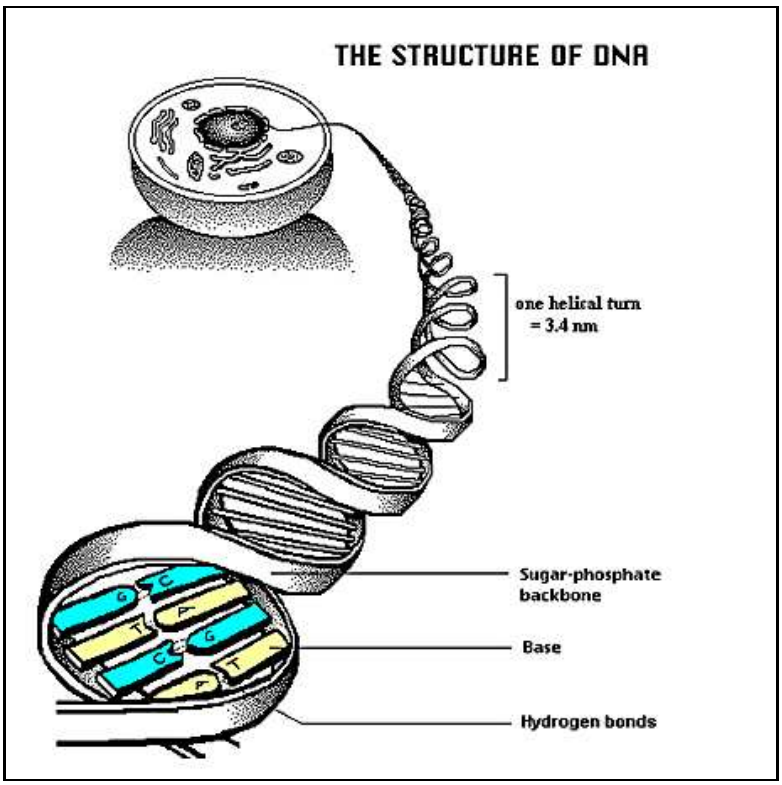

Figure 2.1: DNA double-helix, picture courtesy of the National Health Museum (http://www.accessexellence.org).

Year	Discovery
1865	Genes are particulate factors
1903	Chromosomes are hereditary units
1910	Genes lie in chromosomes
1913	Chromosomes contain linear arrays of genes
1944	DNA is the genetic material
1945	A gene codes for a protein
1953	DNA is a double helix
1961	Genetic code is triplet
1977	DNA can be sequenced
1997	Genomes can be sequenced

Table 2.1: A brief history of genetics (Lewin, 2000).

double helix (see figure 2.1). DNA consists of the four bases Adenine (A), Guanine (G), Thymine (T), Cytosine (C), along with the socalled DNA backbone which consists of alternating series of pentose (sugar) and phosphate residues. Along the strings it is the different permutations of the bases that make up the particular sequence of the DNA, the bases are connected to another base on the other DNA string through hydrogen bindings. The backbone of the two strings in the DNA molecule consists of the sugars connected to each other by phosphate bindings, the bases are again connected to the sugar. Each string of the DNA molecule is called a "strand".

Every base (A,T,C,G) has its complementary base. A is complementary to T and vice versa, the same goes for C and G. The two strands in the DNA molecule are arranged in such a way that a base nearly always has its complementary base on its position on the other strand. If there is an exception, it is likely to originate in some damage to the DNA. Two complementary bases are called a base pair. The strands are therefore said to be complementary. The stability of the DNA is to a great extent caused by the hydrogen-bindings between the complementary bases on the two strands. Between C's and G's there is a triple hydrogen bond, while between A's and T's there is only a double hydrogen bond, thus A-T bindings are weaker than C-G bindings.

DNA serves only as an informational database for the organism. The DNA itself does not perform any special tasks. Data is collected from the DNA according to the organisms' needs. This data collection is called transcription, because the information on the DNA is transcribed into an RNA molecule. (More on RNA in section 2.1.3.)

The cells in an organism have a limited lifetime. New cells are made by dividing existing cells. During such a cell division the genome must be duplicated. This happens through a complicated process called DNA replication (see figure 2.2). After the DNA replication the double-helix is transformed into two identical double-helices. Each of the two new DNA molecules now has one strand each from the original DNA molecule.

2.1.3 RNA

RNA is an abbreviation of Ribonucleic Acid. RNA exists in many different forms, and is therefore annotated by different prefixes: mRNA (messenger RNA), tRNA (transfer RNA), rRNA (ribosomal RNA) and many other. The prefix is given according to the specific task the RNA has. mRNA is the kind of RNA that has been the object for most research done on RNA until today. This is because mRNA is the only RNA that is translated into protein, and proteins has for a long time been regarded as the most important molecules in organisms. mRNA and proteins has therefore been the main target for researchers looking for causes and remedies to

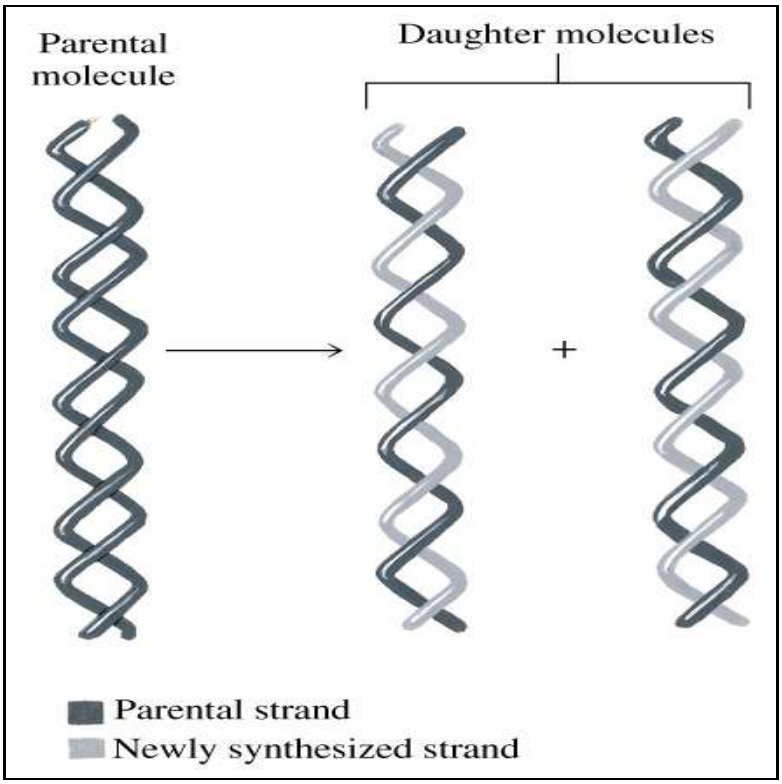

Figure 2.2: Replication of DNA, picture courtesy of the National Health Museum (http://www.accessexcellence.org).

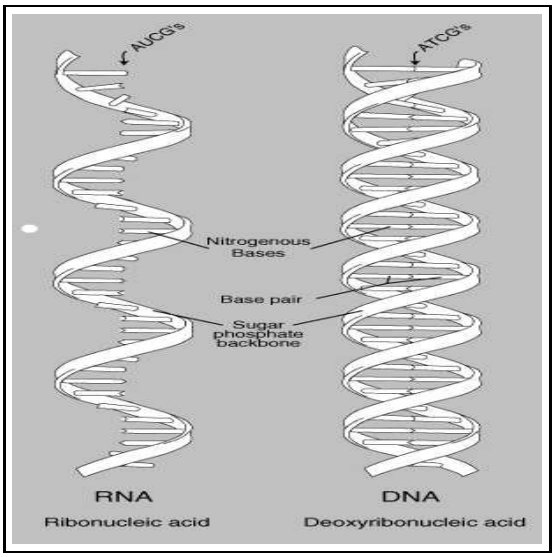

Figure 2.3: RNA and DNA, picture courtesy of the National Health Museum (http://www.accessexcellence.org).

different diseases.

The common abbreviation ncRNA means non-coding RNA, i.e. RNA that is not translated into protein. ncRNA is sometimes also named fRNA (functional RNA), that is because all ncRNA actually perform some task in the cell without being translated into a protein first.

RNA is built basically in the same way as DNA (see figure 2.3). However there are three major differences:

- RNA contains ribose, not deoksyribose.
- RNA contains the base Uracil (U), instead of Thymine (T). U is identical to T apart from lacking one methyl-group.
- RNA is normally single-stranded. (Does not have a complementary strand with hydrogen bindings between the complementary bases.)

While the bases in DNA are bound together and protected in the double-helix, the bases in the RNA lie in the open, and are unprotected to the surrounding environment. DNA is therefore very stable and can stand a lot of “beating” before it dissolves, while RNA is unstable, and usually has a limited lifetime before it dissolves.

Since the making of proteins from mRNA has been considered the most important function originating in the genome, the mRNA coding genes of the genomes also have been the main research object in the mapping of genomes until now. ncRNAs have also avoided much research because their coding areas on the DNA are very hard to detect.

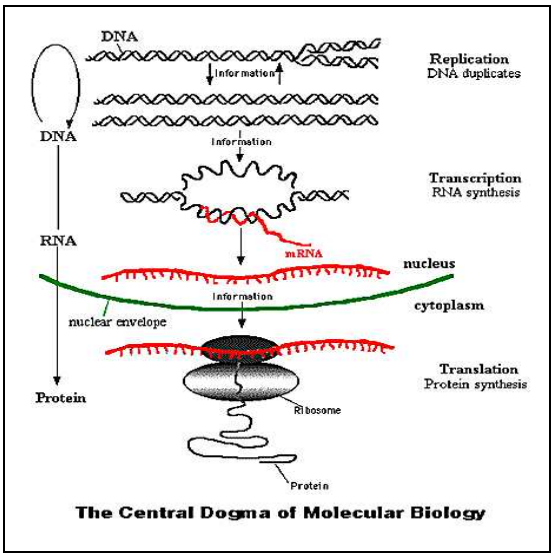

Figure 2.4: The central dogma of molecular biology, picture courtesy of the National Health Museum (http://www.accessexcellence.org).

Due to the big difference between the previously estimated number and the new approximation of protein coding genes, more and more focus is directed towards the genes that code for ncRNAs, and more specific towards localizing these genes on the DNA, and to find their actual funtion in the organism. More about this in section 2.5.

2.1.4 The central dogma of molecular biology

To create a protein, the instructions in the DNA have to reach the ribosomes in the cytoplasm of the cell. This happens by making an accurate copy of the gene that contains the recipe for this protein. This process is called transcription. The copy is produced by an enzyme called RNA polymerase. This enzyme enters the DNA molecule where the gene begins and transcribes the gene into mRNA. Then the mRNA is transported out of the cell nucleus to the ribosomes where the mRNA is translated into proteins. This translation is based upon triples of bases in the mRNA that code for amino acids, and the amino acids bind together and make up protein. When the complete mRNA has been translated, the amino acids make up a protein. Such a production of proteins is called a protein synthesis. This making of proteins is referred to as "The central dogma of molecular biology" (see figure 2.4).

Historically the phrase "gene" was only used for sequences of the DNA molecule that code for mRNA, but in this thesis the term "gene" will be used about a sequence of the DNA molecule that code for some

type of RNA. It will be stated whether it is an mRNA or an ncRNA gene we are talking about when it is not given by the context.

2.2 Discovery of ncRNAs

In the end of March, 2004, 182 completely sequenced genomes have been published (including 4 chromosomes); most of these are from bacteria (Bernal et al., 2001; Kyrpides, 1999). Furthermore, more than 900 genomes are in the process of being sequenced . A majority of these genomes contains varying amounts of DNA which have yet no known function. These regions are often referred to as "junk" DNA, and they make the task of locating the protein coding areas of the genome especially challenging. Such analysis has shown that the number of genes in an organism is not necessarily linked to the size of the genome. The human genome is believed to have only seven times as many genes as yeast, although the human genome is about 200 times longer.

As more genomes become available, there has also been an increase in the number of known RNAs which do not participate in protein synthesis. When the *Escherichia coli (E.coli)* genome was published, it was found to contain at least 4290 protein coding genes (Blattner et al., 1997). Several genes coding for stable functional RNAs have since been found in the *E.coli* genome, including 86 tRNAs, 22rRNAs and 10 other RNAs (Wassarman et al., 1999). The RNA products of these genes seem to be involved in RNA processing, mRNA stability, translation, protein stability and secretion. Several such genes have also been seen in other organisms; Argaman et al. (2001) identified 16 new such genes in yeast, and Olivas et al. (1997) have found 201 candidate ncRNA genes in mice.

Many ncRNAs have been discovered by accident while searching for protein coding genes. One of the reasons for this is that until recently RNAs were generally thought to have no important functions other than in the protein synthesis. This is reflected in the amount of research done in this area, both on the biological and bioinformatical side. The problem of protein coding genes has been studied thoroughly within bioinformatics, resulting in such programs as GENSCAN (Burge and Karlin, 1997), and VEIL (Henderson et al., 1997), which can be used to locate protein coding genes. No such programs do yet exist for locating general ncRNA genes. The emphasis has until now been upon developing speciality tools, such as tRNAscan-SE (Lowe and Eddy, 1999) developed to locate tRNA genes. Tools for locating possible ncRNA genes in genomes could help finding more of these genes, and thus lead to a greater understanding of how they work.

2.3 Transcription

All types of RNA are transcribed from DNA (except in some viruses). This process is in *E.coli* catalyzed by an enzyme named RNA polymerase (see figure 2.5). There are some 7000 such RNA polymerase molecules in every cell in *E.coli*, where about 2000 to 5000 of these are synthesizing RNA at any one time, the number depending on the growth conditions (Lewin, 2000). The RNA polymerase molecule is capable of recognizing the region upstream of a gene, and it binds itself to the DNA molecule at these regions called "promoter regions". This binding of the RNA polymerase to the DNA is called "initialisation of the transcription". At this time the RNA polymerase covers the DNA from about the -55 to the +22 position, relative to transcription starting point. The next step for the RNA polymerase is to break the two strands of the DNA molecule away from each other, to create a transcription bubble. The transcription is now ready to begin, and the RNA polymerase releases its contact with the -55 to the -35 region.

During the transcription RNA is created by adding one nucleotide at the time, building the complementary strand of the "template strand" of the DNA, this transcription happens at a speed of $\approx$40 nucleotides per second at 37 ℃ according to Lewin (2000). When the transcripted RNA chain is about 15 to 20 nuclotides long, the RNA polymerase releases more of its connection to the DNA, and also releases its socalled σ (sigma) unit. The RNA polymerase now consists of what is called the "core enzyme", and has a connection with the DNA of about 30-40 bp.

The resulting new RNA strand is an exact copy of the "coding strand" of the DNA, except for the exchange of U's for the T's. This motion of the RNA polymerase along the DNA strands growing an RNA chain is called "elongation". The hybrid of the template strand DNA and the newly produced RNA is thought to be about 3 to 9 basepairs long (Lewin, 2000). As the RNA polymerase moves along the DNA, the template strand loosens the new RNA and rebinds to the coding strand. Finally, when the RNA is about to become complete, the RNA polymerase must detect this to end the transcription. There are many ways in which the RNA polymerase can be told to end the transcription. This ending of the elongation is enforced by "terminators". When the elongation has terminated, the transcription bubble disassociates and the final part of the DNA rebinds, so that the DNA again is a stable helix.

2.3.1 Promoter regions

The mission of a promoter region is to make the RNA polymerase start the transcription of the DNA molecule at exactly the right position. This process is called "template recognition" i.e. the RNA polymerase recog-

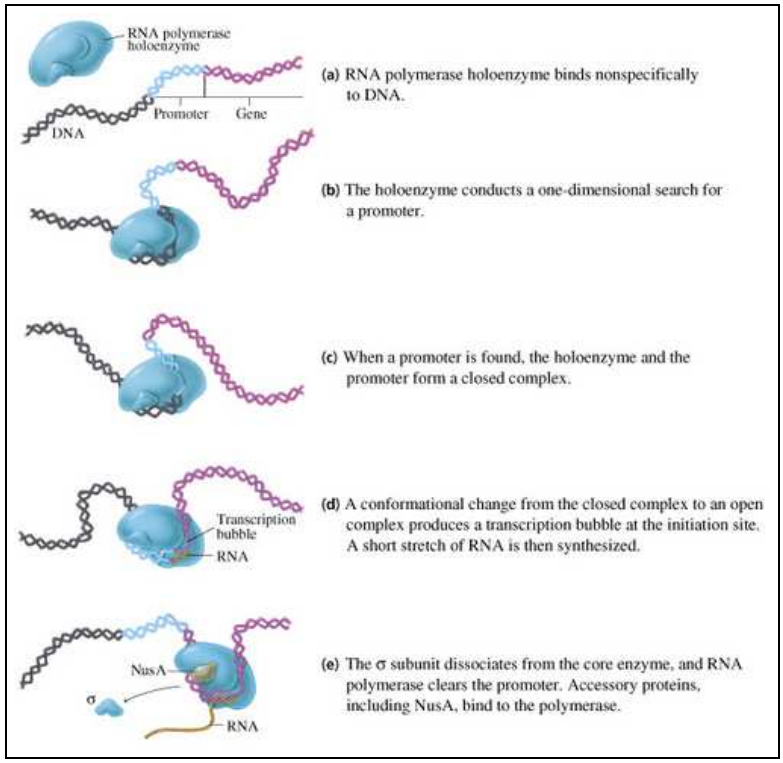

Figure 2.5: Transcription initiation i *E.coli*, the function of the σ sub unit can easily be seen in this illustration, picture courtesy of Principles of Biochemistry (http://cwx.prenhall.com/horton/medialib/media_portfolio/index.html).

nizes the upstream region of the gene that is to be transcribed. The actual region where the RNA-polyerase binds to the DNA molecule is called “binding site”.

RNA polymerase is made up of five different subunits. It is the σ subunit that enables the template recognition. The σ sub-unit locates the correct binding site by searching for certain conserved regions which are located upstream of the gene, namely the promoter regions. The search is based on complementarity; if a piece of the DNA string is very close to or an exact complementary match to the sigma unit, then the sigma unit can bind to this DNA region, and a template is recognized. In bacteria there are different σ sub-units, they all do the same work, but they recognize slightly different promoters, and are triggered by the phase of the cell (Lewin, 2000). This means that if a cell changes from growth- to vegetative-phase it might automaticly use a different σ sub-unit to recognize promoters. This makes gene expression maximized for the new condition.

Promoter regions vary from organism to organism, but there are similarities between organisms of the same family. In *E.coli* there are seven main types of promoter regions, and a consensus region has been found for some of these sequences (see table 2.2). The sigma factor used in the major part of the transcriptions in *E.coli* is the σ^{70}, (Kundu et al., 1997) (this is when the cell is in growth phase). During searches for genes in *E.coli* the consensus promoter region derived from genes recognized by this σ^{70} unit has been widely used, and the better the alignment with the consensus promoter region is, the more likely there is a gene following this promoter.

The length of the template strand associated with the RNA polymerase in *E.coli* is about 60 base pairs. To make up a single region inside these 60 base pairs that is significant, i.e. a sequence not likely to occur often at random, 12 consequtive more or less conserved basepairs are needed. Surprisingly, there have been found no extensive conservation over these 60 basepairs, not even of a region of 12 basepairs needed to establish sufficient significance (Lewin, 2000). Today much of the sequence of the binding site is considered to be irrelevant because of the lack of conservation, but there are some short stretches of the promoter region that show significant conservation, and these small stretches seem critical for the function of the promoter region.

To describe the different promoter regions they have been aligned according to which σ subunit that recognizes it, and the result that is a maximized homology sequence is called the “consensus sequence” of this promoter. Most promoters are therefore described through their class’ consensus regions, and the distance between them. To keep the two conserved regions from each other, the one nearest the gene is named “-10 region” and the one furthest apart “-35 region”, because of

Subunit	Phase	-35 Sequence	Separation	-10 Sequence
σ^{70}	general	TTGACA	16-18 bp	TATAAT
σ^{54}	nitrogen	CTGGNA	6 bp	TTGCA
$\sigma^{38}(\sigma^{S})$	Stationary	not known	not known	not known
σ^{32}	heat shock	CCCTTGAA	13-15 bp	CCCGATNT
$\sigma^{28}(\sigma^{F})$	flagellar	CTAAA	15 bp	GCCGATAA
$\sigma^{24}(\sigma^{E})$	heat shock	not known	not known	not known
$\sigma^{19}(\sigma^{fecl})$	iron transp.	not known	not known	not known

Table 2.2: The σ-factors and their consensus sequences in *E.coli* (Lewin, 2000; EcoCyc, 2004; Ussery, 1999)

their distances from the start codon on the DNA. In the promoter region of *E.coli* some other small conserved regions have been recognized, but these conserved regions are so insignificant that they also might occur by chance, therefore they are not very useful to gene searcing (Lewin, 2000).

An optimal promoter to be recognized by the most used *E.coli* RNA polymerase σ subunit, σ^{70}, has a six nucleotide sequence (TTGACA) starting 35 nucleotides upstream from the gene and another sequence seven nucleotides upstream (TATAAT), with a separation of 17 basepairs between the two conserved regions (Lewin, 2000). In table 2.2 there is a list of *E.coli* sigma factors and their promoter consensus sequences. For three of the sigma factors data on their corresponding consensus sequences have not been found. Two of these consensus sequences (σ^{38} and σ^{24}) have been established during this study, while insufficient data about σ^{28} has excluded it from this search (see table 4.3 on page 44).

A promoter region is needed for a gene to become expressed, mutations in the promoter regions might therefore affect the capability of a gene to become expressed. The most usual result is a downmutation, that is, a mutation that makes the promoter sequence less like the consensus region, the opposite, which is less likely to happen, is an upmutation.

A promoter region seems to occure upstream of every protein-coding gene in *E.coli*, and the same RNA polymerase that transcribes protein coding genes also transcribes the known ncRNA genes. From this follows that if there are promoter regions in the intergenic regions of *E.coli* it might indicate a binding site for RNA polymerase. If the promoter region really is a transcription site, and it does not code for an mRNA, as it might do, it is very likely that it codes for ncRNA. In any of the two latter cases it would in any case be a discovery of a new gene.

2.3.2 Terminators

To terminate the elongation so that the newly produced RNA can escape from the DNA and the DNA can rebind its strands, terminators are needed.

Termination takes place when the RNA polymerase meets a terminator sequence, stops adding nucleotides to the product and dissociates completely from the DNA template. The order of the last two events is unknown. There are two known main types of transcription terminators in *E.coli* (Lewin, 2000).

The first and main type of transcription termination in *E.coli* is intrinsic termination. This termination is "hard coded" on the template strand, i.e. the RNA polymerase is made to stop its elongation according to the actual nuclotide sequence on the DNA string. Intrinsic termination is often called ρ (rho) independent termination.

The outer elements that can cause termination cases are called termination factors. Termination factors are proteins that, of a yet undiscovered reason, at the right time forces the elongation to stop. The protein having this role in most known protein-dependent terminations is the protein ρ, therefore the names of the two termiantion types are ρ dependent and ρ independent termination.

2.3.3 Intrinsic termination

By far, most known terminations of elongation in *E.coli* are intrinsic (Lewin, 2000). Intrinsic termination is dependent upon the formation of a special structure of the RNA named "hairpin", also called "stem and loop" (see figure 2.6). This happens when the RNA transcript contains an "inverted palindrome" (also called "inverted repeat"), that makes it possible for the RNA to fold up against itself and make a hairpin. In addition to the hairpin structure of the RNA it is followed by a U rich region (see figure 3.1 on page 37). This makes the binding to the complementary A on the template strand very weak. What is thought to be happening is that the RNA polymerase first stops and lingers a bit when the hairpin is produced, the hairpin structure weakens DNA-RNA binding, and when the Uracil rich region follows the RNA polymerase can not hold on to the template strand, consequently the RNA and the template strand breaks away from each other. The transcription has now been terminated.

2.3.4 Rho-dependent termination

Rho dependent termination takes place when the termination of a transcription is dependent of a protein named ρ. However there has to be

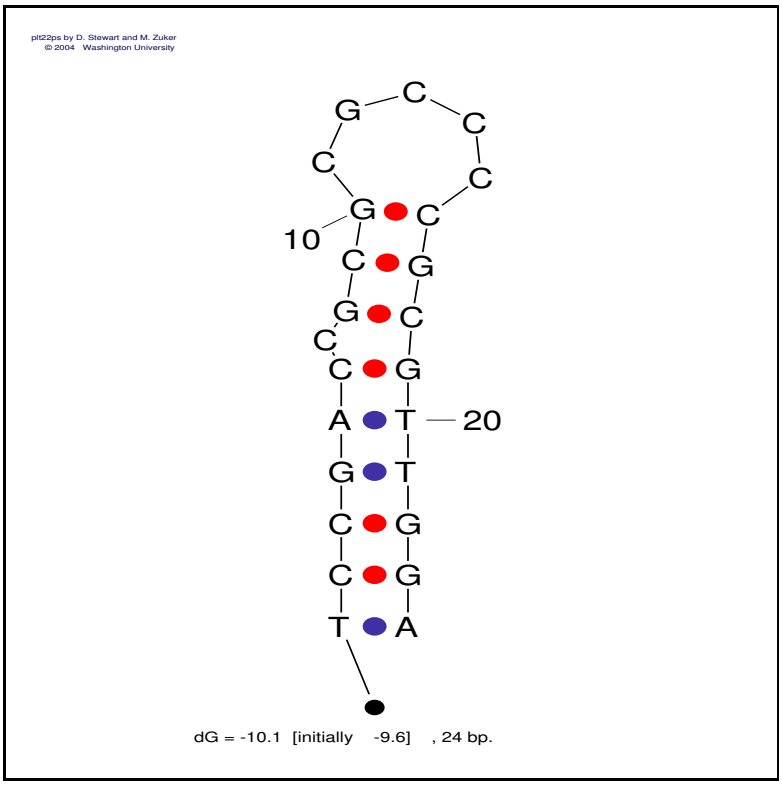

Figure 2.6:]
Hairpin construction (secondary structure), with one gap on left stem (the fourth C from the bottom), picture created using Mfold (Zuker, 2003).

certain special features present on the template strand to allow the termination factor to act. The sequence required for the ρ dependent termination is 50-90 bases long, and lies upstream of the actual termination site. The common feature of this sequence is that the transcripted RNA is rich in C residues and poor in G residues. A general rule of the efficiency of a ρ-dependent terminator is that it increases with the length of the C-rich/G-poor sequence (Lewin, 2000).

The ρ protein makes the transcription stop just like the hairpin structure does. The ρ protein connects to the newly produced RNA string and at the "right time" it catches up with the RNA polymerase running along the DNA strand. When the "right time" actually occurs is decided by termination signals on the template strand. When the RNA polymerase reaches a termination signal it usually lingers a bit, and that is the time when the ρ protein catches up and makes the transcription terminate. ρ is a "helicase", which means that it actively breaks base pairs, in this case between the template and transcript, resulting in termination of transcription. The terminaton signal that makes the RNA polymerase slow down so that the ρ protein catches up is the C-rich/G-poor region required for the ρ-termination to take place (Lewin, 2000).

2.4 Earlier studies on ncRNA

In this section some of the most important and recent work done on ncRNA in *E.coli* will be presented. It will cover both searching for ncRNA and also finding characteristics of ncRNAs. There has been some work on ncRNAs in other bacterias but *E.coli* has been the main target genome, as it also is in this project.

2.4.1 Rivas and Eddy, 2000

Rivas and Eddy (2000) wanted to locate ncRNAs by searching for sequences that create significant secondary structures, in accordance with suggestions of this strategy in the literature. They published a paper named "Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs". Their conclusion was basically what the title says. Namely that a distinct, stable secondary structure is important in most noncoding RNAs, but the secondary structures of the ncRNAs are not sufficiently different from the predicted stability of a random sequence. This conclusion makes this search strategy unusable as a general stand-alone gene-finding approach.

2.4.2 Wassarman *et al.*, 2001

Wassarman et al. (2001) published a paper named "Identification of novel small RNAs using comparative genomics and microarrays". Their strategy was to look for high conservation of small RNAs among closely related bacterial species combined with analysis of transcripts detected by high-density oligonucleotide probe arrays. The search was applied to the *E.coli* genome. They reported on the existence of 23 new RNA species, and 17 of the 23 they found are likely to be novel functional small RNAs.

2.4.3 Argaman *et al.*, 2001

In June 2001, a paper named "Novel small RNA-encoding genes in the intergenic regions of *E.coli*" was published by Argaman et al. (2001). They described a computational strategy for locating ncRNAs by looking for transcription signals and genomic features of known ncRNAs. Their search was very restrictive and it lead to the prediction of 24 putative ncRNA-encoding genes, of which 23 where tested experimentally. They reported the discovery of 14 genes encoding novel small RNAs in *E.coli*.

Their search was a two phase search. First they searched the intergenic regions of *E.coli* looking for the DNA promoter sequence recog-

nized by the major polymerase sigma factor, σ^{70}, and ρ-independent terminators. Intergenic sequences with 50 to 400 base-pairs between a promoter and a terminator region were then aligned with intergenic regions of other bacterias. Those regions from the *E.coli* genome where they could find significant conservation with other intergenic regions from other bacterias were to become the final candidates.

2.4.4 Carter *et al.*, 2001

A machine learning process using neural networks and support vector machines was the strategy behind the creation of an ncRNA locating program developed in this study. *E.coli* was the bacterium used for the development of the program, but the program is also applicable on other bacterial and archaeal genomes. Jackknife testing has shown that the program seems to be fairly accurate, especially while combining these predictions with parameters such as known RNA sequence motifs and the calculated free energy of folding. The program is publicly available, and has located hundreds of candidates, of which very few are experimentally tested. The success rate of this program has by others been shown to be lower than for other approaches for locating novel ncRNA genes, Chen et al. (2002). The findings of (Chen et al., 2002) might originate in the fact that neural networks tend to make a too well fit with the training set, and the corresponding Jackknife test to include the training set data.

2.4.5 Rivas and Eddy, 2001

Their article describes the development of a program called QRNA, the purpose of the program was to locate ncRNAs, Rivas et al. (2001). The algorithms behind this program are based upon comparative sequence analysis. The main idea was to look at the difference between the conserved regions of a coding and a structural (noncoding) RNA. The programs use three different algorithms and describe the different RNAs by using stochastic context free grammars. Tests run by Rivas and Eddy suggest that this QRNA program detects noncoding RNA genes with a fair degree of reliability. The program has suggested several hundred candidates for ncRNAs. However the number of false positives seems to be higher than initially suggested, (Chen et al, 2002).

2.4.6 Chen *et al.*, 2002

Chen et al used basically the same search strategy as (Argaman et al., 2001), namely to search for a σ^{70} promoter within a short distance of a ρ-independant terminator. However, their search criterias were

far from as strict as the ones used by (Argaman et al., 2001). The search of Chen et al resulted in the initial identification of 227 candidate sRNAs. Of these 227 candidates 44 were found to be potential novel non-translatable sRNA genes i.e. ncRNAs. These 44 were found by subtracting the ones of the 227 that were found to be some other type of RNA.

2.4.7 Tjaden et al, 2002

The work of Tjaden et al. (2002) was done to identify the transcriptome of *E.coli*. The transcriptome is all the parts of the DNA that are transcripbed. Oligonucleotide probe arrays were used in the search and several interesting discoveries were made. For a researcher interested in ncRNAs the most interesting discoveries included the discovery of 317 novel transcripts, ranging in length from 50 to 400 bp, with unknown functions, and furthermore the suggestion that several of these transcripts are ncRNAs.

2.4.8 Hershberg et al, 2003

The paper published by Hershberg et al. (2003), was on a survey of 55 known ncRNAs in *E.coli*. The main goal of the survey was to identify common charateristics, and much of their data originated in results from the work published in the papers mentioned above, especially the 55 verified ncRNAs of *E.coli*. Finding characteristics is a good help in understanding this unique gene family and also a great aid for improving the methods of predictions and identification of other ncRNAs in different genomes. According to the approach of this study their most interesting findings were that all known ncRNAs in *E.coli* are located in intergenic regions, most ncRNAs are in intergenic regions ranging from 300 to 900 bp, usually only one ncRNA per intergenic regions and the distribution of ncRNA genes between the leading and lagging strands are about the same. Intergenic regions shorter than 300 bp seem too short for ncRNAs and the intergenic regions longer than 900 bp are usually dominated by repetitive sequences in *E.coli*.

2.4.9 Vogel et al, 2003

By exploiting cDNA cloning techniques Vogel et al. (2003) have managed to locate novel ncRNAs (sRNAs) that have not been predicted earlier. Their work has brought the current number of verified ncRNAs in *E.coli* up to 62. Their research showed that some ncRNAs were encoded from independent genes, while others were processed from mRNA leaders or trailers. This indicates parallel transcriptional output of mRNA and

ncRNA in bacteria. The characterization of ncRNAs analyzed in their work suggested that the definition of an ncRNA is more complex than previously assumed. In the paper Vogel et al. (2003) presents novel distinct ncRNA species, and they also report on their expression patterns, metabolic stability and precise genomic location. Their most important findings are the suggestion of the parallel transcriptional output, the new characteristics of the novel ncRNAs and also their documentation of the successfull use of experimental RNomics.

2.5 ncRNAs today

rRNA and tRNA have together with the mRNA ruled the world of RNAs since the discoveries of the ribozymes (rRNA) in 1981, and also lead to the "RNA world" hypothesis for the origin of proteins, the building blocks of life (Riddihough, 2002)(see figure 2.4 on page 11).

To clearify the explosion of the interest in and knowledge about ncRNAs, the *E.coli* genome is a example. The *E.coli* genome is perhaps the best annotated genome in the world, and over a period of about 30 years (1970-2000) no more than 10 entirely new ncRNA genes had been discovered in *E.coli* (Argaman et al., 2001). The major breakthrough came when Argaman et al. (2001) published a paper where they reported on the discovery of 14 novel ncRNAs.

The discoveries of the functional RNAs have opened up "The other RNA world" to scientists. The interest in the other RNA world has in part been fueled by two related discoveries: the identifiction of large numbers of very small RNAs of approx. 22 nucleotides in length, called microRNAs (miRNAs), in such diverse organisms as *Caenorhabiditis elegans* (a small intestinal worm) and humans, and these molecules' very important function in the process of targeting and destroying homologous mRNA, viral RNA and other RNAs (Riddihough, 2002).

2.6 Known functions of ncRNAs

The roles of the different ncRNAs vary as much as their structures and range from the purely structural to the purely regulatory (Riddihough, 2002). Today, the full importance of the ncRNAs is not understood due to the lack of data on both functions (for many ncRNAs the functions are still unknown) and the extent of the "RNome", the RNA equivalent of the proteome. ("Proteome" is the common name for all proteins in a cell.)

ncRNAs are known to affect many different processes in the cell including plasmid replication, phage development, bacterial virulence and developmental control (Hershberg et al., 2003). Evidence of ncRNAs playing a role in RNA processing and modification also been provided (Storz, 2002). Wassarman et al. (1999) have published a review on small ncRNAs in which different ncRNAs and their functions (some unknown) were listed, and Szymanski and Barciszewski (2002) have released a table with function-classifications of non-protein coding RNA transcripts. In the article by (Wassarman et al., 1999) the name sRNA is used instead of ncRNA. sRNA is a usual name for ncRNAs in bacteria, and is also used by Hershberg et al. (2003). The "s" means small, and originates in the gene length that ranges from 50 to 400 nucleotides.

During the work of Argaman et al. (2001), they found that several of their detected ncRNAs had a significant increase of the expression level during phase transition, specially upon entry into stationary phase. This suggests that ncRNAs play an important role in integrating cellular responses to changing environments, i.e. fine tuning of gene expressions during phase transition (Argaman et al., 2001). This means that many of the ncRNAs found by Argaman et al. (2001) are suggested to be of importance to the bacterial physiology.

The vague knowledge of ncRNAs can easily be described by the fact that for 42 of the 62 discovered ncRNAs, the functions are still unknown (Hershberg et al., 2003).

2.7 Estimates on the number of ncRNAs in genomes

The first ncRNAs were discovered in the 1960s, they were discovered because of their high level of expression. However the number of discovered ncRNAs was very low for many years and the vast numbers that appear to be encoded by a genome were still hidden (Storz, 2002). Some estimates of the number of ncRNAs in *E.coli* range from 50 - 200 and in *C.elegans* from hundreds to thousands (Storz, 2002). Others estimate the number of ncRNAs to be 200 or more in *E.coli*, but still accounting for no more than 5% of the total number of genes and about 0.2 % of the transcriptional output (Mattick, 2003). A recent compilation of the result of the work of others holds more than 1000 candidates to ncRNA genes in *E.coli* (Hershberg et al., 2003). Thus there is little doubt that there exists many unverified ncRNAs and that there might be many more not even suggested as candiates yet. In a review on ncRNA genes by Eddy (2001) there are mentioned different opinions on the number of ncRNAs in *E.coli* ranging from 50 to 370. As there are 62 verified ncRNAs in *E.coli* already these estimates might have to be raised to fit better with the more than 1000 suggested candidates.

However, it is still believed that there are less ncRNA coding genes than coding genes in bacteria, and the main reason why may be expressed like this: ”Even though RNA has lots of good attributes there are many fewer ncRNAs than proteins, this is because of the superiority the proteins get via the robustness and versatility of the polypeptides of the protein, compared to the polynucleotides of the ncRNAs” (Riddihough, 2002).

In higher organisms the estimates of the number of ncRNAs relatively to the number of genes in the genome is much higher than for bacteria. It is estimated that about 98% of the transcriptional output from the genome is non-protein coding RNA in eukaryotes, this includes introns and transcripts from non-protein coding genes (Mattick, 2003).

These non-protein coding genes account for 50-75% of all transcription in higher eukaryotes, (Szymanski and Barciszewski, 2002). When we remove the tRNAs, rRNAs and introns from this bulk of non-protein coding DNA sequences, there are still a very large number of transcribed nucleotides left. If many of these transcribed regions that do not code for proteins actually are some type of ncRNA it would resolve a part of the discrepancy between the estimates of mammalian gene numbers based on genome sequence analysis (30 - 40.000) and cDNA cluster analysis (65 - 70.000) by indicating a whole new set of genes that do not code for proteins (Mattick, 2003).

2.8 Verified ncRNAs today

Below is a table with description of the 62 verified ncRNAs in *E.coli.*

sRNA gene	Adjacent genes	Strand[a]	Length	3' end position
tpk11	*dnsK/dnaJ*	→ → →	370	-
tp2	*pdhR/aceE*	→ ← →	120	-
t44	*map/rpsB*	← → →	135	189847
C0067	*yafT/yafU*	← → ←	124	238856
sraA / t15	*clpX/lon*	→ ← →	120	-
ffs	*ybaZ/ybaA*	← → →	138	475785
rybA	*ybiP/ybiQ*	→ ← →	205	-
rybB / p25	*ybjK/ybjL*	→ ← ←	80	-
sraB /pke20	*yceF/yceD*	← → →	160	1145980
C0293	*icd/ymfD*	→ → ←	72	1196009
C0299	*hlyE/umuD*	← → →	78	1229930
IS061	*abgR/ydaL*	→ ← →	158	1403676[b]
C0343	*ydaN/dpbA*	→ → →	74	1407461
IS063 / tke8	*ompN/ydbK*	← → ←	117	1435259[b]
C0362	*fdnI/yddM*	→ → ←	385	1550410
dicF	*rzpQ/dicB*	→ → →	53	1647458
rydB/ tpe7/ IS082	*ydiC/ydiH*	← ← ←	67	1762726
rprA/IS083	*ydiK/ydiL*	→ → →	105	1768500
ryeB/tpke79	*pphA/yebY*	← ← ←	100	-
sraC/ *ryeA* /tpke79 /IS091	*pphA/yebY*	← → ←	249	1921338
C0465	*tar/cheW*	← → ←	77	1970840
IS092	*yecJ/yecR*	← ← →	165	1985862[b]
dsrA	*dsrB/yedP*	← ← →	85	2023250
IS102	*yeeP/flu*	→ → →	203	2069540[b]
ryeC/ tp11	*yegL/yegM*	← → →	143	-
ryeD/ tpe60	*yegL/yegM*	← → →	137	-
ryeE	*yegQ/ogrK*	→ → ←	47	-
micF	*ompC/yojN*	← → →	93	2311196
tpke70	*ddg/yfdZ*	→ ← ←	40	
C064	*sseA/sseB*	→ ← ←	86	2651472
IS128	*sseA/sseB*	→ → ←	208	2651743[b]
ryfA/ tp1	*sseA/sseB*	→ → ←	302	2652177
tke1	*yfhK/purL*	← ← ←	158	2689212
ssrA	*smpB/intA*	→ → →	363	2753974
sraD	*ygaG/gshA*	← → ←	70	2812897
C0664	*ygbD/hypF*	→ → ←	57.5	2833189

sRNA gene	Adjacent genes	Strand[a]	Length	3' end position
csrB	*yqcC/syd*	← ← ←	360	2922178
gcvB/ IS145	*gcvA/ygdI*	← → ←	205	2940922
sraE/ rygA/ t59	*aas/galR*	← ← →	88	2974124
rygB/ t59	*aas/galR*	← ← →	83	2974332[b]
ssrS	*ygfE/ygfA*	→ → →	183	3054185
rygC/t27	*ygfA/serA*	→ → ←	139	-
C0719	*yghK/glcB*	← → ←	221	3119595
tp8/ c0730	*yqiK/rfaE*	→ ← ←	144	3192737
sraF/ tpk1/ IS160	*ygjR/ygjT*	→ → →	189	3236203
rnpB	*yhaC/yhaD*	→ ← ←	377	3267857
sraG/ p3	*pnp*rpsO/	← → ←	174	3309039
sraH/ ryhA	*elbB/arcB*	← → ←	108	3348325
sraI/ ryhB/ IS176	*yhhX/yhhY*	← ← →	94	3578554
IS183	*yhiW/yhiX*	← → ←	113	3662604[b]
sraJ/ ryiA/ k19	*aslA/hemY*	← → ←	172	3984216
spf	*polA/yihA*	→ → ←	109	4047585
sraK/ ryiB/ tpk2/ *csrC*	*yihA/yihI*	← → →	245	4048860
oxyS	*argH/oxyR*	→ ← →	110	4155864
sraL/ ryjA	*soxR/yjcD*	→ ← →	140	4275506
SroA			93	75608
SroB			84	506428
SroC			163	686066
SroD			86	1886126
SroE			92	2638706
SroG			147	3182734
SroH			161	4188065

Table 2.3: Summary of the 62 verified ncRNA genes in *E.coli*, were the first 55 ones are from Hershberg et al. (2003) and the last 7 ones from Vogel et al. (2003). a) The first arrow is the direction of the upstream gene, the second is the direction of the ncRNA gene, while the third is the direction of the downstream gene

Chapter 3

Search algorithm

In this chapter the major difficulties in searching for ncRNAs will be briefly mentioned, and the search algorithms will be presented. The implemented program is written in java and the code is about 3 000 lines.

3.1 Why novel ncRNAs are hard to detect compared to protein coding genes

The major problem with detecting ncRNAs by searching for transcription signals is the lack of start and stop codons that are widely used in computational searches for protein coding genes (McCutcheon and Eddy, 2003; Carter et al., 2001). Neither can searching for codon usage skews be applied to ncRNA searches because the nucelotides in a ncRNA gene do not code for amino acids. The remaining signals on the DNA string, such as promoters and terminators, are not as easily recognizable and therefore less reliable factors (Carter et al., 2001).

When it comes to detecting ncRNAs in the laboratory there is one major problem; ncRNAs tend do be degraded very quickly, that is, the actual transcript is unstable and dissolves before it can be detected (personal correspondence with K.I.Kristiansen, 2004). This calls for a different solution to the problem of detecting ncRNAs, and one solution might be computational searches like in this study.

3.2 Where to search for ncRNA in the *E.coli* genome

In the *E.coli* genome there are about 4290 predicted protein coding genes (predicted because only some of them are experimentally verified), and accordingly as many promoters. Other verified genes in *E.coli* are the 7 operons, each containing the code of three known rRNAs. There are,

in addition to this, 86 tRNA genes and also 62 verified ncRNA genes ("ncRNA" here does not include tRNA and rRNA)

The length of the *E.coli* genome is about 4 500 000 bp. The genes coding for proteins make up about 89% of these basepairs, and the rest are divided into intergenic regions. This is where the ncRNA genes are thought to be located. These intergenic regions are the regions previously annotated as "junk-DNA", but now they seem to be containing more than than just junk. In bacteria a single gene lies on one strand of the DNA. Different genes can have different directions as they can be located on any of the two strands, but one gene is on one strand and has one direction in *E.coli*. This well organized system makes intergenic regions easily defined and easier to locate in bacteria compared to higher-level organisms.

Because of the restrictions that a protein coding region on one strand sets upon the complementary strand (namely the complementarity), it is not likely that there are ncRNA genes that are overlapping protein coding genes on either of the two strands. Therefore a search should be concentrated on the intergenic regions, where intergenic is defined as the regions of the genome where none of the two DNA strands encode a protein.

Blattner et al. (1997) located the protein coding genes in *E.coli* genome by basicly searching for start and stop codons. These codons had to be in the same reading frames, and the minimum distance between start and stop codon to make the sequence become a hypothetical protein coding gene was set to 100 triplets. This implies that a minimim length protein coding gene contains 300 nucleotides, which again gives a protein consisting of 100 amino acids. Recently as many as 500 of the hypothetical protein coding genes of *E.coli* have been claimed to be false (Skovgaard et al., 2001). One reason for this is the cut off value of 100 triplets. Today there are known protein coding genes that are shorter than 100 triplets, and also many of the previously predicted protein coding genes, longer than 100 triplets, have been proven not to code for proteins. The cut off were more or less set because 100 is a nice number. The uncertainty here originates in the fact that although the *E.coli* genome is totally sequenced, only about one third of the hypothetical protein coding genes of the *E.coli* genome are experimentally verified.

A recent study by Vogel et al. (2003) (see section 2.4.9 on page 22), has shown that there exists ncRNAs that are results of a parallel output of a transcription, where an ncRNA is processed from mRNA leaders or trailers. These types of ncRNAs will not be detected in this study, since the target search is on novel trancription sites.

The conclusion of where to search in the *E.coli* genome is that the main target of the search is the intergenic regions. There might be hypothetical proteins (not verified, but predicted proteins) that actually are

ncRNAs in stead of proteins, but as these sequences already are identified as transcription sites they are less interesting to a search for novel trancription sites. Still the program will be able to search the entire genome, but then all transcription sites detected will be suggested as ncRNA coding candidates while most of these actually code for proteins.

The actual search string used in this study was produced by using a file from GenBank (Burks et al., 1985) to find all coding sequences in the *E.coli K-12* genome (including genes for tRNA and rRNA). These regions were removed from the genome. The genome and the annotation was collected from GenBank on the 30th of March 2004. As mentioned above only regions being intergenic on both strands were included in the final search string, and the minimum length of an intergenic region was according to Hershberg et al. (2003) set to 300 bp. The remaining DNA string to search in consists of 490 intergenic regions having a total of 228 793 nucleotides.

3.3 How many nucleotides make up an ncRNA

An important part of a ncRNA-gene search is where to make a cut-off concerning what candidates to include. The cut-off discussed here is about the gene length, that is, the number of nucleotides between the promoter and the corresponding terminator. Argaman et al. (2001) discarded all promoters and terminators having less than 50 or more than 400 base pairs between them. Their target was to localize small RNAs, sRNAs, which they assumed to have a length of 50 to 400 base-pairs. This choice of minimum and maximum length has also been suggested and used by Tjaden et al. (2002). The length of possible ncRNA candidates in this study will be set to from 40 to 400 bp. 400 as a max already seems to have good margins, the minimum cutoff is set to 40 to achieve some slack due to how the promoter search algorithm in this program works. Shorter ncRNAs (typically snoRNAs) are known to be present in eukaryotes, but are not known to be, or likely to be in procaryotes (personal correspondence with K.I.Kristiansen, 2004).

3.4 Structure of the search algorithm

3.4.1 Input

The user input will basically be any genomic sequence written in a FASTA format. The user supplies a FASTA file with one or more sequences, preferably intergenic sequences from a bacteria closely related to *E.coli*. Then the user can decide which of the implemented promoter consensus

sequences that are to be searched for, and which threshold value the program should use for each consenus sequence during the search. There are no options adjusting the search criteria of the terminators.

3.4.2 Preprocessing

The preprocessing of the FASTA file is taken care of by a BioJava package created with the single purpose of reading DNA sequences of this format. This package delivers the DNA string and the name of the sequence to the main search program.

3.4.3 Performing the search

One DNA sequence is searched at a time. First a forward search, then the reversed complimentary sequence is created and searched.

The search is divided in three parts.

- Search and score possible promoters in the sequence (see section 3.7).
- Search and score possible terminators in the sequence (see section 3.8).
- Compute the final candidates from the candidates found in step one and two above (see next section).

3.5 Computing the final candidates

The computation of the final ncRNA candidates is relatively easy after the promoter and terminator candidates have been located. The algorithm will look for a pair of promoter and terminator candidates that might “fit” together. A “fit” is when the length between the promoter and the terminator is inside the pre-set threshold range (40-400 bp). When such a fit is found a final candidate has been located, the candidate’s data can now be printed.

3.6 Output

The search program implemented as part of this study is meant to become part of a larger program performing different searches with the same goal, namely to locate ncRNAs, and to deliver a consensus answer. The output must therefore be readable for a master program that looks at all the different candidate suggestions and computes the candidates

that are the “consensus candidates”. In addition to this, this program will be usable as a “stand-alone” program.

The output will be a list of candidate genes, the candidates will have a candidate number, name of the sequence it is located in, a score (explained in section 4.1.4 on page 45), coordinates for its location in the sequence, the length and a direction. The name of the ncRNA candidate are related to the first words of the description line of the FASTA inputfile. If the inputfile is named : “>+mesJ_to_+cutF” , meaning that this FASTA sequence contains the intergenic region between the genes mesJ and cutF (the + is according to the direction). Then any candidates located on this string will be named “+mesJ_to_+cutF” and any candidates on the reversed complement to this string will be named “+mesJ_to_+cutF.reversed”.

The candidates will be sorted by the coordinates, this is to make the comparisons with the other search program’s similar files easier. The coordinates will be according to the input sequence, and not general gene-coordinates. There will also be an output file in FASTA format containing the candidate sequences between promoter end and terminator start.

If the input FASTA file header is formatted exactly like this:
>{sequence name}_{description}_{start position}_{end position}
the absolute coordinates of the candidates will be given at the end of the annotation line in the FASTA output file. The coordinates are always according to the input string. Given the coordinates in the FASTA file to a candidate on the reverse complimentary string, one must extract the sequence according to the coordinates and then reverse compliment it, to get the string given in the FASTA file.

In addition, there will also be one candidate file for all promoters and one candidate file for all terminators independent of whether or not the promoter- and terminator- candidates have been used in the final ncRNA gene candidate file.

3.6.1 Example of output files

The promoter- and terminator-candidate files contains the same data as respectively the “promoter data” and “terminator data” parts of the extract from the file containing all data of the ncRNA candidates (ncRNACandidate.txt) file that can be found in the following examples.

The coordinates given in the ncRNACandidates.txt file are relative for this intergenic-region. The coordinates given in the FASTA file are absolute, if the input sequences have the description string formatted as described above.

Example of one candidate from the ncRNACandidate.txt file:

```
Gene data for this candidate:
Candidate length = 54
Seq name :  +thrA_to_+cutF
Candidate number = 1

Promoter data for this candidate:
Promoter name :sigma_24_regulonDB
-35 Threshold =46.772800000000004
-35 score = 50.84
Distance between -35 and -10 :16
-10 Threshold =209.37359999999998
-10 score = 223.83999999999997
Relative start of promoter:  925
Relative finish of promoter:  954
-35 region :  ggaaaa
-10 region :  aatctgaa

Terminator data for this candidate:
Hairpin Energy score:  -6.5
U_tail score :  -3.5322355209461764
Rightleg = ctgttt
Leftleg = aatcag
Loop =ggg
U_tail = tttccgcacgacctg
Relative start of terminator= 1008
Relative end of terminator= 1038
Numbers of nucleotides in terminator = 31
```

The FASTA output for this exact candidate looks like this:

```
>+thrA_to_+cutF_3754_3807
agatcacaacgagcaggtcagctttgcgcaagccgtaacccaggggttgggcaa
```

3.7 Searching for promoters

The main problem of the promoter search is not the computiational part, which itself is a complicated computational problem, but knowing what to look for!

Argaman et al. (2001) focused their promoter-search on finding DNA sequences recognized by the major *E.coli* RNA polymerase σ (sigma) factor (σ^{70}). This σ factor is annotated as a consensus region of the protein coding genes in *E.coli*. The σ^{70} consensus sequence will be one of the main characteristics to search for in the DNA string. However, as previously mentioned, there are at least 7 different sigma factors in *E.coli* (Owens et al., 1998), and as far as possible all the corresponding consensus sequences should be searched for. Accordingly a main part of this study will be to verify today's consensus regions for promoters, and try to find new and better ones.

3.7.1 Different genes are recognized by different σ factors

Another σ factor known to play an important role in *E.coli* is the σ^{32}. The known transcripts that originate from genes with a promoter recognized by the σ^{32} are so-called heat-inducible proteins. That is proteins produced when the bacteria is subject to a heat shock.

It is very likely that heat shock induced ncRNAs exist, as well as other ncRNAs that might only exist in a certian cell phase (personal correspondence with K.I. Kristiansen, 204). Considering the fact that it is the consensus sequence recognized by σ^{70} that has been used in most earlier searches (Argaman et al., 2001; Chen et al., 2002) the consensus sequences of other σ factors will therefore play an important role in this search.

3.7.2 Definition of a promoter candidate

To be identified as a promoter sequence (i.e. a promoter candidate) that is recognized by a given sigma factor there are certain characteristics the sequence must have. First of all the -10 region has to score above the threshold of the consensus region for this sigma factor. Then there are defined maximum and minimum lengths between the -10 and -35 regions, and inside this range there has to be found a possible -35 region with a score above the threshold. If such a -35 region is present inside the range-bounds of the -10 region a promoter candidate has been located.

3.7.3 Searching for promoter candidates

There are no ways to give a well estimated guess of where in an intergenic sequence a possible promoter candidate might hide. However, there are some hints; an ncRNA is not likely to lie very close to one of the ends of the intergenic region, and it is more likely to hide in a sequence with a length of 300 to 900 bp than in shorter or longer sequences. ncRNAs are not likely to lie in shorter sequences because the sequence is simply too short, while the reason not to hide in longer sequences is that these are very dominated by repetitive sequenes (Hershberg et al., 2003). Since the results mentioned above only can be seen as guidelines and not absolutes a brute force search is needed to ensure that all possibilities are checked. This brute force search can be described in the following pseudo code:

```
public void promoterSearch(Sequence s)
   for (all consensus promoter sequences to search for)
       startPoint = first symbol in s;
       while (more symbols in sequence)
          if (sequence from startpoint to n has a score
             higher than -10 sequence threshold score)
             for(all possible -35 sequences, according
                to the allowed distances)
                if (-35 sequence score > -35 threshold score)
                   store candidate;
       startPoint = next symbol in s;
end promoterSearch
```

The search actually checks every possible substring on the DNA string that might make up a -10 region. If the score of this substring exceeds the threshold score, all possible -35 regions to this -10 region are tested and scored, according to which distances from -35 to -10 that are allowed. If there exists a possible -35 region to this -10 region, with a score exceeding the -35 threshold a possible candidate is located, and stored in a promoter candidate vector. For the actual scoring system see section 4.1.4 on page 45.

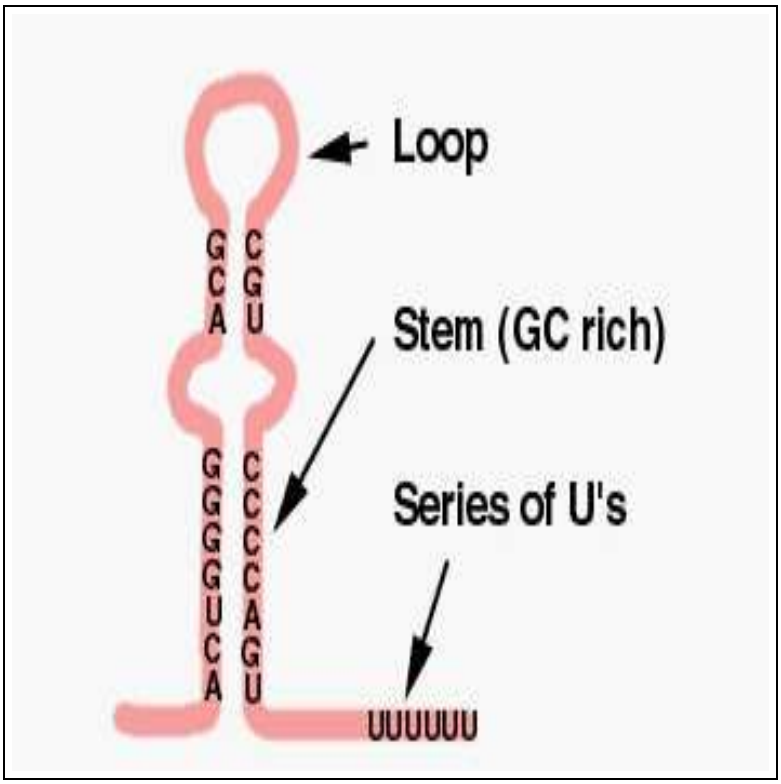

Figure 3.1: Secondary structure of a typical terminator hairpin with a poly U-tail.

3.8 Searching for terminators

The problem of locating terminators is slightly different from the promoter problem. Terminators do not have a consensus sequence, but rather a consensus secondary structure (see subsection 3.8.1). There are two main types of terminators of which one type is dependent upon not only the secondary structure, but also exterior termination factors. These latter terminator sequences seem very hard to locate, if not impossible, when looking only at the DNA sequence (Yada et al., 1999; Richardson, 2002).

The first of the main types of terminators make the elongation stop by the construction of a hairpin (stem-loop) construction, see figure 3.1, on the newly transcripted RNA. When the hairpin region is right it makes the RNA and DNA break away from each other, and the DNA rebinds. The special features of the gene sequences that can make up a stem-loop construction are the main target of the terminator search.

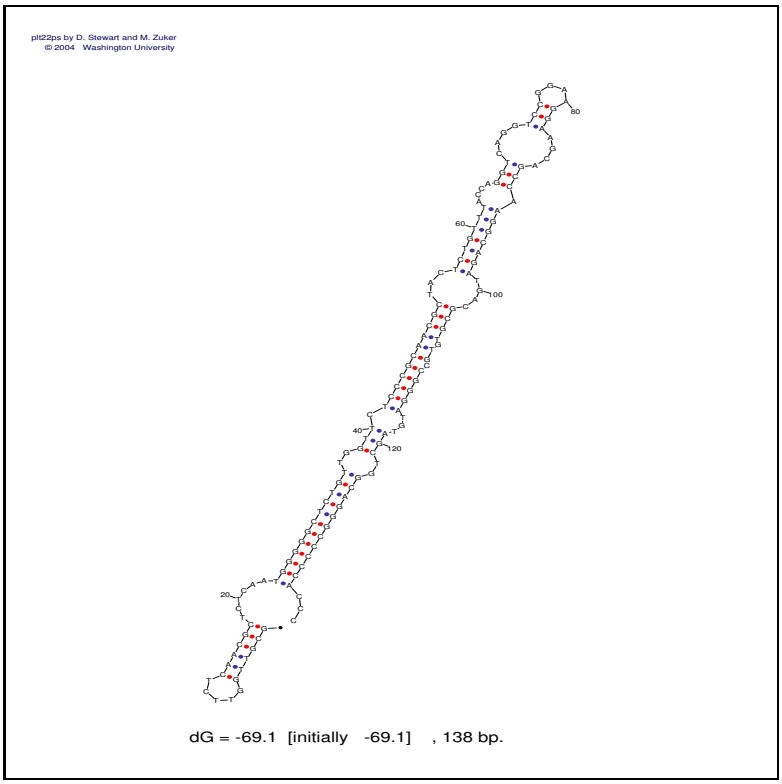

Figure 3.2: Secondary structure of the known ncRNA “ffs”, ncRNA sequence from Hershberg et al. (2003), picture created using Mfold (Zuker, 2003).

3.8.1 Secondary structure folding

A DNA transcript, RNA, struggles to fold itself up to create a stable molecule (see figure 3.2). The structure of a folded RNA can be described two-dimensionally, and such a description is called the secondary structure of the RNA. The target structure for this search is, as mentioned above, called a hairpin, or a stem-loop structure. Such a structure can not form without the nucleotide string being an inverted palindrome with an independent part in the middle of the string, this independent part makes up the loop structure. It is not the actual composition of nucleotides on the string but rather the positioning of the nucleotides that decides whether or not a hairpin structure can be created. The actual base-pair composition of the hairpin is only interesting when it comes to the quality and strength of the hairpin.

3.8.2 Definition of a terminator candidate

As for the promoter candidates there are different criteria that a sequence must fulfill to be considered a terminator candidate. The first criteria is to have a minimum of three consecutive T's no more than five nucleotides downstream from the end of the hairpin structure. The more T's the better in the 15-nucleotide sequence immidiately following the hairpin, having three as an absolute minimum (see figure 3.1 on page 37). In the hairpin structure the loop can range from 3 to 10 nucleotides, and the stem length is constrained to be in the range of 4 to 20 nucleotides. These constraints are the same as the ones used by Ermolaeva et al. (2000), in their search for terminators in intergenic regions. The hairpins are allowed no more than one gap in total (see figure 2.6 on page 18). When one region is found to fulfill all these demands it is considered a terminator candidate, if several candidates are covering approximately the same nucleotide sequence the best scored one remains the final candidate. For the actual hairpin score function see section 4.2 on page 47.

3.8.3 Searching for terminator candidates

The same rules as for the promoters apply for where to search for the terminators, hence a brute force search is again needed. The structure of a ρ-independent terminator has a distinctive U-residue near its end (see figure 3.1). That implies that on the DNA strand coding for this terminator, there will be at least 3 consecutive T's, which in the RNA molecule gives a U-residue. Therefore step one of the brute force is to search the string for sequences of 3 or more consecutive T's. When such a sequence is located the terminator score is computed for every possible terminator constructed by choosing maximum 55 nucleotides upstream from the consecutive T sequence. These possible terminators are constructed so that they fulfill the criteria described above.

Then the tail score is computed, that is the score of the 15 nucleotides immediately following the terminator constructed in part two of the algorithm. These 15 nucleotides contain the three consecutive T's that actually invoked the search to go on from part one of the algorithm. If the total score of the tail and the terminator structure is better than the threshold value of the search, this sequence is considered a terminator candidate.

The actual terminator-search algorithm could be described as this in pseduo-code:

```
public void terminatorSearch(Sequence s)
   startPoint = first nucleotide in s;
   while (more nucleotides in s)
      if(found 3 consecutive T's starting with startPoint)
        for(all 6 different endpoints of terminator structure)
           for(stem length from 4 to 20)
              for(loop length from 3 to 10)
                 for(all gap positions, including no gap)
                    calculate value of terminator score();
                    if(terminator score > terminator threshold)
                      calculate the tailscore();
                      if(tailscore > tailscore threshold)
                        candidate located, store it in
                        terminator candidate vector();
      startPoint = next nucleotide in s;
end terminatorSearch
```

Chapter 4

Scoring system

In this chapter the work to produce the novel promoter score function is described. The score function used for the terminators is also explained.

4.1 Scoring the promoter candidates

The patterns that are searched for during the promoter search are the known consensus sequences of the regions recognized by the different polymerase sigma factors in E.coli. There are seven known sigma factors today (Owens et al., 1998). However, during this study sufficient data to create a consensus sequence has only been found for five of these sequences (Lewin, 2000; Salgado et al., 2000; Pedersen et al., 2000; Ussery, 1999). However, the consensus sequences vary accoring to how hard they are to locate, the dataset used to define them and the fact that there are uncertainties in the datasets. Therefore eight different consensuses are implemented, but they only cover five of the seven different sigma consensus sequences (see table 4.3).

The idea behind the scoring function is to give a weighted score to all nucleotides according to how important they seem to be in the the consensus sequence. We can measure the importance of a certain nucleotide in a specific position by calculating its information value (see section 4.1.3).

4.1.1 Creating consensus sequences

To create a consensus sequence we need a dataset of sequences that together is supposed to make up a consensus, i.e. they are similar but not exactly the same. We then run the sequences through an algorithm for multiple sequence alignment producing a set of sequences positioned in a way which maximises similarity (see figure 4.1). All alignment methods used to create consensus sequences in this study are based on the

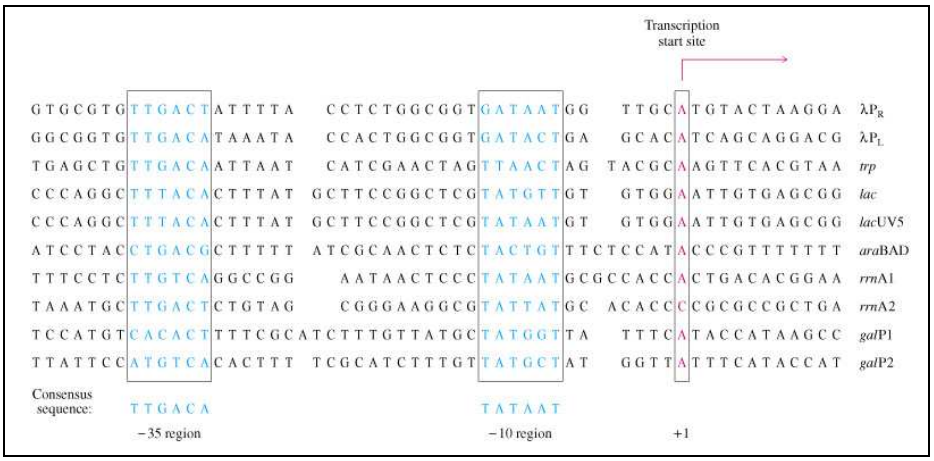

Figure 4.1: Extract of multiple sequence alignment of *E.coli* promoters, picture courtesy of Principles of Biochemistry (http://cwx.prenhall.com/horton/medialib/media_portfolio/index.html).

algorithm published by Smith and Waterman (1981).

In the set of aligned sequences certain nucleotide positions would have the same nucleotide in about every sequence. These positions are the important ones, and we say that they are highly conserved. Then the consensus of the dataset is extracted by looking at what parts of the aligned sequences that are most conserved. Logo plots are often used to visualize consensus sequences (see section 7.1 on page 66).

4.1.2 Consensus sequences are evolving

During the last years several promoter sequences have been verified, and in many of these the -10 and -35 regions are less conserved than previously estimated. This makes the overall significance of the promoter regions weaker, which again gives a weaker consensus. In table 4.1 there is an overview of the development of the understanding of the conservation of the promoter consensus sequences in *E.coli*. It should be noted that there has been a remarkable weakening of the consensus sequences over the last few years, and that the actual consensus sequences are changing due to new data. This implies that the transcription signals represented by the promoter consensus sequence might be too weak to become a good search criteria.

The consensus sequences used in this study are from three different sources. There is one promoter consensus sequence that is not connected to one specific sigma subunit (in this study called σ^x). This promoter consensus sequence origins in data from Lisser and Margalit

Consensus (-35,-10)	**T**	**T**	**G**	**A**	**C**	**A**		**T**	**A**	**T**	**A**	**A**	**T**
Hawley and McClure	82	84	79	64	54	45		79	95	44	59	51	96
Harley and Reynolds	78	82	68	58	52	54		82	89	52	59	49	89
Lisser and Margalit	69	79	61	56	54	54		77	76	60	61	56	82
David Ussery	46	55	23	25	17	43		40	46	33	36	23	53
Regulon DB	31	36	19	14	19	21		51	54	32	40	41	53

Table 4.1: This table shows the figures on the conservation of promoter consensus sequences have changed since 1983. The level of conservation is given as the nucleotide frequency in %. The first 3 inputs are for the non-sigma specific case, while the 2 latter are the consensus sequences recognized by σ^{70}. The 3 first sequences are reproduced from Lisser and Margalit (1993), while the fourth is from Pedersen et al. (2000) and Ussery (1999), and the fifth from Salgado et al. (2000). It should be noticed that the consensus sequences are different in the two latter ones, individually and compared to the three others.

(1993), the sequences used in their study were all confirmed promoter sequences. The promoter sequences from the Regulon database (Salgado et al., 2000) used to compile the three different consensus sequences ($\sigma^{70}b$, $\sigma^{32}b$ and σ^{24}) are from predicted promoters. The last four consensus sequences ($\sigma^{70}a$, $\sigma^{32}a$,σ^{54} and σ^{38}) are from the work of Pedersen et al. (2000) and Ussery (1999), they do not state whether their sequences are from verified or predicted promoters. However, according to the size of their dataset it is most likely that their sequences are from predicted promoters, because they have more sequences than the number found in this study on verified promoters in *E.coli*. The number of sequences in each dataset can be found in table 4.2. Due to different consensuses from the different datasets, there are for some of the sigma factors more than one consensus sequence. It will always be obvious due to the context which dataset a consensus sequence origins from. In table 4.3 the consensus sequences used in this study are listed. For the remaining σ factors, σ^{28} and σ^{19}, there were not sufficient data to create a reasonable scoring function.

4.1.3 Logo plots

A logo plot for a DNA region is basically a plot of every nucleotide position in the sequence showing the information value (see below) of the nucleotide position and the distribution of the different nucleotides in this position. From a logo plot one may easily extract the significance of the sequence. All logo plots used in this study are created from datasets

Promoter type	Dataset size	Reference
$\sigma^{70}a$	3851	Pedersen et al. (2000); Ussery (1999)
$\sigma^{32}a$	75	Pedersen et al. (2000); Ussery (1999)
σ^{54}	177	Pedersen et al. (2000); Ussery (1999)
σ^{38}	68	Pedersen et al. (2000); Ussery (1999)
σ^{x}	298	Lisser and Margalit (1993)
$\sigma^{70}b$	4379	Salgado et al. (2000)
$\sigma^{32}b$	48	Salgado et al. (2000)
σ^{24}	33	Salgado et al. (2000)

Table 4.2: Table of the sizes of the datasets and a reference to the datasets from which the promoter consensus sequences originate. The promoter types are defined according to which sigma factor that recognizes it. (Some are marked with an *a* or a *b* to distinguish them from each other throughout this study.)

Factor	-35 Sequence	Separation	-10 Sequence
$\sigma^{70}a$	TTTTAA	17 -19 bp	AAAATT
$\sigma^{32}a$	TTTAAAAA	9 - 11 bp	CCCAATTT
σ^{54}	TGGGCA	4 - 7 bp	CTGGC
σ^{38}	TTTTAA	16 - 18 bp	TAAATTT
σ^{x}	TTGACA	15 - 21 bp	TATAAT
$\sigma^{70}b$	TTTTTT	15 -21 bp	TAAAAT
$\sigma^{32}b$	TAAAAA	10 - 12 bp	CCCCATT
σ^{24}	GGAAAA	16 - 18 bp	AGTCTGAA

Table 4.3: The σ-factors and their consensus sequences used in this study.

already run through a multiple alignment.

Information value is a term from Information Theory, and in this setting it describes the value provided by having information on a nucleotide in a certain position. In this study the target is a sequence that has a certain similarity with the consensus sequence, it is therefore interesting to know which of the nucleotides in the consensus that are the most important ones. For example, if there is an equal chance for any nucletide in a certain position in the consensus, the value of knowing that there is an A there is zero. Accordingly if there is 96 % chance for an A in a position (percentage is computed from the multiple alignment) it would mean a lot knowing if there really is an A there. The whole idea behind this novel scoring function is to use the entropy of every nucleotide in its position. The main idea is to compute the maximum entropy in a positon and subtract the actual entropy of every nucleotide in that position. To use this method, equal probability (1/4) of every nucleotide in every position must be assumed. This assumption is usual to make when dealing with intergenic regions, in coding regions this assumption do not hold because of codon skews. The idea described above can be used to assign a nucleotide in a given position an information value. The information value is described in bits and ranges from two to zero, having zero as a minimum.

To compute this information value, I, we use this formula :

$$I = 2 + \sum_x P_x log_2 P_x = \sum_x P_x log_2(4P_x) \tag{4.1}$$

where $x \in A, T, C, G$,
P_x = the probability of nucleotide x in position i, $P_x \in [0, 1]$,
then the information value, I, can be calculated for every position, i, in the sequence:

The distribution of the different nucleotides in every position, i, is then calculated by $I_x = P_x \cdot I$, where I = $\sum_x I_x$ for this position i.

An information value has been assigned to all nucleotide positions in the multiple alignment. Also the frequencies of every nucleotide in every position is given in percent, so now the logo plot can be created.

From the data collected and processed to create the logo plot, a score table containing the information value of every nucleotide position in the consensus can be extracted. The score table also contains the individual frequency of each nucleotide in every position (see section 7.1 on page 66).

4.1.4 Scoring the possible promoter sequences

From the information values and nucleotide frequencies explained above a scoring function can be created.

Considering the information values and the frequencies, a set of nucleotides of the same length as for instance the σ^{70} -10 consensus region, can be extracted and assigend a score:

$$Score = \sum^{i} I_i P_{x,i} \tag{4.2}$$

,where
$i \in$[0, length of -10 region] i.e i equals the position in the -10 region
$x \in A, T, C, G$
$P_{x,i}$ = the probability of nucleotide x in position i, $P_x \in [0, 1]$
I_i = Information value of nucleotide position i
x,i denotes nucleotide x in position i

This scoring function scores the nucleotides independently of each other, but it is based upon a multiple sequence alignment. Thus by looking at how the score-data were produced, it is easily seen that the context is taken into consideration. A sequence can not become a promoter candidate unless it is very similar to the consensus sequence, and therefore many of the nucleotides with a high information value will have to be present to achieve a score above the threshold.

This scoring function is a refinement of the often used log-odds-score; which only takes the frequencies and not the information value in every nucleotide position into consideration.

If any of the nucleotides have zero occurences in any one position in the multiple sequence alignment, there will be a problem with a division by 0. This problem is usually solved by adding a “pseudocount” of one occurence of each nucleotide in every position before the actual counting begins. This has been done in the datasets where this problem has occured. The “pseudocount” can be justified by the fact that our consensus sequence is based on only a small part of all the promoter sequences that exist.

The total score of a possible promoter sequence candidate is the sum of the -10 region score and the -35 region score, and the higher the score is, the better the candidate is.

The distance between the -35 and -10 regions is also a possible target for scoring and evaluating the data. It seems that for the σ^{70} consensus sequence a distance of 17 nucleotides between the two regions is optimal (Owens et al., 1998), having a maximum and minimum distance of respectively 15 to 21 nucleotides. However there are not sufficient data available to create a good scoring function for this distance, so only plain maximum and minimum distances have been set for the different consensus sequences. Accordingly, there is no score function that considers this distance in the program developed in this study.

4.2 Scoring terminator candidates

The search and scoring algorithm for the terminators is basically the same as the one used by Ermolaeva et al. (2000). However, there are fewer known facts about the input sequence in this study. The adjacent genes and their direction are not known, thus a part of their scoring algorithm cannot be applied here. In the stem of the hairpin the different nucleotide pairs are scored differently; a G-C pair is the most stable nucleotide pair and is assigned the score S_{gc}, followed by A-T with the assigned score S_{at}. (T is used here since the search is performed directly on the DNA, without transcription T → U.) In RNA there is also weak interaction between G and T, this pair is given the score S_{gt}. All other pairs are considered a mismatch and scored by S_{mm}. A gap in the stem on either side of the hairpin structure weakens the structure further, only one gap per hairpin is allowed, and it is scored by S_{gp}. As for the loop stucture it is more stable the closer it gets to the minimum length, therefore each nucleotide of the loop is assigned with score S_{lp}. The energy score of the hairpin structure is computed by combining the scores of the pairs in the stem.

$$\text{Energy} = S_{gc} \cdot x_1 + S_{at} \cdot x_2 + S_{gt} \cdot x_3 + S_{mm} \cdot x_4 + S_{gp} \cdot x_5 + S_{lp} \cdot x_6$$

x_1, x_2 and x_3 are the counts of the G-C, A-T and the G-T nucleotide pairs in the stem, x_4 and x_5 are the numbers of mismatches and gaps, and x_6 are the number of nucleotides in the loop.

The energy scoring function is designed to separate hairpins from non-hairpins by giving the non-hairpins a score below the energy threshold for a hairpin.

The optimizing problem of the parameters in the scoring function was solved using a decision tree by Ermolaeva et al. (2000). The assigned parameter scores give this energy function:

$$\text{Energy} = 2.3x_1 - 0.9x_2 + 1.5x_3 + 3.5x_4 + 6.0x_5 + 1.0x_6 - 5.7$$

These parameters are the ones used in all terminator hairpin energy calculations in this program. As mentioned above the parameters have been optimized for this specific task using a decision tree. Their values might look strange, but all of x_1, x_2 and x_3 have a score better than the mismatch score, *mm*. (The score parameters have been verified by personal correspondance with Dr. Ermolaeva.) The lower the energy score is, the more stable is the hairpin structure.

To score the tail of the hairpin the length and distance from the hairpin are the important factors to look at. The closer to the hairpin a poly-

T stretch is, and the longer it is, the better. The tail scoring function is also the same as the one used by Ermolaeva et al. (2000).

$$Tailscore = -\sum_{n=1}^{15} x_n$$

For all T residues in the 15-nucleotide region immediately following the hairpin structure where $x_0 = 1$ and

$$x_n = \begin{cases} x_{n-1} \cdot 0.9 & \text{if the } n\text{th nucleotide is a T} \\ x_{n-1} \cdot 0.6 & \text{if the } n\text{th nucleotide is other than T} \end{cases}$$

A low tailscore value indicates a T-rich tail of the terminator candidate. The sum of these two scores decides whether or not this particular region is to be considered a terminator candidate.

Chapter 5

Program complexity and runtime

In this chapter the complexity of the program developed in this study will be discussed along with examples of actual runtimes. Initially, running the program on regular computers versus running it on a larger grid-system will be discussed.

5.1 Running the program

The program developed in this study demands all the CPU capacity while running on a regular computer. A typical program runtime on a 2.4 GHz, pentium 4 CPU having 512 Mb of RAM is about 16 minutes, this is when searching for all the eight possible promoter consensus sequences while using the thresholds found in table 6.1, and the input string being the 490 intergenic regions of *E.coli*..

At the Department of Informatics (IFI) a grid system for distributing processes to idle CPU's has recently been installed. The system has been named Condor. Condor has at its disposal about 260 CPUs, more than 80GB of memory and a total capacity of 110 GFlops. A nice feature of the Condor program is that the user might launch many executions of the program at the same time using one single submit file. The submit file can use several input files and thereby function as a parallelization of the program. To avoid using most of the Linux cluster for general shell-login at IFI the Condor program was used during the final part of this study. Previous executions of the program, especially while running several programs in parallel, caused a huge slow-down of the Linux cluster it was running on. The runtime of the program on the Condor system is usually less than 10 minutes, even if there are more than ten instances of the program that are running in parallel. The runtime varies a bit due to the number and performance of the idle processors.

5.2 Introduction to runtime and complexity calculations

An often used description of short nucleotide sequences is m'mers. A sequence of m nucleotides can be named an m'mer, and in a DNA sequence a new m'mer is found every time we make a single step on the DNA. The correct number of m'mers in a sequence of n nucleotides is $n - m + 1$. As long as m is not too big, the number of m'mers in a DNA sequence is approximately equal to the number of nucleotides in the sequence. For example in a DNA sequence of 500 nucleotides there will be 496 5'mers, which is $\approx$500.

In the case when one also considers the reverse complementary DNA string there will be about twice the amount of m'mers in the sequence as there are nucleoties. In all equations in this chapter n is the number of nucleotides in the DNA input sequence. The following runtimes are set while searching a FASTA file containing the intergenic regions of *E.coli* consisting of 490 DNA sequences and a total of 228 793 basepairs. For all runtimes given in this chapter the thresholds used are the ones found in table 6.1 on page 56. All computations and actual runtimes are computed and measured due to running the program on the machine kaksi.uio.no, which is a Dell 2650 server having an Intel Xeon CPU of 2.00 GHz, with hyperthreading and 2GB of RAM. The runtime calculations have been tested for this machine and follow a linear function at least until a 2.7 Gbase input, see next sections. On different machines the figures will differ depending on the amount of available RAM, however, the runtime can still be described by a funtion linear to the input length.

5.3 Complexity and runtime of the promoter search

A typical -10 region has a length of about 6 nucleotides, and there are eight -10 regions that we look for. There are $4^6 = 4096$ possible six'mers in a DNA sequence. When searhing for σ^{24} using a threshold of 97%, the program has a possible hit on a -10 region for 4 different 6'mers. Dividing the number of possible 6'mers with the number of wanted 6'mers it turns out that for the random case there should occur one hit for every 1024th nucleotide in the input string. For every such hit the continuing search for a -35 region uses less than 100 operations. (This can be deduced from the pseudocode in section 3.7.3 on page 36.)
Let Z_p denote the time consumption of a -35 region search, and Z denote the time consumption of one single operation. The complexity of the promoter search can now be described as:

$$2 \cdot n \cdot Z + (2 \cdot n/1024)Z_p. \tag{5.1}$$

where the factor 2 originates from the fact that the complementary string is also searched.
Since $(2 \cdot n/1024)Z_p \ll n \cdot Z$ (except from the case of a very short input string) , the complexity of the promoter search is $2 \cdot n \Rightarrow n$, which gives a runtime estimate of O(n).
If more than one consensus sequence is searched for, it is easily seen that the complexity of the promoter search is still deducable O(n).

As shown above, the increase in time consumption is a linear function of n. While running the program the promoter search for all the promoter consensus sequences, using the thresholds found in table 6.1, the program uses a little less than 2 minutes per megabase of the input string. This gives a total of about 25 seconds to search the intergenic regions of *E.coli*. The runtime also includes writing the promoter candidates to a file.

5.4 Complexity and runtime of the terminator search

The calculation of the terminator search complexity and runtime is exactly as for the promoter search above. The only differences are the constants of the initial hit ratio, and number of operations per initial hit. In the random case every 64th triplet will consist of three T's. The time consumption of every such hit is constant of a little less than 5500 operations, let the time consumption of a triple T hit be denoted by Z_t. (This can be deduced from the pseudocode in section 3.8.3 on page 39.) Unless the input string is very short Z_t can be neglected, just as for Z_p, this gives a terminator search complexity of :
$2 \cdot n \Rightarrow n$, which again implies a runtime estimate of O(N).
While running the program the terminator search uses about 19 minutes per megabase of the input string. That gives a runtime of 6.08 minutes when searching the intergenic regions of *E.coli*. As for the promoter search, this runtime includes writing the terminator candidates to a file.

5.5 Complexity and runtime of the final candidate computation

The complexity of this method is linear, and the runtime is very dependent upon how many promoter and terminator candidates that are found in the different input sequences, which again is heavily dependent upon

the thresholds. Worst case scenario is one single long DNA input sequence, because then all promoter candidates (P) will be checked with all terminator candidates (T). Then the number of comparisons to be made is equal to the number of promoter candidates times the number of terminator candidates.
This gives a complexity of $P \cdot T$.
Since $P \cdot T$ in absolutely most cases will be $\ll n \cdot Z$ unless there is a very special input string. Accordingly the complexity of the candidate computation can be set to n for convenience.
From this follows a runtime estimate of O(N).
Anyhow, the most CPU demanding operation in this method is to print the FASTA formated output file. The size of this file is dependent on the numbers of hits, which is easily seen as linear to the input string, unless the input string is strongly manipulated. However the runtime of this final candidate computation and file writing has little influence on the total runtime (see next section). The complexity of the FASTA print method can also be set to n, since the number of candidates to print is linear to the input string, and the complexity of one print operation is in the worst case n

The actual runtime of this method with the given thresholds is a about 3.30 minutes per megabase of the input string, which gives a total of 45 seconds runtime when searching the intergenic regions of *E.coli*. This also includes printing the ncRNA candidates to the text file and the FASTA file.

5.6 Total complexity and program runtime

The initiation of the scoring tables is the first thing the program does, the computational time of this operation is a constant k, and in most cases $k << n \cdot Z$, so this constant can be ignored. In the cases where $k << n \cdot Z$ is false, the complete runtime of the program would be close to 0.

The seemingly big difference in actual runtime from the promotersearch and the final candidate computation to the terminatorsearch is not expressed by the big O-notation. This is so because constants are thrown away in the big O-notation. However the actual difference in runtime can be explained as follows:

We denoted the complexity of a -10 region hit in the promoter algorithm to be Z_p, representing a constant of about 100 operations. Moreover we set the complexity of an initial hit in the terminatorsearch to be Z_t, representing a constant of about 5500 operations. Z is still the runtime of one single operation.
This gives us:

Promoter search runtime :

$$Runtime(promoter) = 2 \cdot n \cdot Z + (2 \cdot n/1024) \cdot 100 \cdot Z \approx 24seconds \tag{5.2}$$

Terminator search runtime :

$$Runtime(terminator) = 2 \cdot n \cdot Z + (2 \cdot n/64) \cdot 5500 \cdot Z \approx 368seconds. \tag{5.3}$$

Using equation 5.3 to compute Z, this Z could be inserted in equation 5.2 and the result will be the approximately 24 seconds that the runtime actually is. Another way to explain it is that the promoter search runs approximately 15 times faster than the terminator search, because of the differences in the initial hitratio and number of operations per hit. A way to exploit the fact that the terminator search is a dominant runtime consumer is discussed in section 9.3.

The discussion above explains why the total runtime of the entire program can be estimated as O(N), even though the terminator search is the absolute dominating factor of the runtime. As shown in this chapter the complexity of the program is n, and the runtime is linear to the length of the input string.

Chapter 6

Analysing the program output

In this section methods used to analyse the program output is described. First of all, the idea behind the threshold values used will be explained. All the corresponding results are found in section 7.2 on page 75 and in section 7.3 on page 75.

6.1 Setting the threshold values

With the terminator cut-off set to -6 for the hairpin energy score and -2 for the tailscore, the program created in this study locates 1.309 terminator candidates when searching the intergenic regions of *E.coli*. These cut-off values give many false positives, but also include almost every good terminator candidate in the dataset (Ermolaeva, personal correspondence, 2003). A better terminator search is impossible to achieve without knowing more about the input sequences than what we request in this program.

The most important part of the search is therefore, as previously mentioned, the promoter search. After establishing the eight promoter consensuses that we wish to search for, a threshold must be set regarding when to include or exclude a promoter candidate. Due to different sizes of the datasets the sigma consensus sequences are derived from (see table 4.2 on page 44), the information values of the nucleotides in the different consensus sequences vary (see section 4.1 on page 41). The promoter threshold is set as a percentage of the score a consensus sequence gets when run through the scoring algorithm.

If the dataset that gave the consensus included many sequences a cutoff value of 97% seems to open up for many more hits than the same cut-off will give on a consensus coming from a smaller dataset. This originates from the fact that in the large datasets the consensus is weak, measured in bits, compared to the smaller datasets (see section 4.1.3). Thus many more sequences are included in a 97% similarity cut-off on

Promoter type	Threshold (%)	# ncRNACandidates
$\sigma^{70}a$	95	125
$\sigma^{32}a$	82	52
σ^{54}	93	74
σ^{38}	91	87
σ^{x}	88	67
$\sigma^{70}b$	95	162
$\sigma^{32}b$	83	79
σ^{24}	85	96

Table 6.1: The thresholds used when searching for candidates for further analyzis.

a low information value consensus than on a consensus with higher information values. This also implies that a threshold of 97% would score more sequences above threshold on a weak -35 region consensus than on a strong -10 sequence. This flexibility can justify that the same threshold percentage goes for both the -10 and the -35 region in the same consensus sequence, although their total information value migth be different. To find a reasonable threshold value to the different consensus sequences used here the program was run several times with thresholds giving several hundreds or thousands of candidates up to a threshold giving 0 candidates or the threshold being 100% (see table 6.2).

It should be noted that in some cases the number of promoters is less than the number of ncRNA candidates. This is because there might be several terminators that could fit with a promoter.

To set a reasonable threshold table 6.2 was used. For the most used sigma factors ($\sigma^{70}a$ and $\sigma^{70}b$) the thresholds were set so that they included more candidates than for the other σ factors. The general idea of the thresholds was to set them so that the number of candidates included was about 70 (+/-30) per specific promoter consensus sequence. The thresholds were chosen like this to create a candidate set that had a reasonable size for the time available in this study. The thresholds chosen can be found in table 6.1

Table 6.2 contains a list of how many hits every promoter sequence type achieves using different thresholds. The number of final ncRNA candidates when using a given threshold is also included.

Promoter type	Threshold (%)	Promoter hits	ncRNA candidate hits
σ^{70}a	89	1127	624
σ^{70}a	90	739	411
σ^{70}a	91	610	331
σ^{70}a	92	610	331
σ^{70}a	93	506	270
σ^{70}a	94	243	125
σ^{70}a	95	243	125
σ^{70}a	96	229	119
σ^{70}a	97	54	19
σ^{70}a	98	54	19
σ^{70}a	99	24	8
σ^{70}a	100	24	8
σ^{32}a	80	207	91
σ^{32}a	81	173	80
σ^{32}a	82	107	52
σ^{32}a	83	78	44
σ^{32}a	84	59	25
σ^{32}a	85	42	18
σ^{32}a	86	32	14
σ^{32}a	87	20	9
σ^{32}a	88	14	7
σ^{32}a	89	10	2
σ^{32}a	90	6	0
σ^{32}a	91	4	0
σ^{32}a	92	2	0
σ^{32}a	94	2	0
σ^{32}a	95	1	0
σ^{32}a	96	0	0
σ^{54}	88	378	251
σ^{54}	89	206	143
σ^{54}	90	178	121
σ^{54}	91	144	101
σ^{54}	92	102	74
σ^{54}	93	102	74
σ^{54}	94	71	56
σ^{54}	95	60	52

Promoter type	Threshold (%)	Promoter hits	ncRNA candidate hits
σ^{54}	96	23	20
σ^{54}	97	11	15
σ^{54}	98	8	9
σ^{54}	99	0	0
σ^{38}	87	474	255
σ^{38}	88	379	202
σ^{38}	89	361	198
σ^{38}	90	210	97
σ^{38}	91	157	87
σ^{38}	92	49	18
σ^{38}	93	46	17
σ^{38}	94	37	15
σ^{38}	95	15	3
σ^{38}	96	9	2
σ^{38}	97	9	2
σ^{38}	98	0	0
σ^{x}	84	546	292
σ^{x}	85	436	230
σ^{x}	88	134	67
σ^{x}	89	87	41
σ^{x}	90	60	30
σ^{x}	91	47	26
σ^{x}	92	37	20
σ^{x}	93	18	12
σ^{x}	94	2	1
σ^{x}	95	2	1
σ^{x}	96	1	1
σ^{x}	100	1	1
σ^{70}b	92	814	444
σ^{70}b	93	561	297
σ^{70}b	94	413	232
σ^{70}b	95	278	162
σ^{70}b	96	90	59
σ^{70}b	97	64	38
σ^{70}b	98	22	9
σ^{70}b	99	17	7
σ^{70}b	100	2	0
σ^{32}b	80	327	166
σ^{32}b	81	252	121

Promoter type	Threshold (%)	Promoter hits	ncRNA candidate hits
σ^{32}b	82	208	99
σ^{32}b	83	158	79
σ^{32}b	84	112	44
σ^{32}b	85	77	33
σ^{32}b	86	55	21
σ^{32}b	87	42	18
σ^{32}b	88	32	16
σ^{32}b	89	26	13
σ^{32}b	90	16	8
σ^{32}b	91	9	3
σ^{32}b	92	7	3
σ^{32}b	93	7	3
σ^{32}b	94	5	3
σ^{32}b	95	3	1
σ^{32}b	96	1	1
σ^{32}b	97	0	0
σ^{24}	82	517	233
σ^{24}	83	382	180
σ^{24}	84	255	135
σ^{24}	85	182	96
σ^{24}	86	122	59
σ^{24}	87	84	28
σ^{24}	88	50	15
σ^{24}	89	28	10
σ^{24}	90	19	8
σ^{24}	91	11	5
σ^{24}	92	2	2
σ^{24}	94	2	2
σ^{24}	95	1	0
σ^{24}	97	1	0
σ^{24}	98	0	0

Table 6.2: Table of the number of hits in *E.coli* using different promoter thresholds.

6.2 Comparing results from this study with the random case

For some genes there are several promoters upstream of the gene, but the absolute dominating case is one promoter to one gene. In table 6.3 there is an overview of the number of predicted promoters to known genes. This indicates that we should expect to find only one promoter in front of novel ncRNA genes.

6.2.1 Estimating number of promoter hits in a random DNA sequence

By estimating how many promoter hits the program will give on a random sequence of nucleotides, it is possible to say whether the search criteria are suitable, and if any findings are significant.

First of all, a model of a random string must be created. To become a reasonable model, the random string should maintain the di- and three-nucleotide frequency of the search string. This is because there are dependencies between nearby nucelotides in the string, and it has been proven that to have a decent model of an intergenic DNA string both di- and tri-nucleotide frequencies must be accounted for. The need for taking more than the single-nucleotide frequency into consideration when creating the random model can easily be proven by counting k'mers and looking at dependencies in the k-nucleotide frequencies when adding one nucelotide at the time. Measurements of entropy shows that for the intergenic regions of *E.coli* it is sufficient to include up to tri-nucleotide frequencies to achieve a reasonable model of the input string. This makes the random model not entirely random, but rather a random model that conserves the di- and tri-nucleotide frequencies in the original string. This also implicates that the single nucleotide frequency is maintained.

Promoters per gene	Number of genes
6	3
5	1
4	10
3	62
2	427
1	3506

Table 6.3: Table of the number of predicted promoters per gene in *E.coli* (Salgado et al., 2000).

Promoter type	Threshold (%)	# Promoter hits
$\sigma^{70}a$	100	24
$\sigma^{32}a$	95	1
σ^{54}	98	8
σ^{38}	97	9
σ^{x}	100	1
$\sigma^{70}b$	100	2
$\sigma^{32}b$	96	1
σ^{24}	97	1

Table 6.4: The thresholds used when producing candidates to compare with the random case.

When the frequencies of every single-, di- and tri- nucleotide in the sequence have been found, a simple version of the hidden Markov model (Eddy, 1998) can be used to create a random DNA sequence where these three frequencies are maintained. However, random sequences are not created, because to extract correct statistical data a huge number of random sequences would have to be produced, which again gives more computational work than necessary.

Instead features of the Markov model are used to compute a formula that can estimate the number of hits when searching for one promoter consensus sequence with a given threshold. The threshold used for this comparison between estimates and actual hits was set much higher than the threshold for what is or is not a fair candidate. This was done because a lower threshold would include very many different promoter consensus sequences scoring above the given threshold, and this againg implies an unnecessary big computational work and more possible error sources. A list of the chosen thresholds and their corresponding numbers of promoter hits can be found in table 6.4.

The chosen thresholds will, for most of the promoter consensus sequences, allow more than one sequence to be recognized as a candidate. Using the different thresholds all possible nucleotide words that might make up a -10 or a -35 region having a score at or above the threshold were produced. Then the frequencies of these nucelotide words in the above described random DNA string were computed.

To estimate the frequency of the m'mers in a random sequence this formula can be used:

$$\hat{N}(W) = \frac{N(w_1, w_2, ..., w_k) \cdot N(w_2, ..., w_{k+1}) \cdot .... \cdot N(w_{p-k+1} ..., w_p)}{N(w_2, ..., w_k) \cdot .... \cdot N(w_{p-k+1} ..., w_{p-1})} \tag{6.1}$$

,where

k = the length of the counted m'mers, here k = 3 (see above).
$w \in A, T, C, G$
W = $w_1, w_1, w_2, w_3, ..., w_p,$, (W is a p'mer)
N(W) = actual number of occurences of W in the intergenic DNA of *E.coli*
$\hat{N}(W)$ = estimated number of occurences of W in a random DNA sequence that conserves the features from the original input given by the counted k'mers.

Estimates of the number of occurences of all -10 and -35 region words that score above the chosen thresholds have now been established. The next step is to estimate the number of promoter hits in a random sequence given this k'mer distribution by estimating the number of occurences where a -10 region lies within the allowed distance of a -35 region. These estimates were computed as follows :

W_1 = a word scoring above threshold from the -10 region
W_2 = a word scoring above threshold from the -35 region
$|W_n|$ = the length of W_n
L = 228.793 , i.e. the number of nucleotides in the input as a total
n = 490 , i.e. the number of intergenic regions in the input data
$\bar{x}$ = [*a*,*b*], where *a* is the minimum and *b* is the maximum distance between -35 and -10 region
$\hat{\mu}(W_1\bar{x}W_2)$ = is the expected value of the probability of whether one randomly picked string of length $|W_1| + \bar{x} + |W_2|$ is a string consisting of $W_1, \bar{x}$ and W_2.
$\hat{N}(W_1\bar{x}W_2$ = is the estimated number of occurences of this combination of W_1, W_2 having a distance in the range of [*a*,*b*].
Because $a < 4$ always, and the input sting is intergenic (i.e. not known to contain reading frames) the distribution of possible -10 and -35 regions can be considered independent of each other. This gives

$$\hat{\mu}(W_1\bar{x}W_2) = \hat{\mu}(W_1) \cdot \hat{\mu}(W_2) \tag{6.2}$$

Also we know that

$$\hat{\mu}(W) = \frac{\hat{N}(W)}{L - n \cdot (|W| - 1)} \tag{6.3}$$

where the part "$n \cdot (|W| - 1)$" accounts for the fact that the input string consists of n intergenic regions.
From this follows :

$$\hat{N}(W_1\bar{x}W_2) = \hat{\mu}(W_1\bar{x}W_2) \cdot (L - n(|W_1| + \bar{x} + |W_2| - 1)) \tag{6.4}$$

Since $\bar{x}$ varies over [*a*,*b*] we now have:

$$\hat{N}(W_1\bar{x}W_2) = \sum_{x=a}^{b} \hat{N}(W_1 x W_2) \tag{6.5}$$

Which finally gives :

$$\sum_{x=a}^{b} \hat{N}(W_1 \tilde{x} W_2) = \sum_{x=a}^{b} \hat{\mu}(W_1)\hat{\mu}(W_2) \cdot (L - n(|W_1| + x + |W_2| - 1)) \quad (6.6)$$

As mentioned above there are in most cases more than one sequence in the -10 and -35 region that give a score above the chosen threshold. To get the correct estimate of hits in a random sequence we therefore have to add up all estimated hits for all combinations of words in the -10 and -35 region that have a score above threshold. In table 7.9 on page 75 the estimated number of promoter sequence hits in a random sequence, given the 3'mers, is compared to the actual number of promoter sequence hits produced by the program developed in this study. It should be noted that this analysis is independent of the terminator hits.

6.3 Aligning candidates to intergenic regions in related bacteria

A typical property of a DNA sequence that holds important information is that there are homologs (very similar sequences) to this sequence in the DNA of closely related species. Considering this it is easily seen that a homology search, using the candidates from this program to search intergenic regions of bacteria closely related to *E.coli*, would be a way to measure the quality of the candidates and also the program. Karin Lagesen, a Phd.student at the Bioinformatics group at Rikshospitalet, has developed a program for the general case of multiple sequcene alignments of intergenic regions.

About 850 candidates from the program output have been used to create multiple alignments with intergenic regions from 11 bacteria related to *E.coli* (see appendix B on page 93). The multiple alignments have been created using CLUSTAL X (v1.81). 742 of the candidates used in these alignments results from running the program with the thresholds found in table 6.1, while the last 91 candidates used have been found by selecting thresholds for the sigma consensus sequences so that the number of output candidates is somewhere around 10. These 91 candidates are of course included in the larger dataset of the 742 first candidates.

The alignments were made by using the entire sequence of this program's ncRNA candidates and then aligning this sequence to intergenic regions of the other bacteria. The maximum e-value were set to 1. The output contains, as a maximum, the 10 best alignments for every one of the eleven related bacteria. The output also contains the actual alignment of these hits.

An example of the output from this alignment and a link to all the alignments can be found in section 7.3 on page 75.

Chapter 7

Results

The results of this study can be divided into several parts.

- The development of a new promoter score function taking advantage of the data collected and extracted from previous work to produce logo plots and corresponding score tables.

- The actual program implementing the novel promoter location scoring algorithm, combining it with the previously developed terminator search algorithm.
  At http://folk.uio.no/gardt/Hovedfag/index.html the search program developed in this study can be downloaded along with the necessary BioJava packages. At this site one can also download the Java code, JavaDoc for the program and also the file containing the intergenic regions of *E.coli* that were used in this study.

- The estimates on the number of hits in the random case, when random means a random sequence that conserves some of the most important features of the search string.

- Analysis of how many of the previously verified and suggested ncRNAs that are detected by this program.

- An example of the data produced when candidates have been aligned with intergenic regions of related bacterias. All alignments can be found from an URL given in section 7.3.

- A list containing the 18 suggested candidates for novel ncRNA genes.

In this chapter the results not covered in previous sections will be presented.

7.1 Developing a new promoter score function using logo-plots

There are 62 verified and a little more than 1000 predicted ncRNAs in *E.coli* (Hershberg et al., 2003; Vogel et al., 2003).

To produce logo plots there are many good tools on the Internet. For this specific task the WebLogo plotter was chosen, which is made by a research group at Berkeley, USA (Crooks GE and SEs, 2004; Schneider and Stephens, 1990). The WebLogo program was used to create the logo plots in this study where the multiple sequence alignments were accessible.

A logo plot for each of the 8 implemented consensus sequences was created by using the WebLogo plotter. For 5 of the plots (the five plots with data not from the Regulon database) the actual sequences were not accessible, but the multiple sequence alignment profiles were. For these 5 latter plots the positions outside the actual consensus sequence were unknown and therefore assigned an information value of zero. The data of the plots from Pedersen et al. (2000) and Ussery (1999) were extracted by measuring directly on the plots. This is the reason why the numbers in the scoring function might look a bit "constructed". The size of the datasets connected to every consensus sequence can be found in table 4.2.

By creating these logo plots and the according score tables at least three new weighted consensus sequences have been suggested for the *E.coli* polymerase, this will hopefully open doors to locate novel ncRNA genes in the *E.coli* genome.

The logo plots and the corresponding score tables of the 8 consensus sequences available for use in this program are found on the following pages.

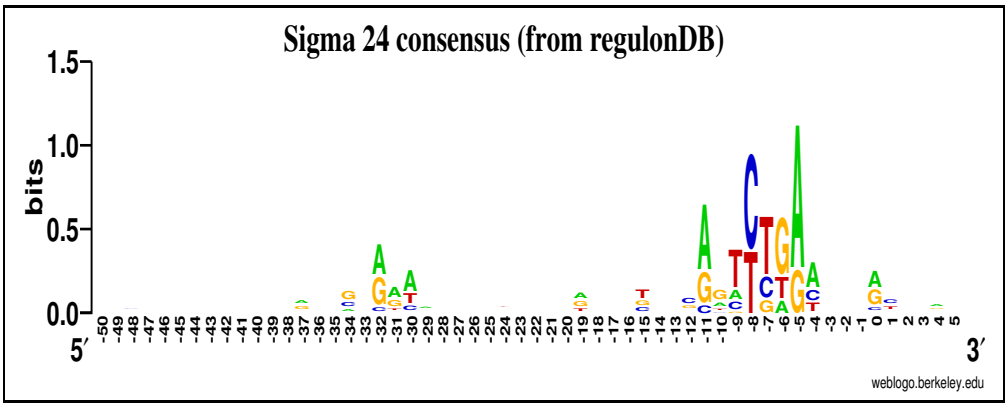

Figure 7.1: Logo plot of 33 sequences recognized by σ^{24}, dataset from the regulonDB (Salgado et al., 2000).

Consensus -35	**G**	**G**	**A**	**A**	**A**	**A**
Information value	0.17	0.08	0.37	0.18	0.26	0.10
Frequency of A	19	27	43	43	49	41
Frequency of T	11	19	5	16	27	24
Frequency of C	27	16	11	11	16	19
Frequency of G	43	38	41	30	8	17

Consensus -10	**A**	**G**	**T**	**C**	**T**	**G**	**A**	**A**
Information value	0.53	0.17	0.36	0.72	0.47	0.47	0.86	0.29
Frequency of A	57	24	19	3	3	16	70	49
Frequency of T	3	16	57	38	57	24	3	21
Frequency of C	11	14	16	57	24	3	3	24
Frequency of G	30	46	8	3	16	57	24	5

Table 7.1: σ^{24} consensus sequence, -35 and -10 region, dataset from the Regulon database (Salgado et al., 2000)

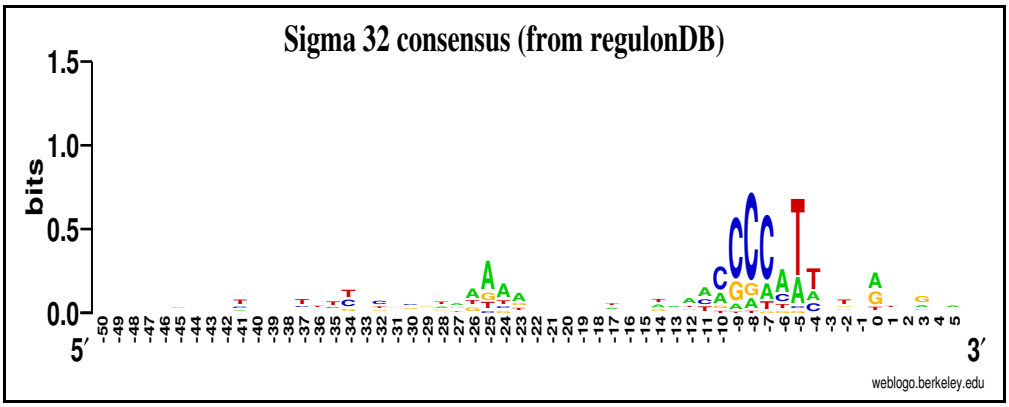

Figure 7.2: Logo plot of 48 sequences recognized by σ^{32}, dataset from the Regulon database (Salgado et al., 2000).

Consensus -35	**T**	**A**	**A**	**A**	**A**	**A**
Information value	0.14	0.13	0.20	0.35	0.23	0.19
Frequency of A	31	37	44	56	50	46
Frequency of T	37	27	25	17	21	21
Frequency of C	12	12	10	10	15	13
Frequency of G	23	27	23	19	15	21

Consensus -10	**C**	**C**	**C**	**C**	**A**	**T**	**T**
Information value	0.32	0.57	0.70	0.58	0.31	0.66	0.30
Frequency of A	27	10	12	19	54	25	25
Frequency of T	12	6	6	12	15	65	48
Frequency of C	52	63	71	65	21	6	23
Frequency of G	12	23	13	6	12	6	6

Table 7.2: σ^{32} consensus sequence, -35 and -10 region, dataset from the Regulon database (Salgado et al., 2000).

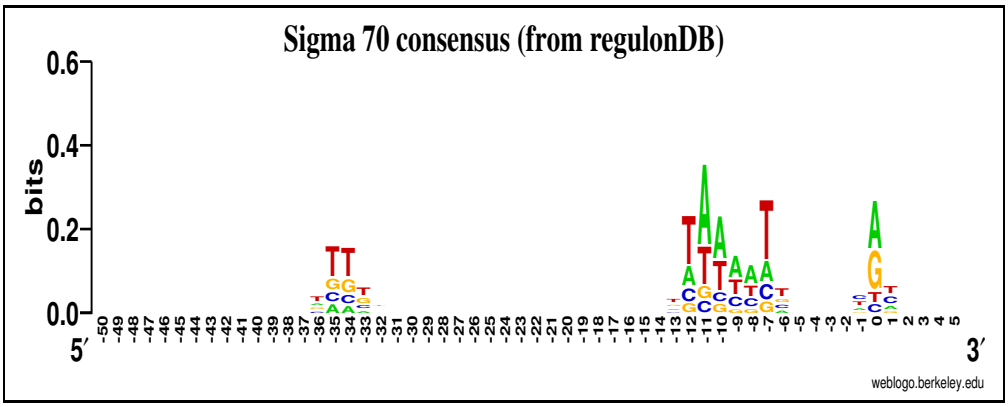

Figure 7.3: Logo plot of 4379 sequences recognized by σ^{70} dataset from the Regulon database (Salgado et al., 2000).

Consensus -35	**T**	**T**	**T**	**T**	**T**	**T**
Information value	0.01	0.05	0.16	0.16	0.06	0.02
Frequency of A	24	22	17	14	17	21
Frequency of T	31	36	47	46	36	32
Frequency of C	22	20	17	17	19	26
Frequency of G	24	21	19	23	27	21
Consensus -10	**T**	**A**	**A**	**A**	**A**	**T**
Information value	0.24	0.36	0.24	0.14	0.12	0.28
Frequency of A	22	54	45	40	41	20
Frequency of T	51	26	32	29	26	53
Frequency of C	16	9	12	20	21	15
Frequency of G	12	10	11	11	13	12

Table 7.3: σ^{70} consensus sequence, -35 and -10 region, dataset from the Regulon database (Salgado et al., 2000)

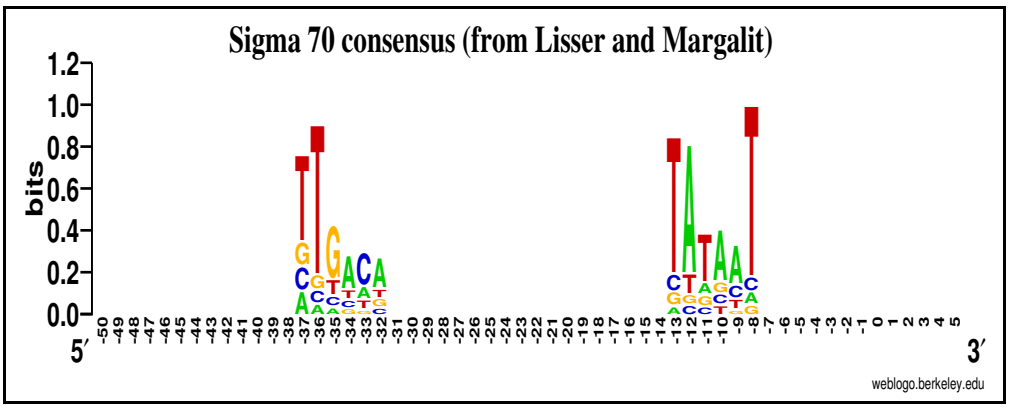

Figure 7.4: Logo plot of 298 sequences recognized by *E.coli* polymerase, dataset from Lisser and Margalit (1993)

Consensus -35	**T**	**T**	**G**	**A**	**C**	**A**
Information value	0.87	0.85	0.40	0.42	0.35	1.04
Frequency of A	10	6	9	56	21	54
Frequency of T	69	79	18	16	16	17
Frequency of C	10	7	12	17	54	13
Frequency of G	10	8	61	11	9	16
Consensus -10	**T**	**A**	**T**	**A**	**A**	**T**
Information value	0.61	0.93	0.44	0.32	0.31	0.28
Frequency of A	5	76	15	61	56	6
Frequency of T	77	12	60	12	15	82
Frequency of C	10	6	11	13	20	7
Frequency of G	8	6	14	14	8	5

Table 7.4: σ^x consensus sequence, -35 and -10 , dataset from Lisser and Margalit (1993)

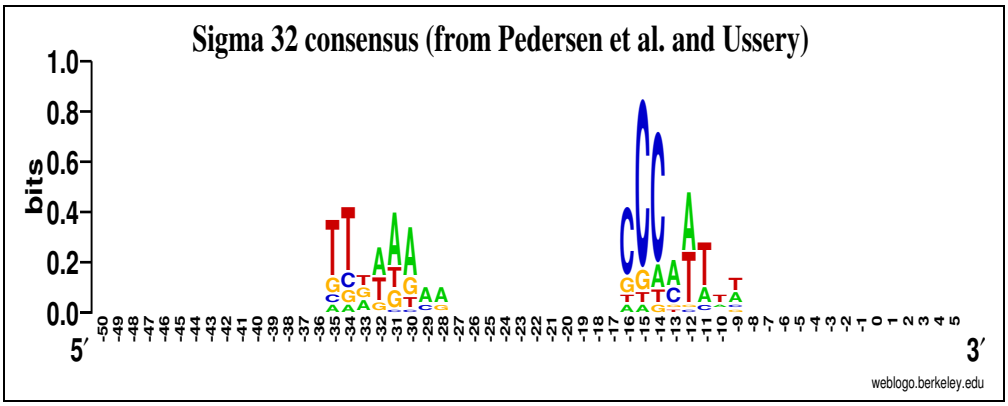

Figure 7.5: Logo plot of 75 sequences recognized by σ^{32}, data from Pedersen et al. (2000) and Ussery (1999).

Consensus sequence	**T**	**T**	**T**	**A**	**A**	**A**	**A**	**A**
Information value	0.37	0.43	0.16	0.27	0.38	0.35	0.12	0.12
Frequency of A	10.8	9.3	31.3	44.4	57.9	57.9	58.3	58.3
Frequency of T	62.2	60.1	31.3	37.0	23.7	14.3	8.3	8.3
Frequency of C	10.8	16.3	6.3	3.7	5.3	5.7	25.0	0.0
Frequency of G	18.9	14.0	3.1	14.8	13.2	22.9	0	25.0
Consensus sequence	**C**	**C**	**C**	**A**	**A**	**T**	**T**	**T**
Information value	0.43	0.86	0.73	0.22	0.49	0.29	0.08	0.15
Frequency of A	9.3	5.0	13.7	50.0	49.0	24.1	25.0	33.3
Frequency of T	9.3	5.8	8.2	9.1	42.9	62.1	50.0	40.0
Frequency of C	65.1	79.1	72.6	31.8	4.1	10.3	12.5	13.3
Frequency of G	16.3	10.5	5.5	9.1	4.1	3.4	12.5	13.3

Table 7.5: σ^{32} consensus sequence, -35 and -15 region, data from Pedersen et al. (2000) and Ussery (1999).

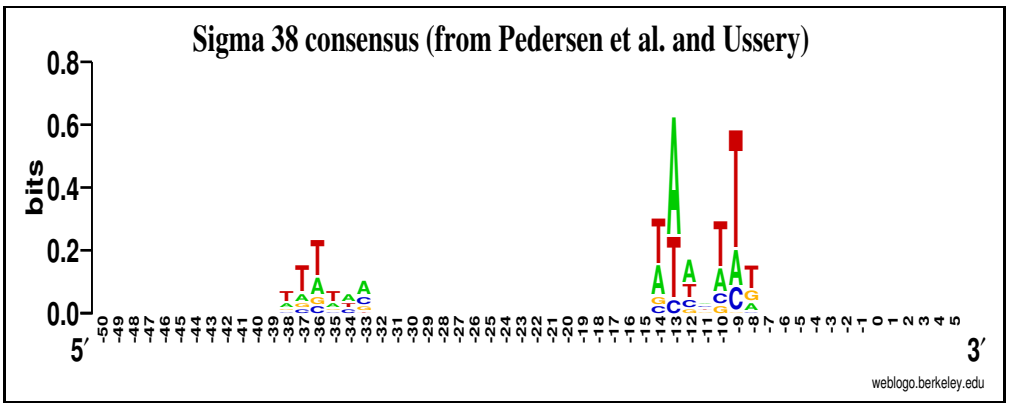

Figure 7.6: Logo plot of 68 sequences recognized by σ^{38}, data from Pedersen et al. (2000) and Ussery (1999).

Consensus sequence	**T**	**T**	**T**	**T**	**A**	**A**
Information value	0.08	0.16	0.24	0.08	0.07	0.11
Frequency of A	25.0	18.6	25.0	25.0	42.3	45.5
Frequency of T	50.0	56.3	50.0	50.0	28.6	9.1
Frequency of C	12.5	12.5	12.5	12.5	28.6	27.3
Frequency of G	12.5	12.5	12.5	12.5	0.0	18.2

Consensus sequence	**T**	**A**	**A**	**A**	**T**	**T**	**T**
Information value	0.30	0.63	0.18	0.04	0.30	0.59	0.15
Frequency of A	33.3	60.3	44.4	25.0	26.7	20.3	20.0
Frequency of T	50.0	31.7	27.8	25.0	50.0	64.4	53.3
Frequency of C	8.3	7.9	16.7	25.0	13.3	13.6	6.7
Frequency of G	8.3	0.0	1.1	25.0	10.0	16.9	26.7

Table 7.6: σ^{38} consensus sequence, - 35 and -10 region, data from Pedersen et al. (2000) and Ussery (1999).

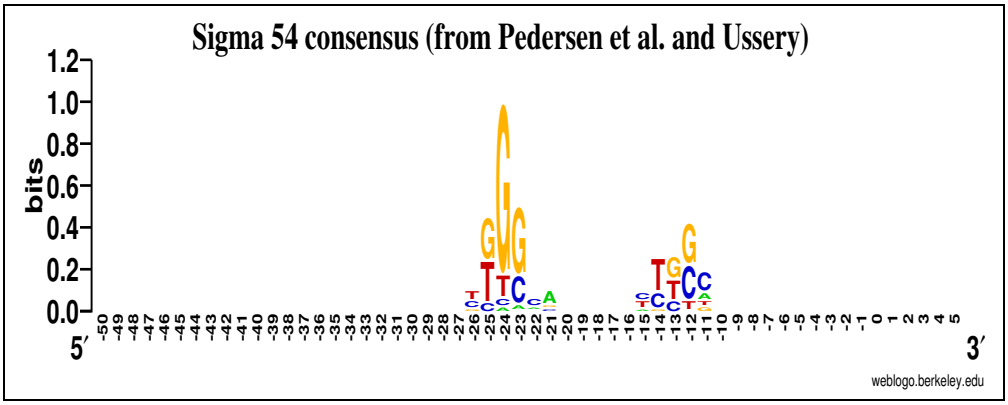

Figure 7.7: Logo plot of 177 sequences recognized by σ^{54}, data from Pedersen et al. (2000) and Ussery (1999).

Consensus sequence	T	G	G	G	C	A
Information value	0.11	0.46	1.00	0.51	0.07	0.11
Frequency of A	0.0	0.0	3.0	5.9	28.6	63.6
Frequency of T	45.5	43.4	11.0	2.0	0.0	0.0
Frequency of C	36.4	10.9	4.0	27.5	57.1	18.2
Frequency of G	18.2	45.7	82.0	64.7	14.3	18.2

Consensus sequence	C	T	G	G	C
Information value	0.10	0.26	0.27	0.43	0.20
Frequency of A	20.0	0.0	0.0	2.3	20.0
Frequency of T	40.0	61.5	37.0	11.6	15.0
Frequency of C	40.0	30.8	22.2	39.6	50.0
Frequency of G	0.0	7.7	40.7	46.5	15.0

Table 7.7: σ^{54} consensus sequence, -24 and -12 region, data from Pedersen et al. (2000) and Ussery (1999).

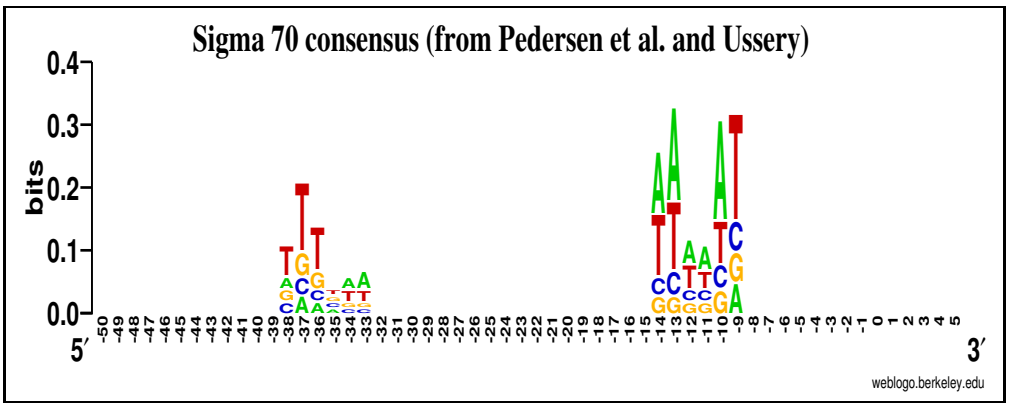

Figure 7.8: Logo plot of 3851 sequences recognized by σ^{70}, data from Pedersen et al. (2000) and Ussery (1999).

Consensus sequence	**T**	**T**	**T**	**T**	**A**	**A**
Information value	0.11	0.20	0.13	0.04	0.06	0.07
Frequency of A	18.2	12.5	11.5	25.0	33.3	42.8
Frequency of T	45.5	55.0	53.8	25.0	33.3	28.6
Frequency of C	18.2	12.5	11.5	25.0	16.7	14.3
Frequency of G	18.2	18.0	23.1	25.0	16.7	14.3
Consensus sequence	**A**	**A**	**A**	**A**	**T**	**T**
Information value	0.25	0.33	0.12	0.11	0.31	0.32
Frequency of A	40.0	45.5	33.3	36.4	22.6	15.6
Frequency of T	40.0	33.3	33.3	27.2	51.2	53.1
Frequency of C	10.0	12.1	16.7	18.2	12.9	15.6
Frequency of G	10.0	9.1	16.7	18.2	12.9	15.6

Table 7.8: σ^{70} consensus sequence, -35 and -10 region, data from Pedersen et al. (2000) and Ussery (1999).

Promoter type	Threshold (%)	Estimated # of hit	Actual # of hits
$\sigma^{70}a$	100	6.7	24
$\sigma^{32}a$	95	0.2	1
σ^{54}	98	0.7	8
σ^{38}	97	0.9	9
σ^{x}	100	0.06	1
$\sigma^{70}b$	100	0.61	2
$\sigma^{32}b$	96	0.18	1
σ^{24}	97	0.04	1

Table 7.9: This table holds a comparison of the hit ratio for this actual program compared to what is expected from the random case.

7.2 Comparing results from this study with the random case

In table 7.9 an overview of the number of estimated and actual hits per promoter consensus sequence when using a given threshold is presented. By comparing the expected and actual number of hits, it seems like most hit ratioes except from $\sigma^{70}a$ and $\sigma^{70}b$ are significant. But even these have a number of hits about three times as high as what was expected. These results are discussed further in chapter 8 on page 83.

7.3 Aligning candidates to intergenic regions in related bacterias

The alignments produced as described in section 6.3 on page 63 can be found at this URL: http://folk.uio.no/karinlag/gard/HTML/ .
The sequences used in the respective alignments is easily read from the filenames. The files are named after which σ factor that recognizes it, when necessary the the σ factor name is followed by an “R” or a “D” to separate the ones origining from the Regulon database (Salgado et al., 2000) and the ones from Pedersen et al. (2000) and Ussery (1999).

A slightly simplified example from an alignment considered as good is included on the next page. In the example the first line contains the name and data the intergenic region the candidate is located in. The next section shows whether or not this intergenic region contains any of the 62 known ncRNAs. Then there is a section with alignment hits between this candidate and genes in other bacteria. The next section is about the actual hits between this candidate and intergenic regions the other

bacteria. Finally the actual alignment of the intergenic hits is included.

```
b1837/pphA_to_b1839_1920997_1921388.reversed
IG region contains the following ncRNAs:
sraC/ryeA/tpke79/IS091 with direction f, at
position 1921090-1921338(94-342)
This ig region matches these genes in these
genomes:

Genome                          Gene name       Evalue
Salmonella_typhimurium_LT2      -               3e-27
Buchnera_aphidicola_Sg          lspA            0.27
Shigella_flexneri_2a_2457T      mutS            0.008
Salmonella_typhi                -               2e-25
Shigella_flexneri_2a            mutS            0.008


Info about the hits:

1_cYersinia_pestis_KIM_2783897_2783967          evalue 6e-22
2_cYersinia_pestis_KIM_2279677_2279747          evalue 0.49
1_Salmonella_typhi_Ty2_1092686_1092756          evalue 2e-25
1_Yersinia_pestis_CO92_2032736_2032806          evalue 6e-22
2_cYersinia_pestis_CO92_2503417_2503487         evalue 0.5
1_cShigella_flexneri_2a_2457T_1856205_1856275   evalue 2e-34
1_cShigella_flexneri_2a_1887011_1887081         evalue 2e-34


CLUSTAL X (1.81) multiple sequence alignment

1_cYersinia_pestis_KIM_2783897 ---------TAAGCCTACAT-TAATACC
1_Yersinia_pestis_CO92_2032736 AAGTGCGGCTGAATAAGCCTACAT-TAATACC
b1837/pphA_to_b1839_1920997_19 AGGGCAAGGCAACTAAGCCTGCAT-TAATGCC
1_cShigella_flexneri_2a_2457T_ AGGGCAAGGCAACTAAGCCTGCAT-TAATGCC
1_cShigella_flexneri_2a_188701 AGGGCAAGGCAACTAAGCCTGCAT-TAATGCC
1_Salmonella_typhi_Ty2_1092686 AGAGCAAGGCGATTTAGCCTGCAT-TAATGCC
2_cYersinia_pestis_KIM_2279677 -GCTGTACCGGTCCTATTTACGTCATACCGCC
2_cYersinia_pestis_CO92_250341 -GCTGTACCGGTCCTATTTACGTCATACCGCC

AACTTTTAGCGCACGGCTCTCT-CCCAAGAGCAGCCATTTCCCTAGACCGAATATCA
AACTTTTAGCGCACGGCTCTCT-CCCAAGAGCCATTTCCCT---------
AACTTTTAGCGCACGGCTCTCT-CCCAAGAGCCATTTCCCT---------
AACTTTTAGCGCACGGCTCTCT-CCCAAGAGCCATTTCCCT---------
AACTTTTAGCGCACGGCTCTCT-CCCAAGAGCCATTTCCCT---------
AACTTTTAGCGCGTTTTTTCATGCCTCATTATGAATTTTT----------
AACTTTTAGCGCGTTTTTTCATGCCTCATTATGAATTTTT----------
```

Promoter type	# candidates	Promoter hit	Terminator hit	Double hit
$\sigma^{70}a$	125	17	7	0
$\sigma^{32}a$	52	3	9	0
σ^{54}	74	10	4	0
σ^{38}	87	9	5	0
σ^{x}	67	3	5	1
$\sigma^{70}b$	162	15	17	0
$\sigma^{32}b$	79	0	6	0
σ^{24}	96	7	8	0
In total	742	64	61	1

Table 7.10: Distribution of hits in this study compared to dataset of 1056 predicted ncRNAs Hershberg et al. (2003). The thresholds are the same as in table 6.1.

7.4 Comparing new and previous candidates

The ncRNA candidates from this program (when using the thresholds from 6.1 on page 56) have been compared to 52 known ncRNAs (Hershberg et al., 2003; Vogel et al., 2003), and to 1056 predicted ncRNAs (Hershberg et al., 2003). Three types of hits have been defined.

- *Promoter hit* - a hit where the suggested candidate from this program has its first nucleotide-position less than 25 bases upstream and no more than 15 bases downstream of the ncRNA it is compared to.

- *Terminator hit* - a hit where the suggested candidate from this program has its last nucleotide-position less than 40 bases upstream and no more than 40 bases downstream of the ncRNA it is compared to.

- *Double hit* - a hit that is both a promoter hit and a terminator hit.

In the two tables 7.10 and 7.11, the distribution of hits from the candidates suggested by this new program is compared to the 1056 predicted ncRNAs and the 52 verified ncRNAs respectively. There seems to be only a small amount of overlap between the ncRNA candidates found in this study and previous studies. This implies that the known and suggested ncRNAs have weak transcriptional signals, or that the search criteria of this study are unsuitable.

Promoter type	# candidates	Promoter hit	Terminator hit	Double hit
$\sigma^{70}a$	125	1	0	0
$\sigma^{32}a$	52	0	0	0
σ^{54}	74	1	0	0
σ^{38}	87	0	0	0
σ^{x}	67	2	1	0
$\sigma^{70}b$	162	0	3	0
$\sigma^{32}b$	79	0	0	0
σ^{24}	96	2	5	0
In total	742	6	9	0

Table 7.11: Distribution of hits in this study compared to dataset of 52 known ncRNAs Hershberg et al. (2003); Vogel et al. (2003). The thresholds are the same as in table 6.1.

7.4.1 Testing program on verified ncRNA sequences

One natural question to ask is "Why is the program locating so few of the known ncRNAs"? To answer this the 70 nucleotides upstream and 60 nucleotides downstream of 55 of the known ncRNAs were cut from the DNA string and put into two separate files. In the file containing the promoter regions the program ought to find, only 6 promoters were found. In the terminator file only 6 terminators were found. Moreover there was not one double hit, that is a promoter located upstream and a terminator located downstream of the same ncRNA. This is closely related to the findings above, and suggests exactly the same. Either the transcription signals are weak or the search criteria are unsuitable. This introduces an idea to possible further work; The regions upstream and downstream of the known and suggested ncRNAs should be aligned to search for other, or new signals or consensuses. This problem of not detecting the already known ncRNAs was already suggested when the promoter consensus sequences stated in this study turned out as weak as they were.

7.5 Suggested ncRNA candidates from this study

To give a final suggestion of ncRNA candidates, the candidates produced while having a high threshold where chosen. Then the alignments of these candidates to other related bacterial intergenic regions were studied. The result was a list of 20 ncRNA candidates. Of these candidates, two lie inside the coding regions of already verified ncRNAs. Candidates being part of repetitive sequences have been removed. This list is

found in table 7.12, and the special cases mentioned above are indicated. All these candidates have been tested using Rfam (Griffiths-Jones et al., 2003). Rfam is a database of all known ncRNA families and it has a search function to test your nucleotide sequence against their database. The only hits achieved where the two special situations where the suggested candidate lies inside a verified ncRNA.

Intergenic region	Length	Al. [a]	σ factor	Start	Stop	D. [b]
dapB to carA	88	7	σ^{70}a	29317	29407	→
cysB to acnA	116	7	σ^{70}a	1332972	1333088	→
sseA to sseB	211	9	σ^{70}a	2651401	2651635	←
yeiN/yeiC to fruA	44	4	σ^{70}a	2257650	2257694	→
ykgD to ykgE	112	6	σ^{32}a	320456	320570	←
malP to malT	75	2	σ^{32}a	3550252	3550319	←
yciN to topA	58	5	σ^{54}(1)	1328870	1328928	→
kdsA to chaA	184	5	σ^{54}(1)	1268269	1268455	←
livK to yhhK	109	8	σ^{38}(1)	3595369	3595478	→
glnS to ybfM	166	7	σ^{38}(1)	707084	707252	←
kdpA to ybfA	210	9	σ^{38}(1)	728026	278238	←
bax to malS	200	5	σ^{38}(1)	3734852	3735052	→
ykgM to eaeH	121	7	σ^{38}(1)	313369	313490	→
yhiW to yhiX *	152	8	σ^{X}(2)	3662422	3662574	→
b1688 to b1689	67	7	σ^{32}b	1768418	1768487	←
tra8_3 to b4285	220	9	σ^{32}b	4506210	4506430	→
yhaO to yhaP	41	4	σ^{24}(3)	3255803	3255844	→
b1837/pphA to b1839 **	71	5	σ^{24}(3)	1921161	1921234	←
purL to yfhD/yfhC	94	6	σ^{24}(3)	2693634	2693728	→
ugpB to livF/livG/livM	103	6	σ^{24}(3)	3590165	3590268	→

Table 7.12: This table holds information about the suggested ncRNA candidates from this study.
[a]) Number of bacteria having a fairly similar sequence
[b]) The direction of the candidate
*) This candidate lies inside the coding region of the verified ncRNA IS183 with direction →, at position 3662492-3662604(244-356). According to Rfam this candidate has structural similarities to this ncRNA, IS183.
**) This candidate lies inside the coding region of the verified ncRNA sraC/ryeA/tpke79/IS091 with direction →, at position 1921090-1921338(94-342). According to Rfam this candidate has strucuralk similarities to the verified ncRNA ryeB.

Chapter 8

Discussion and conclusion

In this study a modified version of a previously tested method for locating transcription signals has been applied to locate ncRNAs in the *E.coli* genome. The modification is mainly the novel promoter sequence score function. In this chapter the results and findings from this study will be discussed and followed by a conclusion about this study.

8.1 Discussion

The program created in this study does not find many of the promoter and terminator sequences of the previously known or suggested ncRNAs. This is because these ncRNAs do not have a promoter and/or a terminator sequence that has the features needed to be detected by the program. This implies that the program does not have optimal search criteria, and/or that the transcription signals of ncRNAs might differ from or not be as conserved as for most mRNAs.

According to table 7.9 the search criteria used in the promotersearch in this study give significantly more hits than one would expect from the random case. This goes for most of the promoter consensus sequences that are searched for. This should imply that the consensus sequences searched for contain some kind of information, which is the reason why they are conserved.

The low number of hits on the already verified ncRNAs is explained by their lack of conserved promoter sequences and terminator structures that are located by this search program (see table 7.11). This indicates that the search criteria implemented neither represent the consensus of the promoters of these ncRNAs nor the structure of their terminators. Further work should therefore include looking for novel transcription signals in previously known and suggested ncRNAs. However, the promoter and terminator search criteria implemented do represent promoters and terminators of known mRNAs, and this implies a new

question: "Do ncRNAs have different promoter and terminator features, or are promoters and terminators less important to the transcription process than previously thought"? To create a program that includes a higher percentage of the known and verified ncRNAs in its candidate list the program should have had implemented an algorithm that has been trained on the promoter and terminator sequences of these candidates. This might be done by applying a neural network, creating consensuses sequences based upon known ncRNAs promoter sequences, using context free grammars to create a model of the terminators of these ncRNAs or apply a version of the Markov model to create a model to fit transcription signals into. On the other hand this program has implemented search criteria to five of the seven σ factors in *E.coli*. As only the features of the σ^{70} factor have been used as search criteria previously, the program might actually present new and interesting candidates.

Approximately 850 ncRNAcandidates have been aligned with intergenic regions from bacteria related to *E.coli*. By combining these alignments with the candidates having the best promoter and terminator score 20 ncRNAcandidates have been suggested (see table 7.12 on page 81). Of these candidates, two are located inside already verified ncRNAs. Candidates having a sequence that is part of repetitive intergenic regions of *E.coli* have also been removed. This leaves 18 candidates for novel ncRNAs suggested in this study, and many more that should be examined in detail during further work. Of these 18 candidates none have a structure recognized by the Rfam database, which contains all known ncRNA structures (Griffiths-Jones et al., 2003). This suggests that the candidates from this study either belong to novel classes of ncRNA or they are false. The two suggested candidates that lie inside verified ncRNAs were recognized by Rfam.

The quality of these candidates is very hard to estimate by computational approaches, so they should be tested in a laboratory. Such a test will first of all check whether the sequence is transcribed.

Further work to improve the program and finding better search criteria is discussed in chapter 9.

8.2 Conclusion

This study has created a novel scoring function for promoters. The score function is based upon known or predicted promoter sequences, mostly from mRNAs. This score function has been implemented in a search algorithm, and it has been combined with a previously used algorithm to locate ρ-independent terminators (Ermolaeva et al., 2000). The resulting ncRNA detection program suggestes several ncRNA candidates. The number of candidates suggested is highly dependent upon the threshold

used with the promoter consensus score function.

The implemented program does not have a high hit ratio on neither the 62 known nor the more than 1000 unverified ncRNAs. This suggests that the search criteria of the program are insufficient and/or wrong, or that ncRNAs do not have conserved promoter sequences and/or terminator structures.

It should be noticed that in the previous studies(see section 2.4 on page 20), where transcription signals have been the main idea behind the searches, the search has in most cases been combined with microarray experiments. So these studies have had the possibility of operating with weaker thresholds because they could exclude candidates based upon microarray results.

Analyses of the hit ratios on a random intergenic DNA string have shown that the search criteria tend to be significant, but not very strong.

A final selection from the candidates has been made by aligning more than 850 candidates with intergenic regions of 11 bacteria closely related to *E.coli*. 20 candidates having both strong transcription signals and a strong conservation (low e-values) have been chosen as final candidates (see table 7.12 on page 81). Of these 20 candidates 18 are not part of any previously verified ncRNAs. It should be noted that several other candidates also made good alignments with intergenic regions of some or all of the other related bacteria. These candidates also deserve closer examination.

All 18 candidates lack structural similarities to known ncRNA families. This implies detection of novel ncRNA families or false candidates. To verifiy or falsify these candidates they should be tested in a laboratory.

The conclusion about the actual search program is that the program finds what it looks for, but the criteria of what to look for seem too weak. Further research in this field will establish better search criteria and then also a more accurate program. If this program is to be used as part of a larger program for finding ncRNAs (see section 1.1 on page 1 and section 9.4 on 89), the thresholds will have to be set at a liberal value, but a good initial selection of candidates by this program will nevertheless represent a remarkable time saver since level two of the larger program has a complexity of n^6.

As an independent program for ncRNA detection, this program is not particularly well suited as of today, however, combined with other analyses it might represent a useful tool. The answer to this question will be given when the 18 novel ncRNA candidates are tested in the laboratory.

This study leaves many questions, one of the most interesting being: “If promoters and terminators of the known and suggested ncRNAs remain undetected in a search for transcription signals, are there then any transcription signals to these genes that remain unknown”?

Chapter 9

Improvements and further work

This study has had a time limit of one and a half year, and it is limited how much one can implement and test during this period. In this chapter possible improvements and ideas are discussed.

9.1 Refining the promoter search

As more and more coding sequences are experimentally verified, their corresponding promoter sequences will also be verified. Then a larger dataset could be collected to describe the consensus sequence of the promoters. This might also provide more detailed criteria to the search that might exclude false positives.

Further research on the topic of promoters will also reveal whether a promoter search on a DNA string really is a good method to search for novel genes. However, the significance of the promoter sequence has been verified on the DNA sequence, so the current search is more than a wild-goose chase.

The most important part of a refined search will most likely be to get a better defined consensus sequence, that is, higher information values on the nucleotide positions, and also include other transcription signals that today do not have enough significance compared to the random case.

9.2 Refining the terminator search

The search algorithm of the terminator search of this program has previously been used as part of a larger terminator search algorithm (Ermolaeva et al., 2000). The part implemented here was the initial part of

the algorithm, i.e. it included about every good terminator candidate, but also included many false positives. One of the criteria to this program was that the input sequence should be a DNA sequence, and that was all the program was supposed to know concerning the sequence. Therefore the rest of the algorithm could not be implemented, because more knowledge of the sequence is needed. In addition to this the ρ-dependent terminators are not searched for by this program.

A further development of this program would naturally be to demand more sequence information, that is sequence coordinates, to be able to implement the rest of the scoring algorithm used by Ermolaeva et al. (2000). The program should also in some way try to include a search for ρ-dependent terminators. These terminators are by less common than the intrinsic terminators in *E.coli* (Lewin, 2000), but still essential to the bacteria (Richardson, 2002). Therefore a ρ-dependent terminator search function should be implemented. Such a function might become very useful when searching DNA from other bacteria.

The terminator search might also be improved by looking at the scoring function. First of all, as previously mentioned, the parameters of the scoring function look a bit peculiar. This might origin in a common problem with decision trees; they might actually produce an answer that fits the particular model very well, but when used to describe situations slightly different from the model, they fail. Secondly, there are two separate thresholds used in the terminator search. Presently a terminator has to score above both thresholds separately, it might be an idea to unite these thresholds. This will for example enable a strong hairpin with a weak tail to pass the test. The reason behind these suggestions to improvements is the fact that the termination signals of the previously known and suggested ncRNAs are weak, and that these improvements might make them easier to detect.

9.3 Speeding up the program

The program has a structure that makes it very easy to convert into a parallel program. The search is first of all divided into two main parts: promoter search and terminator search. The promoter search is easily divided in two, the search on the input string and on the complementary. Then every single input sequence could be searched in parallel, and every search for the different consensus sequence could also be executed in parallel.

The terminator search could also be divided in two, according to the input string and the complimentary string. As for the promoters every input string could be searched in parallel. Also if one wishes to search many times in the same material with different threshold values or dif-

ferent promoter consensus sequences, the terminator candidates should be stored to avoid doing the same search over again. This will give significant results concerning the runtime of a second iteration of the program. As seen in section 5.6 on page 52 it would speed up the program with a factor of about 6.
The final part of the search, where the program looks for a promoter - terminator match, could also be run in parallel for every sequence and direction.
If for example, there are 10 input sequences, and the search is for 5 of the different promoter consensus sequences, the promoter search could be run in parallel on
$10 \cdot 2 \cdot 5 = 100$ processors.
The terminator search would only need $10 \cdot 2 = 20$ processors,
and the final search would need $10 \cdot 2 = 20$ processors.
This gives an initial search using 120 processors in parallel and a final search using 20 processors.

A parallelization like this would speed the program significantly, and is fairly easily conducted in a super-computer environment, or in a grid system like Condor (see section 5.1 on page 49). A speed-up like this opens up for far more complicated search criteria without a remarkable increase in the computational time.

9.4 Further work

First of all this program should be speeded up accoring to the outline described above. If the number of processors is insufficient, the program should be speeded up by using as many processors as possible. Moreover, it should be possible to import your own consensus sequence to search for, this is because the consensus sequences of the promoters might be different from bacteria to bacteria.

Secondly, this program is meant to become part of a larger three level program dedicated to search for ncRNAs. The program described in this study is meant to be the initial search of the input DNA, and is supposed to pick out a range of candidates allowing a fairly high level of false positives. The idea behind this initial part of the search is to avoid running very time consuming computational calculations on parts of the input DNA that can be marked as uninteresting by this program.

The idea of the second level of this larger program is to try to match a candidate sequence with a secondary structure that an ncRNA is likely to have. ncRNAs are dependent on a stable secondary structure to become a stable functionale molecule, and the secondary structure is better conserved than the primary structure. Consequently a basepair may be exchanged with another basepair, or the positions of the bases could

be switched, still keeping the same secondary structure, but having a primary structure insufficiently similar to other ncRNAs in related species. This indicates that a similarity search would not recognize the novel ncRNA because of the lack of primary structure similarity. A problem with this search strategy is that it only recognizes structures similar to already known classes of RNA, thus novel types with novel structures will not be detected by this search.

The third part of the search would do the same as the previously mentioned program used in this study to locate conserved regions in DNA of other related bacteria, compared to the candidate sequences. This third level of the program would both process the data from the first level and also the candidates from the second level, having those as a favorite. A search for homologues exclusively in the sequences suggested by level 2 (originating in level 1) is not an optimal solution, this is because there might be sequences with a high conservation and strong transcription signals, but with an unknown secondary structure. This is the case when a novel ncRNA with a new secondary structure is detected, as for example if any of the 18 suggested candidates from this study actually is an ncRNA.

Further work would therefor also include uniting these three levels and make them work together. Then a user interface will have to be created, and the whole program should be made accessible for other users by making it a web application.

Appendix A

Definitions

In this chapter certain words and expressions from molecular biology will be explained. This will be of use to the reader unfamiliar to basic molecular biology.

- *5' and 3'* - these are the names of the ends of a DNA or RNA strand. A piece of DNA or RNA is read from the 5' to the 3' end. The names are read "five prime" and "three prime". The names stem from the naming of atoms in the sugar ring in the DNA backbone.
- *cDNA* (complementary DNA) - is DNA produced from a RNA transcript by using reverse transcription.
- *Codon* - a codon is a triplet of nucelotides that codes for a special signal, usually it is a signal for which aminoacid it is translated into. There are also codons that tell the translation machinery when to start and stop. Amino acids are the building blocks of protein.
- *Conserved* - a nucleotide is conserved if the same nucleotide is found at the same position in the same gene in several genomes of related species. Important parts of the genome tend to be conserved.
- *Consensus sequence* - a consensus sequence contains the most common symbol in each position of the sequences in the set.
- *E-value* - an e-value is a number describing the number of expected hits in the random case. Often used to describe the quality of alignments.
- *FASTA* - is a file format widely used for files containing gene sequences. The file is formatted as this:
  ">{sequence name}{sequence annotation}(max one line)
  {the actual sequence, 60 bases per line}"

- *fRNA* - is a common name used for functional RNA, includes small RNA (sRNA), small nucleolar RNA (snoRNA) and many more types. Widely used for all RNA except for mRNA, rRNA and tRNA (Carter et al., 2001).
- *Hybridization* - a hybridization takes place when two single strands of nucleotides connects to each other, due to their base-complementarity, creating a stable double stranded construction.
- *ncRNA* - all types of RNA that are not mRNA.
- *Nucleotide* - nucleotides consist of a base, sugar and phosphate. They are recognizable by the differencies of the bases. There are four different nucleotides A,T,C, and G in DNA. In RNA there are A,C,G and T is replaced by U.
- *Oligonucleotide* - "oligo" for short, is a short (e.g. 20) sequence of consecutive nucleotides formin a small DNA or RNA molecule.
- *Operon* - is a unit of bacterial gene expression and regulation, including structural genes and control elements in DNA recognized by regulator gene products. In other words an operon is a sequence on the DNA that is transcribed as a whole, but contains two or more different coding regions. The transcript is processed into several parts after the transcription.
- *Polymerase* - a polymerase is a large and complex protein that has a very important role in the DNA transcription: it reads the DNA and "produces" RNA.
- *Reading frame* - A sequence of codons of three nucleotides that are located right next to each other, but not overlapping, and on the same strand, are said to be in the same reading frame.
- *Sigma factor* - a sigma factor is one of the five sub-units in the complex polymerase protein.
- *Triplet* - a triplet is three consecutive nucleotides in a string of nucleotides
- *Upstream and downstream* - from a position on a DNA or RNA strand nucleotides on the 5' (lefthand) side of a chosen nucleotide or gene are upstream, accordingly downstream is on the 3' (righthand) side .

Appendix B

The 11 related bacteria used in the alignments

- *Buchnera aphidicola NC* (van Ham et al., 2003)
- *Buchnera aphidicola SG* (Tamas et al., 2002)
- *Buchnera sp AP* (Shigenobu et al., 2000)
- *Salmonella typhi* (Parkhill et al., 2001a)
- *Salmonella typhi Ty2* (Deng et al., 2003)
- *Salmonella typhimurium LT2* (McClelland et al., 2001)
- *Shigella flexneri 2a* (Jin et al., 2002)
- *Shigella flexneri 2a 2457T* (Wei et al., 2003)
- *Wigglesworthia brevipalpis* (Akman et al., 2002)
- *Yersinia pestis CO92* (Parkhill et al., 2001b)
- *Yersinia pestis KIM* (Deng et al., 2002)

Bibliography

Akman, L. , Yamashita, A. , Watanabe, H. , Oshima, K. , Shiba, T. , Hattori, M. , and S., Aksoy . Genome sequence of the endocellular obligate symbiont of tsetse flies, wigglesworthia glossinidia. *Nat Genet*, 32(3): 402–7, 2002.

Argaman, L. , Hershberg, R. , Vogel, J. , Bejerano, G. , Wagner, E.G. , Margalit, H. , and Altuvia, S. . Novel small RNA-encoding genes in the intergenic regions of Escherichia coli. *Curr Biol*, 11(12):941–50, 2001.

Bernal, A. , Ear, U. , and Kyrpides, N. . Genomes online database (gold): a monitor of genome projects world-wide. *Nucleic Acids Research*, 29: 126–127, 2001.

Blattner, F.R. , Plunkett, 3rd, G. , Bloch, C.A. , Perna, N.T. , Burland, V. , Riley, M. , Collado-Vides, J. , Glasner, J.D. , Rode, C.K. , Mayhew, G.F. , Gregor, J. , Davis, N.W. , Kirkpatrick, H.A. , Goeden, M.A. , Rose, D.J. , Mau, B. , and Shao, Y. . The complete genome sequence of Escherichia coli K-12. *Science*, 277(5331):1453–74, 1997.

Burge, C. and Karlin, S. . Prediction of complete gene structures in human genomic dna. *J Mol Biol*, 268(1):78–94, 1997.

Burks, C. , Fickett, J.W. , Goad, W.B. , Kanehisa, M. , Lewitter, F.I. , Rindone, W.P. , Swindell, C.D. , Tung, C.S. , and Bilofsky, H.S. . The genbank nucleic acid sequence database. *Comput Appl Biosci*, 1(4):225–33, 1985.

Carter, R.J. , Dubchak, I. , and Holbrook, S.R. . A computational approach to identify genes for functional rnas in genomic sequences. *Nucleic Acids Research*, 29(19):3928–3938, 2001.

Chen, S. , Lesnik, E.A. , Hall, T.A , Sampath, R. , Griffey, R.H. , Ecker, D.J. , and Blyn, L.B. . A bioinformatics based approach to discover small rna genes in the escherichia coli genome. *BioSystems*, 65:157–177, 2002.

Crooks GE, Hon G, Candonia JM and SEs, Brenner . Weblogo : A sequence logo generator. *In progress* http://weblogo.berkeley.edu/logo.cgi. 2004.

Deng, W. , Burland, V. , Plunkett, G, 3rd. , Boutin, A. , Mayhew, GF. , Liss, P. , Perna, NT. , Rose, DJ. , Mau, B. , Zhou, S. , Schwartz, DC. , Fetherston, JD. , Lindler, LE. , Brubaker, RR. , Plano, GV. , Straley, SC. , McDonough, KA. , Nilles, ML. , Matson, JS. , Blattner, FR. , and Perry, RD. . Genome sequence of yersinia pestis kim. *J Bacteriol*, 184(16):4601–11, 2002.

Deng, W. , Liou, S.R. , Plunkett, G.3rd. , Mayhew, G.F. , Rose, D.J. , Burland, V. , Kodoyianni, V. , Schwartz, D.C. , and Blattner, F.R. . Comparative genomics of salmonella enterica serovar typhi strains ty2 and ct18. *J Bacteriol*, 187(7):2330–7, 2003.

EcoCyc. Ecocyc, http://biocyc.org/ecoli/new-image?object=rna-polymerases. *ASM News*, 70(1):25, 2004.

Eddy, S.R . Profile hidden markov models. *Bioinformatics*, 14(9):755–763, 1998.

Eddy, S.R. . Non-coding RNA genes and the modern RNA world. *Nat Rev Genet*, 2(12):919–29, 2001.

Ermolaeva, M.D. , Khalak, H.G. , White, O. , Smith, H.O. , and Salzberg, S.L. . Prediction of transcription terminators in bacterial genomes. *J Mol Biol*, 301(3836):27–33, 2000.

Griffiths-Jones, S. , Bateman, A. , Marshall, M. , Khanna, A. , and Eddy, S.R. . Rfam: an rna family database. *Nucleic Acids Research*, 31(1): 439–441, 2003.

Henderson, J. , Salzberg, S. , and Fasman, K.H. . Finding genes in DNA with a Hidden Markov Model. *J Comput Biol*, 4(2):127–41, 1997.

Hershberg, R. , Altuvia, S. , and Margalit, H. . A survey of small rna-encoding genes in escherichia coli. *Nucleic Acids Research*, 2003.

Jin, Q. , Yuan, Z. , Xu, J. , Wang, Y. , Shen, Y. , Lu, W. , Wang, J. , Liu, H. , Yang, J. , Yang, F. , Zhang, X. , Zhang, J. , Yang, G. , Wu, H. , Qu, D. , Dong, J. , Sun, L. , Xue, Y. , Zhao, A. , Gao, Y. , Zhu, J. , Kan, B. , Ding, K. , Chen, S. , Cheng, H. , Yao, Z. , He, B. , Chen, R. , Ma, D. , Qiang, B. , Wen, Y. , Hou, Y. , and Yu, J. . Genome sequence of shigella flexneri 2a: insights into pathogenicity through comparison with genomes of escherichia coli k12 and o157. *Nucleic Acids Res*, 30 (20):4432–41, 2002.

Klug, W.S and Cummings, M.R. . Essentials of genetics. *ISBN 0-13-371147-1*, 2nd edition:0–250, 1996.

Kundu, T.A. , Kusano, S. , and Ishihama, A. . Promoter selectivity of escherichia coli rna polymerase of sigma f holoenzyme involved in transcription of flagellar and chemotaxis genes. *Journal of bacteriology*, 179(13):4264–4269, 1997.

Kyrpides, N. . Genomes online database (gold): a monitor of complete and ongoing genome projects world wide. *Bioinformatics*, 15:773–774, 1999.

Lander, ES. , Linton, LM. , Birren, B. , Nusbaum, C. , Zody, MC. , Baldwin, J. , Devon, K. , Dewar, K. , Doyle, M. , FitzHugh, W. , Funke, R. , Gage, D. , Harris, K. , Heaford, A. , Howland, J. , Kann, L. , Lehoczky, J. , LeVine, R. , McEwan, P. , McKernan, K. , Meldrim, J. , Mesirov, JP. , Miranda, C. , Morris, W. , Naylor, J. , Raymond, C. , Rosetti, M. , Santos, R. , Sheridan, A. , Sougnez, C. , Stange-Thomann, N. , Stojanovic, N. , Subramanian, A. , Wyman, D. , Rogers, J. , Sulston, J. , Ainscough, R. , Beck, S. , Bentley, D. , Burton, J. , Clee, C. , Carter, N. , Coulson, A. , Deadman, R. , Deloukas, P. , Dunham, A. , Dunham, I. , Durbin, R. , French, L. , Grafham, D. , Gregory, S. , Hubbard, T. , Humphray, S. , Hunt, A. , Jones, M. , Lloyd, C. , McMurray, A. , Matthews, L. , Mercer, S. , Milne, S. , Mullikin, JC. , Mungall, A. , Plumb, R. , Ross, M. , Shownkeen, R. , Sims, S. , Waterston, RH. , Wilson, RK. , Hillier, LW. , McPherson, JD. , Marra, MA. , Mardis, ER. , Fulton, LA. , Chinwalla, AT. , Pepin, KH. , Gish, WR. , Chissoe, SL. , Wendl, MC. , Delehaunty, KD. , Miner, TL. , Delehaunty, A. , Kramer, JB. , Cook, LL. , Fulton, RS. , Johnson, DL. , Minx, PJ. , Clifton, SW. , Hawkins, T. , Branscomb, E. , Predki, P. , Richardson, P. , Wenning, S. , Slezak, T. , Doggett, N. , Cheng, JF. , Olsen, A. , Lucas, S. , Elkin, C. , Uberbacher, E. , Frazier, M. , Gibbs, RA. , Muzny, DM. , Scherer, SE. , Bouck, JB. , Sodergren, EJ. , Worley, KC. , Rives, CM. , Gorrell, JH. , Metzker, ML. , Naylor, SL. , Kucherlapati, RS. , Nelson, DL. , Weinstock, GM. , Sakaki, Y. , Fujiyama, A. , Hattori, M. , Yada, T. , Toyoda, A. , Itoh, T. , Kawagoe, C. , Watanabe, H. , Totoki, Y. , Taylor, T. , Weissenbach, J. , Heilig, R. , Saurin, W. , Artiguenave, F. , Brottier, P. , Bruls, T. , Pelletier, E. , Robert, C. , Wincker, P. , Smith, DR. , Doucette-Stamm, L. , Rubenfield, M. , Weinstock, K. , Lee, HM. , Dubois, J. , Rosenthal, A. , Platzer, M. , Nyakatura, G. , Taudien, S. , Rump, A. , Yang, H. , Yu, J. , Wang, J. , Huang, G. , Gu, J. , Hood, L. , Rowen, L. , Madan, A. , Qin, S. , Davis, RW. , Federspiel, NA. , Abola, AP. , Proctor, MJ. , Myers, RM. , Schmutz, J. , Dickson, M. , Grimwood, J. , Cox, DR. , Olson, MV. , Kaul, R. , Raymond, C. , Shimizu, N. , Kawasaki, K. , Minoshima, S. , Evans, GA. , Athanasiou, M. , Schultz, R. , Roe, BA. , Chen, F. , Pan, H. , Ramser, J. , Lehrach, H. , Reinhardt, R. , McCombie, WR. , de la Bastide, M. , Dedhia, N. , Blocker, H. , Hornischer, K. , Nordsiek, G. , Agarwala, R. , Aravind, L. , Bailey, JA. , Bateman, A. , Batzoglou, S. , Birney, E. ,

Bork, P. , Brown, DG. , Burge, CB. , Cerutti, L. , Chen, HC. , Church, D. , Clamp, M. , Copley, RR. , Doerks, T. , Eddy, SR. , Eichler, EE. , Furey, TS. , Galagan, J. , Gilbert, JG. , Harmon, C. , Hayashizaki, Y. , Haussler, D. , Hermjakob, H. , Hokamp, K. , Jang, W. , Johnson, LS. , Jones, TA. , Kasif, S. , Kaspryzk, A. , Kennedy, S. , Kent, WJ. , Kitts, P. , Koonin, EV. , Korf, I. , Kulp, D. , Lancet, D. , Lowe, TM. , McLysaght, A. , Mikkelsen, T. , Moran, JV. , Mulder, N. , Pollara, VJ. , Ponting, CP. , Schuler, G. , Schultz, J. , Slater, G. , Smit, AF. , Stupka, E. , Szustakowski, J. , Thierry-Mieg, D. , Thierry-Mieg, J. , Wagner, L. , Wallis, J. , Wheeler, R. , Williams, A. , Wolf, YI. , Wolfe, KH. , Yang, SP. , Yeh, RF. , Collins, F. , Guyer, MS. , Peterson, J. , Felsenfeld, A. , Wetterstrand, KA. , Patrinos, A. , Morgan, MJ. , Szustakowki, J. , de Jong, P. , Catanese, JJ. , Osoegawa, K. , Shizuya, H. , Choi, S. , and Chen, YJ . Initial sequencing and analysis of the human genome. *Nature*, 15(409):860–921, 2001.

Lewin, B. . Genes 7. *ISBN: 0-19-879276*, page 3, 2000.

Lisser, S. and Margalit, H. . Compilation of e.coli mrna promoter sequences. *Nucleic acids research*, 21(7):1507–1516, 1993.

Lowe, T.M. and Eddy, S.R. . A computational screen for methylation guide snoRNAs in yeast. *Science*, 283(5405):1168–71, 1999.

Mattick, J.S. . Challenging the dogma: the hidden layer of non-protein-coding rnas in comlex organisms. *BioEssays*, 25:930–939, 2003.

McClelland, M. , Sanderson, KE. , Spieth, J. , Clifton, SW. , Latreille, P. , Courtney, L. , Porwollik, S. , Ali, J. , Dante, M. , Du, F. , Hou, S. , Layman, D. , Leonard, S. , Nguyen, C. , Scott, K. , Holmes, A. , Grewal, N. , Mulvaney, E. , Ryan, E. , Sun, H. , Florea, L. , Miller, W. , Stoneking, T. , Nhan, M. , Waterston, R. , and Wilson, RK. . Complete genome sequence of salmonella enterica serovar typhimurium lt2. *Nature*, 413 (6858):852–6, 2001.

McCutcheon, J.P. and Eddy, S. . Computational identification of non-coding rnas in saccharomyces cervisiae by comparative genomics. *Nucleic Acids Research*, 31(14):4119–4128, 2003.

Olivas, W.M. , Muhlrad, D. , and Parker, R. . Analysis of the yeast genome: identification of new non-coding and small ORF-containing RNAs. *Nucleic Acids Res*, 25(22):4619–25, 1997.

Owens, J.T. , Chmura, A.J , Murakami, K. , Fujita, N. , Ishihama, A. , and Meares, M.F. . Mapping the promoter dna sites proximal to conserved regions of sigma 70 in an escherichia coli rna polymerase-lacuv5 open promoter complex. *Biochemistry*, 37:7670–7675, 1998.

Parkhill, J. , Dougan, G. , James, KD. , Thomson, NR. , Pickard, D. , Wain, J. , Churcher, C. , Mungall, KL. , Bentley, SD. , Holden, MT. , Sebaihia, M. , Baker, S. , Basham, D. , Brooks, K. , Chillingworth, T. , Connerton, P. , Cronin, A. , Davis, P. , Davies, RM. , Dowd, L. , White, N. , Farrar, J. , Feltwell, T. , Hamlin, N. , Haque, A. , Hien, TT. , Holroyd, S. , Jagels, K. , Krogh, A. , Larsen, TS. , Leather, S. , Moule, S. , O'Gaora, P. , Parry, C. , Quail, M. , Rutherford, K. , Simmonds, M. , Skelton, J. , Stevens, K. , Whitehead, S. , and Barrell, BG. . Complete genome sequence of a multiple drug resistant salmonella enterica serovar typhi ct18. *Nature*, 413(6858):848–52, 2001a.

Parkhill, J. , Wren, BW. , Thomson, NR. , Titball, RW. , Holden, MT. , Prentice, MB. , Sebaihia, M. , James, KD. , Churcher, C. , Mungall, KL. , Baker, S. , Basham, D. , Bentley, SD. , Brooks, K. , Cerdeno-Tarraga, AM. , Chillingworth, T. , Cronin, A. , Davies, RM. , Davis, P. , Dougan, G. , Feltwell, T. , Hamlin, N. , Holroyd, S. , Jagels, K. , Karlyshev, AV. , Leather, S. , Moule, S. , Oyston, PC. , Quail, M. , Rutherford, K. , Simmonds, M. , Skelton, J. , Stevens, K. , Whitehead, S. , and Barrell, BG. . Genome sequence of yersinia pestis, the causative agent of plague. *Nature*, 413(6855):523–7, 2001b.

Pedersen, AG. , Jensen, LJ. , Brunak, S. , Staerfeldt, HH. , and Usserey, D. . A dna structural atlas for escherichia coli. *J. Mol Biol*, 299(4):907–930, 2000.

Richardson, J.P . Rho-dependent termination and atpases in transcript termination. *Biochimica et Biophysica*, 1577:251–260, 2002.

Riddihough, G. . The other rna world. *Science*, 296(5571):1259, 2002.

Rivas, E. and Eddy, S.R. . Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 16(7):583–605, 2000.

Rivas, E. , Klein, R.J. , Jones, T.A. , and Eddy, S.R. . Computational identification of noncoding RNAs in E. coli by comparative genomics. *Curr Biol*, 11(17):1369–73, 2001.

Salgado, H. , Santos-Zavaleta, A. , Gamo-Castro, S. , Millan-Zarate, D. , Blattner, F.R. , and Collado-Vides, J. . Regulondb(version3.0):transcriptional regulation and operon organization in escerichia coli k-12. *Nucleic Acids Research*, 28:65–67, 2000.

Schneider, T. D. and Stephens, R. M. . Sequence logos: A new way to display consensus s equences. *Nucleic Acids Res.*, 18:6097–6100, 1990.

Shigenobu, S. , Watanabe, H. , Hattori, M. , Sakaki, Y. , and Ishikawa, H. . Genome sequence of the endocellular bacterial symbiont of aphids buchnera sp. aps. *Nature*, 7(407):81–6, 2000.

Skovgaard, M. , Jensen, L.J. , Brunak, S. , Usserey, D. , and Krogh, A. . On the total number of genes and their length distribution in complete microbial genes. *Trends in Genetics*, 17(8):425–428, 2001.

Smith, T.F. and Waterman, M.S. . Identification of common molecular subsequences. *J Mol Biol.*, 147(1):195–7, 1981.

Storz, G. . An expanding universe of noncoding RNAs. *Science*, 296 (5571):1260–3, 2002.

Szymanski, M. and Barciszewski, J. . Beyond the proteome: non-coding regulatory rnas. *Genome Biology*, Genome Biology 2002(3(5)):reviews 0005.1–0005.8, 2002.

Tamas, I. , Klasson, L. , Canback, B. , Naslund, A.K. , Eriksson, A.S. , Wernegreen, J.J. , Sandstrom, J.P. , Moran, N.A. , and Andersson, S.G. . 50 million years of genomic stasis in endosymbiotic bacteria. *Science*, 28(296):2376–9, 2002.

Tjaden, B. , Saxena, R.M. , Stolyar, S. , Haynor, D.R. , Kolker, E. , and Rosenow, C. . Transcriptome analysis of escherichia coli using high-density oligonucleotide probe array. *Nucleic Acids Research*, 30(17), 2002.

Ussery, David . Bioinformatics lecture notes 12-nov-99. *http://www.cbs.dtu.dk/staff/dave/MScourse/promoters.html*, pages 1–6, 1999.

van Ham, RC. , Kamerbeek, J.and Palacios, C. , Rausell, C. , Abascal, F. , Bastolla, U. , Fernandez, J.M. , Jimenez, L. , Postigo, M. , Silva, F.J. Tamames, J. , Viguera, E. , Latorre, A. , Valencia, A. , Moran, F. , and Moya, A. . Reductive genome evolution in buchnera aphidicola. *Proc Natl Acad Sci U S A.*, 100(2):581–586, 2003.

Vogel, J. , Bartels, V. , Tang, T.H. , Churakov, G. , Slagter-Jager, J.G , Huttenhofer, A. , and E.G.H., Wagner . Rnomics in escherichia coli detects new srna species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Research*, 31(22):6435–6443, 2003.

Wassarman, K.M. , Repoila, F. , Rosenow, C. , Storz, G. , and Gottesman, S. . Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev*, 15(13):1637–51, 2001.

Wassarman, K.M. , Zhang, A. , and Storz, G. . Small RNAs in Escherichia coli. *Trends Microbiol*, 7(1):37–45, 1999.

Wei, J. , Goldberg, MB. , Burland, V. , Venkatesan, MM. , Deng, W. , Fournier, G. , Mayhew, GF. , Plunkett, G, 3rd. , Rose, DJ. , Darling, A. , Mau, B. , Perna, NT. , Payne, SM. , Runyen-Janecky, LJ. , Zhou, S. , Schwartz, DC. , and Blattner, FR. . Complete genome sequence and comparative genomics of shigella flexneri serotype 2a strain 2457t. *Infect Immun*, 71(5):2775–86, 2003.

Yada, T. , Nakao, M. , Totoki, Y. , and Nakai, K. . Modelling and predicting transcriptional units of escherichia coli genes usin hidden markov models. *Bioinformatics*, 15(12):987–993, 1999.

Zuker, M. . Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):3406–15, 2003.