Computational approaches for in-depth analysis of cDNA sequence tags

PER UNNEBERG

Doctoral Thesis
Stockholm, Sweden 2004
Abstract

Major recent improvements in biotechnology have led to an accelerated production of DNA sequences. The completion of the human genome sequence, along with the genomes of more than two hundred other species, has marked the arrival of the genome era. The ultimate goal is to understand the structure and function of genomes and their genes. This thesis has focused on the computational analysis of complementary DNA (cDNA) sequences. These are copies of mRNA transcripts that correspond to the coding regions of genomes.

Studying the expression patterns of genes is essential for understanding gene function. Many gene expression profiling techniques generate short sequence tags that derive from transcripts. A pilot study was performed to assess the feasibility of using the pyrosequencing platform for gene expression analysis. The sequences generated by pyrosequencing in most cases (≈ 85%) were long enough (> 18 nucleotides) to uniquely identify the corresponding transcripts through database searches. Aspects of transcript identification by short sequence tags were further investigated in a number of public databases, revealing that a tag length 16–17 nucleotides was sufficient for unique identification.

Longer transcript representations are obtained from expressed sequence tag (EST) sequencing. Method development for the analysis and maintenance of large EST data sets has been performed on data from poplar, which is a tree of commercial interest to the forest biotechnology industry. In 2003 a large EST-sequencing project reached > 100 000 reads, providing a unique resource for tree biology research. ESTs have been grouped into clusters and singletons that represent potential genes. Preliminary analyses have estimated gene content in *Populus* to be very similar to that of model organism *Arabidopsis thaliana*.

EST data collections provide a rich source for mining polymorphisms. A software application was developed and applied to EST data from two *Populus* species, and candidate single nucleotide polymorphisms (SNPs) were recorded. A study of genetic variation between the species revealed a striking similarity, with orthologous pairs being > 98% identical on the protein level.

**Keywords:** cDNA, EST, gene expression, SNP, SAGE, polymorphism, assembly, clustering, DNA sequencing, pyrosequencing, mRNA transcript, orthology, tree biotechnology, restriction enzyme
LIST OF PUBLICATIONS

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


# Contents

<table>
<thead>
<tr>
<th>LIST OF PUBLICATIONS</th>
<th>iv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contents</td>
<td>v</td>
</tr>
</tbody>
</table>

## 1 INTRODUCTION

1 Introduction

1.1 From lab bench to large-scale industry .......................... 3
1.2 The genome era ........................................... 4
1.3 The role of cDNA sequencing ............................. 7

## 2 cDNA sequence tags

2.1 DNA sequencing technologies ................................. 9
   2.1.1 Sanger sequencing .................................... 9
   2.1.2 Pyrosequencing ..................................... 11
2.2 From mRNA to cDNA .......................................... 11
   2.2.1 mRNA ................................................. 11
   2.2.2 cDNA libraries ....................................... 13
2.3 Tag sequencing protocols .................................... 14
   2.3.1 Expressed sequence tags .............................. 14
   2.3.2 SAGE and related methods ......................... 16
   2.3.3 Pyrosequencing and MPSS ............................ 16

## 3 Data organisation and data flow

3.1 From sequencing machine to human-readable format .......... 19
3.2 Clustering and assembly of EST data ........................... 19
   3.2.1 Filtering steps ...................................... 21
   3.2.2 Clustering techniques ................................ 21
   3.2.3 Transcript assembly .................................. 21
3.3 Databases .................................................. 22
   3.3.1 Sequence tag databases ............................... 22
   3.3.2 Gene index databases ................................. 24
## CONTENTS

3.3.3 Genetic variation databases ........................................... 25

4 Analysis and applications of cDNA sequences .......................... 27
   4.1 Identification and characterisation of cDNA sequences through
database searches ..................................................... 27
      4.1.1 Identification via tag mapping ............................ 28
      4.1.2 Identification via sequence alignment .................... 28
      4.1.3 Characterisation via homology ............................ 29
   4.2 Expression profiling and gene discovery .......................... 29
      4.2.1 Expression profiling ....................................... 30
      4.2.2 Gene discovery ........................................... 31
   4.3 Genome annotation .............................................. 31
   4.4 Analysis of genetic variation ..................................... 32

II PRESENT INVESTIGATION ................................................. 35

5 Analysing gene expression with short sequence tags (I, II) ............ 39
   5.1 Using pyrosequencing for gene expression analysis (I) ............ 39
   5.2 The importance of reliable databases (II) .......................... 39
      5.2.1 Restriction enzyme analysis .............................. 40
      5.2.2 Transcript identification using sequence alignments .......... 41
      5.2.3 Tag mapping ........................................... 42

6 An EST resource for Populus (III,IV,V) ................................ 43
   6.1 SNP discovery in EST data (IV) .................................. 44
   6.2 Comparison of two poplar EST data sets (V) ....................... 45

7 Concluding remarks ...................................................... 49

Abbreviations .................................................................. 51

Acknowledgements ......................................................... 53

REFERENCES .................................................................. 55

Publications (I-V) ............................................................. 73
Part I

INTRODUCTION
Chapter 1

Introduction

1.1 From lab bench to large-scale industry

In a matter of two generations, biotechnology has evolved from lab bench research conducted by committed scientists to a large-scale industry. Progress, especially during the last two decades, has accelerated at a tremendous rate, culminating in the recent completion of the Human Genome Project (HGP). During this period, we have witnessed the birth and evolution of the genomics research field, which promises to revolutionise many aspects of our lives (Collins et al., 2003a).

A number of key discoveries made during the 1940s–70s identified the basic components of cellular life and laid the foundations for modern biotechnology. Deoxyribonucleic acid (DNA) became the focus of attention in modern genetics in 1944 when Oswald Avery showed that DNA is the basis of heredity (Avery et al., 1944). Almost ten years later, in a classic paper Watson and Crick determined the structure of DNA (Watson and Crick, 1953). Crick later also postulated the central dogma of molecular biology which describes the information flow from gene through messenger RNA (mRNA) to protein in the cell (Figure 1). The link between DNA and protein became fully established in 1966 when the genetic code was determined (Nirenberg, 1963). Recombinant biotechnology emerged with the discoveries of reverse transcriptase and restriction enzymes.

Various technical improvements have further enhanced the scope of biotechnology. A major step was taken 1977 when two DNA sequencing technologies emerged (Sanger et al., 1977b; Maxam and Gilbert, 1977). Shortly afterwards, the complete sequences of bacteriophage ΦX174 (Sanger et al., 1977a) and human mitochondrial DNA (Anderson et al., 1981) were published, foreshadowing the large-scale sequencing projects that were shortly to be initiated. The development of the polymerase chain reaction (PCR) (Saiki et al., 1988), together with the introduction of automated sequencing methods (Smith et al., 1986), set the stage for the era of genomics.
1. INTRODUCTION

Figure 1. Schematic overview of the central dogma. Genetic material DNA is transcribed into mRNA, which in turn is translated into protein. DNA can also make copies of itself via a process termed replication. The process of reverse transcription, in which mRNA is used as a template to generate complementary DNA (cDNA), is also illustrated.

1.2 The genome era

The concept of large-scale sequencing became emphasised in 1990 with the presentation of the first five-year plan of the HGP, a public genome project to sequence the human genome and that of five model organisms Escherichia coli, Saccharomyces cerevisiae, Drosophila melanogaster, Candida elegans, and mouse. Although targeted for completion in 2005, we now know that the goal of the HGP was achieved ahead of schedule as the international consortium marked the end of the project in 2003. Two years earlier, the preliminary results of both the public (Lander et al., 2001) and private (Venter et al., 2001) efforts were published.

Not only was the HGP completed ahead of schedule, but the project delivered more than expected, and cost less than projected (Collins et al., 2003b). However, the first free-living organism to be sequenced was not part of the HGP. In 1995, the genome sequence of Haemophilus influenzae was published by Fleischmann et al., catalysing the development of new methods which led to increased generation of sequence data (Venter et al., 2003). According to the Genome Online Database (GOLD) (Bernal et al., 2001), 210 genomes have been published, and 1 166 genomes are currently being sequenced. Table 1 on the facing page summarises the eukaryotic genome sequences that have been completed to date.

The increased rate of sequencing throughput has led to a data explosion, generating a treasure trove of biological information. Nucleotide database growth
### Table 1. Table of published eukaryotic genomes.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome size (Mb)</th>
<th>dbEST entries (Dec. 2003)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. cerevisiae</em> (baker’s yeast)</td>
<td>12</td>
<td>3 000</td>
<td>Goffeau et al. (1996)</td>
</tr>
<tr>
<td>C. elegans</td>
<td>12</td>
<td>299 000</td>
<td>The <em>C. elegans</em> Sequencing Consortium (1998)</td>
</tr>
<tr>
<td><em>D. melanogaster</em> (fruit fly)</td>
<td>137</td>
<td>382 000</td>
<td>Adams et al. (2000)</td>
</tr>
<tr>
<td><em>A. thaliana</em> (thale cress)</td>
<td>115</td>
<td>322 000</td>
<td>The <em>Arabidopsis</em> Genome Initiative (2000)</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>3 000</td>
<td>5 658 000</td>
<td>Lander et al. (2001)</td>
</tr>
<tr>
<td><em>Guillardia theta</em></td>
<td>0.55</td>
<td>450</td>
<td>Douglas et al. (2001)</td>
</tr>
<tr>
<td><em>Encephalitozoon cuniculi</em></td>
<td>2.5</td>
<td>0</td>
<td>Katinka et al. (2001)</td>
</tr>
<tr>
<td><em>Schizosaccharomyces pombe</em></td>
<td>14</td>
<td>8 100</td>
<td>Wood et al. (2002)</td>
</tr>
<tr>
<td><em>Oryza sativa japonica</em> (rice)</td>
<td>420</td>
<td>284 000</td>
<td>Goff et al. (2002)</td>
</tr>
<tr>
<td><em>Oryza sativa indica</em> (rice)</td>
<td>420</td>
<td>0</td>
<td>Yu et al. (2002)</td>
</tr>
<tr>
<td><em>Plasmodium falciparum</em></td>
<td>22.9</td>
<td>20 200</td>
<td>Gardner et al. (2002)</td>
</tr>
<tr>
<td><em>Plasmodium yoelii</em></td>
<td>23.1</td>
<td>13 900</td>
<td>Carlton et al. (2002)</td>
</tr>
<tr>
<td><em>Anopheles gambiae</em> (malaria mosquito)</td>
<td>0.28</td>
<td>150 000</td>
<td>Holt et al. (2002)</td>
</tr>
<tr>
<td><em>Takifugu rubripes</em></td>
<td>365</td>
<td>25 900</td>
<td>Aparicio et al. (2002)</td>
</tr>
<tr>
<td><em>Mus musculus</em> (mouse)</td>
<td>2 800</td>
<td>4 243 000</td>
<td>Waterston et al. (2002)</td>
</tr>
<tr>
<td><em>Ciona intestinalis</em> (sea urchin)</td>
<td>116.7</td>
<td>684 000</td>
<td>Dehal et al. (2002)</td>
</tr>
<tr>
<td><em>Neurospora crassa</em></td>
<td>43</td>
<td>28 000</td>
<td>Galagan et al. (2003)</td>
</tr>
<tr>
<td><em>Caenorhabditis briggsae</em></td>
<td>104</td>
<td>2 400</td>
<td>Stein et al. (2003)</td>
</tr>
<tr>
<td><em>Bombyx mori</em> (silkworm)</td>
<td>530</td>
<td>117 000</td>
<td>Mita et al. (2004)</td>
</tr>
<tr>
<td><em>Cyanidioschyzon merolae</em></td>
<td>16.5</td>
<td>0</td>
<td>Matsuzaki et al. (2004)</td>
</tr>
<tr>
<td><em>Cryptosporidium parvum</em></td>
<td>10.4</td>
<td>570</td>
<td>Abrahamsen et al. (2004)</td>
</tr>
<tr>
<td><em>Ashbya gossypii</em></td>
<td>9</td>
<td>0</td>
<td>Dietrich et al. (2004)</td>
</tr>
<tr>
<td><em>Candida glabrata</em></td>
<td>12.3</td>
<td>0</td>
<td>Dujon et al. (2004)</td>
</tr>
<tr>
<td><em>Kluyveromyces lactis</em></td>
<td>10.6</td>
<td>0</td>
<td>Dujon et al. (2004)</td>
</tr>
<tr>
<td><em>Yarrowia lipolytica</em></td>
<td>20.5</td>
<td>0</td>
<td>Dujon et al. (2004)</td>
</tr>
<tr>
<td><em>Debaryomyces hansenii</em></td>
<td>12.2</td>
<td>0</td>
<td>Dujon et al. (2004)</td>
</tr>
<tr>
<td><em>Phanerochaete chrysosporium</em></td>
<td>30</td>
<td>110</td>
<td>Martinez et al. (2004)</td>
</tr>
</tbody>
</table>
is exponential, as illustrated by looking at the completion rate of entire genomes (Figure 2). This highlights the importance of advances in information technology and their contribution to genomics (Kanehisa and Bork, 2003), without which projects like the HGP would hardly have been completed on time.

Focus is now shifting from generating DNA sequence data to analysis of the completed genome sequences at hand. As such, raw genome sequences are of little value (Clark, 1999). Functional annotation, in which genes and regulatory elements are identified and characterised, is a primary goal in making sense out of sequence (Jareborg and Durbin, 2000). However, due to the complex arrangement of exons and introns in eukaryotes, \textit{ab initio} prediction of genes is extremely difficult (Guigo et al., 2000; Claverie et al., 1997). Gene prediction methods have recently improved through the introduction of comparative analysis of several genomes (Flicek et al., 2003; Parra et al., 2003), although mammalian gene prediction is not completely accurate (Wang et al., 2003; Brent and Guigo, 2004).
1.3 The role of cDNA sequencing

At the onset of the HGP, gene discovery was still a labour-intensive and time-consuming task. However, the introduction of high-throughput EST sequencing (Adams et al., 1991) provided a rapid method for gene discovery. One of the main advantages of cDNA sequencing is that direct access is gained to the coding regions of a genome. In humans only 1.5% of the genome is protein-coding (Lander et al., 2001), prompting several researchers to advocate the inclusion of a large cDNA sequencing effort in the HGP. Furthermore, compared to genome sequencing, it is easier, quicker and cheaper to sequence cDNA collections. Finally, EST sequences were found to be of great help in the annotation of genomes (Venter et al., 2003). With information gained from EST sequencing projects (Boguski et al., 1993; Schuler et al., 1996; Boguski and Schuler, 1995), comprehensive gene catalogues and databases have been compiled.

Since a cDNA represents a gene sequence, high-throughput EST sequencing in effect serves as a means to study global gene expression analysis. The publication by Adams et al. eventually led to the development of three important gene expression profiling methods, namely serial analysis of gene expression (SAGE) (Velculescu et al., 1995), high-density oligonucleotide microarrays (Lockhart et al., 1996; Wodicka et al., 1997) and cDNA microarrays (Schena et al., 1995). The SAGE method was an early example of a methodology that uses shorter sequence tags to identify cDNAs, increasing the number of samples that can be analysed in one sequencing reaction.

The plentitude and high information content of cDNA data makes in silico analysis important. For instance, data mining of EST data sets has been applied to the study of genetic variation such as SNPs (Marth et al., 1999; Picoult-Newberg et al., 1999). Development of new cDNA analysis methods and analyses of existing data sets continues to be an important component of biotechnology research.
Chapter 2

cDNA sequence tags

The major theme of this thesis is the study of short sequence tags generated from cDNA clones. This chapter provides a brief overview of DNA sequencing, cDNA production and cloning, and tag sequencing protocols. Although the focus is on in silico analysis of cDNA tags, my intention is to shed light on the nature of the data, thus exemplifying some methodology issues that have implications for tag sequence analysis and interpretation.

2.1 DNA sequencing technologies

In the late 1970s Sanger et al. described a method for determining DNA nucleotide sequences (Sanger et al., 1977b). Although almost 30 years have passed since then, the majority of current DNA sequencing methods are still based on the principles of Sanger sequencing. There have been several suggestions for alternative approaches, one of which is pyrosequencing (Ronaghi et al., 1996). This thesis is concerned with data generated by either Sanger sequencing or pyrosequencing. Hence, other methods will not be described in detail.

2.1.1 Sanger sequencing

Sanger sequencing is an enzymatic method that relies upon DNA polymerase-mediated extension of template DNA. The reaction mix consists of DNA template, an oligonucleotide primer, deoxynucleotides (dNTPs) adenine (A), cytosine (C), guanine (G), and thymine (T), and dideoxynucleotides (ddNTPs), which differ from the dNTPs in that they lack a 3′ hydroxyl group. Incorporation of ddNTP induces base-specific termination of the growing DNA chain. DNA fragments of different sizes are produced which differ at the 3′ end. The fragments are subsequently electrophoretically separated on a matrix, producing a banded pattern that reveals the DNA template sequence (Figure 3).
Numerous technological advances have dramatically increased the throughput and quality of data generated by Sanger sequencing. Most importantly, large-scale sequencing projects would not have been possible without the invention of automated DNA sequencers, in which the analysis of electrophoretic patterns is performed simultaneously with sequencing (Sterky and Lundeberg, 2000). Whereas the original method by Sanger used radioactive labelling of fragments and separated the fragments in four lanes, the introduction of fluorescent dyes has enabled individual labelling of nucleotides and separation in a single lane (Smith et al., 1986). The recent introduction of capillary-based electrophoresis has further reduced sequencing time (Meldrum, 2000). At present, many sequencers produce sequences with a read length of 600 base pairs (bps) or more.
2.1.2 Pyrosequencing

An alternative approach to DNA sequencing was invented by Pål Nyrén and co-workers during the 1980s, laying the foundations for the current method of pyrosequencing. This is based on the principle of converting the pyrophosphate (PP$_i$) that is released upon DNA extension into light (Hyman, 1988; Ronaghi et al., 1996). By adding nucleotides in consecutive order, it is possible to monitor which nucleotides are incorporated by registering the release of PP$_i$.

Pyrosequencing, in its original format, is a solid-phase technique in which three enzymes are used as follows. Given a DNA template, nucleotides and a primer, nucleotides are incorporated and PP$_i$ is released by DNA polymerase. ATP-sulfurylase then converts PP$_i$ and adenosine di-phosphate (ADP) to adenosine tri-phosphate (ATP). In the final light-generating step, firefly luciferase utilises the energy in ATP to oxidise luciferin and generates light (Figure 4). For each round of nucleotide addition, a washing step was originally required to remove free nucleotides. Currently, however pyrosequencing is a four-enzyme technique and the sequencing reaction takes place in liquid phase in a test tube (Ronaghi et al., 1998). The enzyme apyrase effectively degrades unincorporated nucleotides and eliminates the washing step.

Pyrosequencing has several benefits as compared to conventional Sanger sequencing. It is easily automated, accurate and does not require labelling of primers or nucleotides. Furthermore, as the reaction takes place in real-time and data analysis is immediate, the time-consuming electrophoresis is eliminated. Drawbacks include the costs of the biotinylated primers required for template preparation, and the short read lengths of 20–30 bp.

Currently, pyrosequencing is finding extensive use in single nucleotide polymorphism (SNP) genotyping (Fakhrai-Rad et al., 2002). Moreover, the short read length makes pyrosequencing suitable for tag-generating studies. Recently, attention has focused on miniaturising pyrosequencing to increase throughput (Ahmadian et al., 2002; Russom et al., 2003a,b). In addition, a massively parallel system based on pyrosequencing that can routinely sequence 10 Mbp genomes has been created by the company 454 Life Sciences (www.454.com).

2.2 From mRNA to cDNA

The quality of the final sequence data is limited by the quality of the methods used to produce the raw material for sequencing. It is important to keep in mind the biological background of the data, which may have consequences for the subsequent analysis.

2.2.1 mRNA

The differences in gene organisation between prokaryotes and eukaryotes are reflected in the differing fates of their newly synthesised mRNAs. Prokaryote genes
2. CDNA SEQUENCE TAGS

Figure 4. Basic principle of pyrosequencing. When a nucleotide is incorporated, PP\textsubscript{i} is released and subsequently converted to light. Since nucleotides are added in consecutive order, it is possible to deduce the identity of the incorporated nucleotide.

lack introns and are often organised into operons, with several genes being transcribed at once. Consequently, the transcription of prokaryotic genes is straightforward, with transcription and translation often taking place simultaneously (Proudfoot, 2000).

Eukaryotic genes contain introns and various post-transcriptional modifications take place after transcription has generated a precursor mRNA molecule. The 5’ end is capped with 7-methylguanine (Shatkin, 1976), intronic regions are removed by splicing, and the 3’ ends are polyadenylated (Colgan and Manley, 1997). In mature eukaryotic mRNA, the coding sequence (CDS) is flanked by 5’ and 3’ untranslated regions (UTRs) (Figure 5). After modification, the mature mRNA is transported from the nucleus into the cytosol where the mRNA is translated into protein.

Figure 5. Illustration of a mature mRNA. The 5’ end is capped with 7-methylguanine. The CDS is flanked by 5’ UTR and 3’ UTR. Shown is also the polyadenylation signal AAUAAA which triggers the addition of the poly(A)-tail.
In higher organisms, alternative splicing commonly occurs which increases the combinatorial output of a gene (Roberts and Smith, 2002; Modrek and Lee, 2002). A recently described way to increase mRNA variability is through alternate polyadenylation (Beaudoing and Gautheret, 2001; Pauws et al., 2001; Iseli et al., 2002). Both these modifications are tissue-dependent.

Due to the simultaneous expression of several genes, the nucleus contains many different mRNA molecules at different stages of maturity. Termed heterogeneous nuclear RNA (hnRNA), this constitutes less than 3–5% of a cell’s RNA content (Alberts et al., 1994). Furthermore, most mammalian cells express > 10000 genes, represented by approximately $0.4 \times 10^8$ mRNAs (Jackson et al., 2000). The distribution of transcript levels follows a Pareto-like distribution, with a large fraction of transcript species ($\approx 70\%$) being expressed at low levels, and few transcripts expressed at high levels (Kuznetsov et al., 2002). Random sampling of mRNAs therefore quickly leads to redundancy in cDNA libraries, and the probability of not observing a given transcript species can be high unless sample sizes are large (Vingron and Hoheisel, 1999).

### 2.2.2 cDNA libraries

The construction of cDNA libraries from an mRNA population can basically be divided into three phases: mRNA isolation, cDNA synthesis and cDNA cloning. Each step is briefly described here, along with the problems that might arise.

#### mRNA isolation

Several protocols exist for isolating total RNA from cells/tissues. mRNA is purified from total RNA by solid-phase affinity selection. In the case of eukaryotic mRNA, selection is performed on columns or beads with immobilised oligo(dT) probes. To avoid bias against short clones, the purified mRNAs may be size fractionated before library construction.

A small fraction of the mRNA population lacks a poly(A)-tail (Wang et al., 2000), and thus, it is generally not possible to purify all eukaryotic mRNA species with oligo(dT) probes. In addition, problems may arise in downstream data analysis since the purified mRNA sample will contain immature mRNAs, and RNase activity constantly exposes mRNAs to degradation (Das et al., 2001).

#### cDNA synthesis

The general steps of cDNA synthesis are: (1) first-strand synthesis, in which reverse transcriptase synthesises a cDNA-strand complementary to the mRNA; (2) RNA removal with alkali; (3) second-strand synthesis, in which DNA polymerase syntheses the second complementary DNA-strand. The choice of primer for first-strand synthesis governs the properties of the resulting cDNA sequences. Priming with oligo(dT) is required to obtain the entire 3' UTR. However, the oligo(dT)
primer may misanneal and prime at internal A-rich sites in mRNA: studies by Bonaldo et al. suggest that 10–15% cDNAs have 3' truncations. Stable hairpin loops can further hinder the extension of the first strand, which in practice means that full-length cDNAs will be scarce. Alternatively, random primers can be used to increase the coding content as UTR regions are excluded to a higher extent in the resulting fragments. The generated cDNA fragments will correspond to random subsequences of the entire transcript.

2.3 Tag sequencing protocols

cDNA libraries often contain clones with inserts in the range 1–4 kilobases (kb). Given that the read length of DNA sequencers is limited (< 800 bp), there are no efficient methods for systematically obtaining the entire sequence of many cDNA clones (Andersson et al., 1997) (Figure 6).

The last decade has seen the emergence of several methods that generate short sequence tags which correspond to mRNAs. The most important methods are briefly reviewed below.

2.3.1 Expressed sequence tags

When the HGP began to take shape at the end of the 1980s, human gene number estimates were in the range of 50 000–100 000 genes, but only a few thousand were represented in GenBank. As the completion of the human genome sequence was estimated to lie 12–15 years ahead, it was suggested that efforts should also be made to sequence cDNAs in order to increase the coverage of human genes studied.

In 1991, Adams and coworkers accomplished a pilot study where randomly selected cDNA clones were single-pass sequenced from each clone end (Adams et al., 1991). To increase the information content, the cDNA library consisted of randomly primed clones, thus enriching the number of clones with coding content. The partial cDNA sequences were named expressed sequence tags (ESTs)
and proved to be capable of identifying previously uncharacterised genes. Although the principle of ESTs had previously been established (Putney et al., 1983; Milner and Sutcliffe, 1983), this was the first large-scale generation of ESTs. The mRNA population had been randomly sampled, so the cDNA library reflected the transcriptome’s composition. However, the random priming approach is not suitable for quantifying gene expression levels. By using a 3'-directed cDNA library, where cDNAs had been digested with MpoI to ensure unique identification of each mRNA, expression profiling by EST sequencing was introduced in the following year (Okubo et al., 1992). These studies initiated a high number of large-scale EST projects – see chapter 4 for examples.

There are two main problems with ESTs which have to be taken into account during analysis. First, the problems inherent in cDNA library construction are reflected in EST data. Deriving a faithful representation of transcripts is complicated by the presence of contaminants. ESTs are fragmentary in nature since sequencing only generates incomplete reads from clone ends. In addition, sequence bias is introduced through the truncation of 5'-ends along with poly(A)-selection of transcripts. Sequence quality is further reduced by errors introduced by reverse transcriptase. It is estimated that 3% of nucleotides within any EST are incorrect (Hillier et al., 1996). Second, since cDNA libraries reflect the distribution of transcripts in the underlying mRNA population, rare transcripts are poorly represented. Very large samples, often exceeding practical limitations, are needed to observe these transcripts.
2. CDNA SEQUENCE TAGS

2.3.2 SAGE and related methods

Designed to further increase the throughput of cDNA sequencing, serial analysis of gene expression (SAGE) (Velculescu et al., 1995) has become a commonly used alternative for sequencing-based transcriptome profiling (Velculescu et al., 1999; Zhang et al., 1997; Welle et al., 1999; Jones et al., 2001; Fujii and Amrein, 2002; Steen et al., 2002; Blackshaw et al., 2004). SAGE relies on the fundamental assumption that a short sequence tag (9–10 bp) is sufficient to uniquely identify an mRNA transcript. Tags are generated by cleaving the mRNA population with a restriction endonuclease. The tags are then randomly ligated tail-to-tail into “ditags” and amplified by PCR. The amplified ditags are then concatenated, cloned and sequenced (Figure 7). Consequently, substantial numbers of transcripts (> 50) can potentially be identified for each sequencing reaction, thereby ideally increasing throughput by the same order of magnitude compared to EST sequencing.

SAGE is a very potent method, but unfortunately, it is difficult to prepare tag libraries (Yamamoto et al., 2001). In addition, tag distributions are potentially biased, for example by preferential ditag amplification (Spinella et al., 1999), mRNAs lacking a restriction enzyme recognition site will be excluded from analysis, and tag lengths may not be sufficient to uniquely identify the original transcript. Furthermore, the type IIS restriction endonuclease used for tag generation does not always give equally-sized restriction products. Technical difficulties include the need for large mRNA samples, and the fact that a large number of linker-dimers are generated by PCR significantly reduces the number of observables.

Several modifications have since been proposed to circumvent the weaknesses of the original protocol. For instance, microSAGE (Datson et al., 1999) is designed to remedy the sample size problem, and biotinylated PCR primers have been introduced to minimise linker-dimer contamination (Powell, 1998). The majority of improvements, however, have been directed at increasing the length of the tags, and thus boost identification strength (Spinella et al., 1999; Yamamoto et al., 2001; Saha et al., 2002; Chen et al., 2000).

2.3.3 Pyrosequencing and MPSS

Two important alternative tag sequencing protocols are pyrosequencing and massively parallel signature sequencing (MPSS). The principles of pyrosequencing have been outlined in section 2.1.2, and its potential for generating tag sequences was mentioned. The evaluation of a pilot project for gene expression studies is documented in article I appended to this thesis.

MPSS relies on microbead technology to accomplish the in vitro cloning of millions of templates (Brenner et al., 2000). The signature of each clone is then determined by a sequencing technique that is non-gel-based and avoids separation of DNA fragments. However, the cost and complexity of the technique does not make it amenable for high-throughput analysis or diagnostics (Jongeneel et al.,
Tag sequencing protocols

Figure 7. Schematic diagram of the SAGE method. The anchoring enzyme is NlaIII and the type IIS enzyme is BsmFI. Recognition sites for enzymes are illustrated by grey boxes. The linkers A and B are represented with surrounding boxes. The 3’ regions of the linkers contain tagging enzyme recognition sequence. X and Y represent bases in the transcript-derived tag sequences.
2. **CDNA SEQUENCE TAGS**

Alternatives that produce longer signature tags have also been applied to MPSS (Silva et al., 2004).
Chapter 3

Data organisation and data flow

3.1 From sequencing machine to human-readable format

The raw data produced by automated sequencing machines must be processed into signal intensity data which can be transformed via a process called base-calling into a sequence of DNA bases. In conventional Sanger sequencing with four dyes, signal intensity data are usually summarised as “traces” in chromatograms, consisting of four differently coloured curves that correspond to DNA bases. Base-calling can be complicated by several factors, such as poor peak resolution and compression (Ewing et al., 1998).

Early in the history of automated sequencing it became obvious that error probabilities should be assigned to called bases (Bonfield and Staden, 1995). With this in mind, the standard chromatogram format (SCF) file format was developed (Dear and Staden, 1992). Today, PHRED (Ewing and Green, 1998) is by far the most commonly used base-calling program. PHRED reads and processes several chromatogram formats, such as SCF files and chromatograms generated by Applied Biosystems sequencers (www.appliedbiosystems.com). Base calls and quality values are stored in Phred files.

Currently, there is no base-calling software for pyrosequencing. Instead, signal data are summarised in pyrograms and base-calling is done manually.

3.2 Clustering and assembly of EST data

Clustering and assembly of EST sequences is an important data management topic. The steps involved in the procedure are described in the following sections. An overview of the steps is illustrated in Figure 8.
3. Data organisation and data flow

Figure 8. Illustration of the steps involved in cDNA library construction, EST sequencing, and sequence assembly. Exons are represented by boxes, DNA is coloured red, mRNA blue, EST reads green and magenta. 1) Genes are transcribed into mRNA. 2) Splicing occurs, producing mature or in some cases immature mRNAs. 3) mRNA is reverse transcribed to cDNA. In some cases also immature mRNA is reverse transcribed, indicated by the dashed line. 4) cDNAs are cloned. 5) EST sequencing generates 5’ and 3’ reads. 6) EST data are clustered into clusters and singletons.
3.2.1 Filtering steps

Given a set of base-called sequence reads, the first priority is to filter the data. EST data sets inevitably contain unwanted sequence such as ribosomal RNA (rRNA) or host organism (usually *E. coli*) DNA. Furthermore, desired reads are flanked by vector and adaptor sequences. Consequently, the contaminating sequences must be identified and processed (Telles and da Silva, 2001).

Usually, similarity search algorithms such as BLAST (Altschul et al., 1990) are used to identify contaminants. rRNA and host organism reads are discarded from further analyses whereas vector and adaptor sequences are masked. If quality values are available, low-quality regions are often masked.

Low-complexity regions, repeats and poly(A) regions should not be removed *per se* if they are part of *bona fide* cDNAs. However, these regions introduce artificial links between unrelated sequences which may lead to overclustering (see below). Either these regions are annotated and thus overlooked during clustering or they are masked.

3.2.2 Clustering techniques

The general aim of clustering is to group data that are related by using a similarity measure (Jain et al., 1999). Broadly speaking, the criteria for letting two sequences belong to the same cluster range from loose to strict (Burke et al., 1998).

Loose clustering techniques group two sequences if they share overlaps above a given threshold. Alternatively spliced transcripts may end up in the same cluster (Haas et al., 2000). The presence of chimeras and artifacts leads to a tendency of overclustering. Similarity is usually measured by standard algorithms such as BLAST, FASTA (Pearson, 1990) or variants of the Smith-Waterman algorithm (Smith and Waterman, 1981a). In d2_cluster (Burke et al., 1999), a two-phase strategy is employed to speed up clustering. Common subsequences of 13 bases separated by no more than two bases are identified, after which similarity is deduced using a constrained Smith-Waterman algorithm.

Strict clustering applies more stringent criteria to joining sequences, with high levels of sequence identity required over long regions. For instance, the TIGR Gene Indices clustering tools (TGICL) (Pertea et al., 2003) require 95% identity over a region > 40 nucleotides (nts) long, with < 20 bases of mismatched sequence at either end. Consequently, problems with chimeras and contaminants that artificially join clusters are avoided. On the other hand, a gene may be represented by several clusters which makes interpretation more difficult.

3.2.3 Transcript assembly

Redundancy is commonly further reduced and error rates minimised in clustering programs by generating contiguous consensus sequences (“contigs”) for each cluster. In essence, this constitutes the reconstruction of the mRNA sequence from
which the ESTs in the cluster were derived. Contigs are generally longer than constituent ESTs, and therefore contain more information (Ewing et al., 1999–2000). One caveat with using contigs, however, is that they do not correspond to experimental data and may therefore be erroneous.

The most commonly used assembly algorithms in sequencing projects have been Phrap (www.phrap.org) and CAP3 (Huang and Madan, 1999). Phrap relies heavily on quality values and is not well suited for assembly of sequence-only data. Furthermore, Phrap considers sequence regions with only one supporting read as unreliable, leading to a tendency to trim off ends of contigs. Like Phrap, CAP3 was designed for genome assembly. A recent comparison has shown that CAP3 gave the best overall performance for EST assembly (Liang et al., 2000).

Previous algorithms were derived from programs used in genome sequencing to assemble fragments into single contiguous sequences. The primary issue therefore has been to deal with repeat regions in an efficient manner. However, repeats are not a major problem with ESTs. Moreover, with alternative splicing frequencies as high as 60% (Mironov et al., 1999; Modrek and Lee, 2002), a contiguous sequence may not be the best way to represent a cluster, and the performance of genome assemblers on ESTs is questionable (Heber et al., 2002). Recent improvements in the area of transcript assembly have therefore focused on alternative splicing. CAP4 (www.paracel.com) is an improvement of CAP3 in so much as it is specifically designed for EST assembly. It is currently a part of the Paracel TranscriptAssembler (PTA) package, a pipeline that combines filtering, clustering and assembly functions.

3.3 Databases

The rapid increase in DNA sequence production demands efficient and structured storage of data in databases. Furthermore, the interfaces must be user-friendly and provide easy access to pertinent biological information. Concurrently with the growth in sequence data, the number of databases has increased rapidly in recent years (Galperin, 2004), and merely listing, let alone describing all the databases is beyond the scope of this thesis. This section deals with some of the databases that merit further scrutiny and serve as representative sources for characterisation of short DNA sequences. A summary of the databases is shown in Table 2. Protein databases are not described although they are mentioned in section 4.1.

3.3.1 Sequence tag databases

Since its inception in 1988, The National Center for Biotechnology Information (NCBI) has hosted several of the most important nucleotide and protein databases, among them GenBank (Benson et al., 2004). In essence, GenBank consists of nucleic acid sequence data from the scientific community. EST data are the major sources of new sequence records, and there is a database dedicated solely to
Table 2. Listing of some common DNA sequence databases.

<table>
<thead>
<tr>
<th>Database</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Short tag databases</td>
<td></td>
</tr>
<tr>
<td>GenBank, dbEST</td>
<td><a href="http://www.ncbi.nlm.nih.gov/">www.ncbi.nlm.nih.gov/</a></td>
</tr>
<tr>
<td>DDBJ</td>
<td><a href="http://www.ddbj.nig.ac.jp/">www.ddbj.nig.ac.jp/</a></td>
</tr>
<tr>
<td>EMBL</td>
<td><a href="http://www.ebi.ac.uk/embl.html">www.ebi.ac.uk/embl.html</a></td>
</tr>
<tr>
<td>2. Gene index databases</td>
<td></td>
</tr>
<tr>
<td>RefSeq</td>
<td><a href="http://www.ncbi.nlm.nih.gov/RefSeq/">www.ncbi.nlm.nih.gov/RefSeq/</a></td>
</tr>
<tr>
<td>TIGR gene indices</td>
<td><a href="http://www.tigr.org/tdb/tgi.shtml">www.tigr.org/tdb/tgi.shtml</a></td>
</tr>
<tr>
<td>GeneNest</td>
<td>genenest.molgen.mpg.de/</td>
</tr>
<tr>
<td>STACK</td>
<td><a href="http://www.sanbi.ac.za/Dbases.html">www.sanbi.ac.za/Dbases.html</a></td>
</tr>
<tr>
<td>PopulusDB</td>
<td><a href="http://www.populus.db.umu.se/">www.populus.db.umu.se/</a></td>
</tr>
<tr>
<td>3. Polymorphism databases</td>
<td></td>
</tr>
<tr>
<td>SNP consortium</td>
<td>snp.cshl.org/</td>
</tr>
<tr>
<td>HGBASE</td>
<td>hgvbase.cgb.ki.se/</td>
</tr>
<tr>
<td>topoSNP</td>
<td>gila.bioengr.uic.edu/snp/toposnp</td>
</tr>
<tr>
<td>Monsanto Arabidopsis polymorphism and Ler sequence collection</td>
<td><a href="http://www.arabidopsis.org/Cereon/">www.arabidopsis.org/Cereon/</a></td>
</tr>
</tbody>
</table>

ESTs, dbEST (Boguski et al., 1993). As part of an international collaboration, GenBank incorporates sequences submitted to the DNA Data Bank of Japan (DDBJ) (Miyazaki et al., 2004) and the EMBL Data Library (Kulikova et al., 2004). By August 2004, GenBank contained 37.3 million DNA sequences. The sites provide a number of tools for retrieving sequences and for doing similarity searches. To meet the demands for more efficient searches, it is possible to search genome-specific databases at NCBI (Wheeler et al., 2004).

In contrast to EST data, similarity searches cannot be performed against tag sequences generated by SAGE or MPSS because they are too short. One strategy has been to compile dictionaries of tags, where \textit{in silico}-derived tags are mapped to sequences in EST or gene index databases. SAGEmap (Lal et al., 1999; Lash et al., 2000) maps SAGE tags to UniGene clusters (see 3.3.2), constructed for tags generated through the Cancer Genome Anatomy Project (CGAP) (Strausberg et al., 2000). Unambiguous tag-to-gene mapping is difficult for several reasons. First, not all tags are unique to a specific gene (Stollberg et al., 2000). Second, one gene may be represented by multiple clusters, so a single tag may map to more than one cluster. Third, the high error rate in ESTs leads to tags mapping to inappropriate clusters (Pleasance et al., 2003).

Recently, similar tag mappings have utilised the information in the genome sequences (Jongeneel et al., 2003; Iseli et al., 2002; Meyers et al., 2004b), which on...
3. DATA ORGANISATION AND DATA FLOW

average are of higher quality. The use of longer tags increases the specificity of tag mapping, but does not necessarily imply complete tag uniqueness. For instance, 12.5% of 20-bp tags extracted from the A. thaliana genome originated from multiple positions in a study by Meyers et al..

3.3.2 Gene index databases

One consequence of large-scale EST sequencing is data redundancy. Highly expressed genes will be sampled at higher rates, leading to multiple observations of transcripts. In order to reduce redundancy and facilitate interpretation, methods have been designed to cluster (see also section 3.2.2 on page 21) EST data into groups representing distinct genes. A non-redundant gene catalogue for an organism provides an estimate of gene number.

The UniGene database (Schuler et al., 1996; Schuler, 1997) is a system for automatically partitioning sequence data into clusters. Sequences consist of both ESTs and well-characterised mRNAs from dbEST and GenBank. Emphasis is laid on removing low quality and non-gene sequences. UniGene relies on loose clustering, which may lead to the improper joining of clusters due to chimeras and other artifacts. Similarity searches are commonly performed against a non-redundant version of UniGene, in which each cluster is represented by the cluster member with the longest high-quality region of sequence data (Bouck et al., 1999). Consequently, the representative sequence will not cover either the entire gene or the 3’-end in most cases.

Instead of using representative sequences, the Sequence Tag Alignment and Consensus Knowledgebase (STACK) database generates a contiguous sequence for each cluster (Christoffels et al., 2001). In this way, redundancy is eliminated and a larger portion of the transcript is covered. STACK clusters are built from sequences in dbEST that are categorised according to their tissue of origin. This implies that cluster sizes are reduced and that alternative transcripts commonly form separate groups, according to the tissue-specificity of alternative splicing. GeneNest (Haas et al., 2000) is another resource that provides consensus sequences.

TIGR Gene Indices (Quackenbush et al., 2001) provide the largest resource of gene indices in terms of the number of represented species. Indices are built with the TGICL tools using ESTs in dbEST, mRNAs in GenBank and sequences at The Institute for Genome Research (TIGR). Cluster assembly is done with CAP4, resulting in tentative consensus (TC) sequences. The stringent clustering approach leads to closely related genes and splice variants being represented by separate TC sequences.

While previously described databases have high transcriptome coverage, problems inherent to clustering may lead to invalid representations of the original mRNAs. The RefSeq database (Pruitt et al., 2003), on the other hand, is a resource that provides high-quality, manually curated reference sequences for a variety of organisms. Moreover, the amount of additional information, such as annotations of protein domains, is greater for RefSeq records.
3.3.3 Genetic variation databases

Studies of genetic variation in populations are important for several reasons. Analyses of molecular pattern variation hold the potential to reconstruct the evolutionary history of populations (Wakeley et al., 2001; Reich et al., 2002), and polymorphic markers are commonly used for association studies (Botstein and Risch, 2003). Recently, much interest has been turned to single nucleotide polymorphisms (SNPs) due to their abundance (Brookes, 1999) and importance as genetic markers (Kwok and Gu, 1999). In addition, the discovery that haplotypes tend to be inherited in blocks (Patil et al., 2001; Gabriel et al., 2002) has initiated large-scale SNP mapping projects (The International HapMap Consortium, 2003). Consequently, a demand for a SNP catalogue has emerged.

In response to this demand, several SNP databases have emerged, mainly for humans (Sherry et al., 2001; Brookes et al., 2000), but also for other species (Jander et al., 2002). SNPs which occur in coding sequence may induce amino acid changes and are therefore of special interest (Clifford et al., 2004), as exemplified by the topoSNP database (Stitziel et al., 2004).
Chapter 4

Analysis and applications of cDNA sequences

EST datasets have been produced in ongoing projects for a wide variety of source tissues and organisms including humans (Okubo et al., 1992; Hillier et al., 1996), mouse (Marra et al., 1999), *C. elegans* (Parkinson et al., 2001), *D. melanogaster* (Rubin et al., 2000), *Arabidopsis* (Asamizu et al., 2000), rice (Yamamoto and Sasaki, 1997), and *Populus* (Sterky et al., 1998), to mention but a few. The sequence datasets have been used in a wide range of applications. Previously unknown genes have been discovered and characterised, and gene expression profiling has allowed the detailed description of various tissues.

With the continuous completion of more and more genomes, EST data sets have been used in new applications. Deriving from coding regions of a genome, they are valuable aids for determining intron/exon boundaries in genome annotation projects (Imanishi et al., 2004), identifying alternative splicing (Brett et al., 2000; Modrek et al., 2001), mining polyadenylation signals (Beaudoin et al., 2000), determining alternative polyadenylation (Iseli et al., 2002), and for finding polymorphisms such as SNPs (Picoult-Newberg et al., 1999).

This chapter briefly examines applications and analysis methods for tag sequence data that have been employed in the work underlying this thesis. These are tag identification via database searches and sequence alignments, expression profiling and gene discovery, genome annotation and, finally, analysis of genetic variation.

4.1 Identification and characterisation of cDNA sequences through database searches

One of the primary uses of high-throughput tag sequencing is comparing experimental data to databases in order to identify and characterise the transcripts from which the tags derived. To reduce the influence of noise the first step is to search a
4. ANALYSIS AND APPLICATIONS OF cDNA SEQUENCES

database of sequences associated with the organism under study. The unambiguous match of a tag to a database entry identifies the transcript, and may provide functional information for annotated entries. If no annotation is given, or no match is obtained, characterisation is still possible through similarity searches against databases from other genera. Functions may then be inferred in some cases based on homology (Attwood, 2000).

4.1.1 Identification via tag mapping

Methods that generate < 20 bp tags, such as SAGE and MPSS, rely on reference datasets in which in silico-derived tags have been mapped to cDNA sequences (Lash et al., 2000) or genomic sequences (Meyers et al., 2004a). Identification merely involves exact string-matching to the reference dataset (Tuteja and Tuteja, 2004). However, caution must be taken in the interpretation of the tag identification results, for reasons related to the tag-mapping procedure (see section 3.3.1). In addition, sequencing errors will lead to erroneous tag identification. Given that the average error rate per sequenced base is 1%, a 10-bp tag will have a 10% probability of containing an erroneous base. This introduces a potential for misassigning tags, even if recent studies show that the error rate per tag may be as low as 2% (Wang, 2003). Methods that attempt to correct the impact of sequencing errors have been developed (Colinge and Feger, 2001).

4.1.2 Identification via sequence alignment

In the absence of sequencing errors, the string-matching approach would also be applicable to longer tag sequences. Both database entries and tags, however, do contain sequencing errors, as well as polymorphisms, which in practice means identical lookups are scarce. Therefore, for longer tags, such as ESTs, alignment algorithms are employed, and identity is inferred from near-identical matches.

Sequence alignment algorithms have been used since the 1970s for studying the evolutionary relationships between proteins (Ouzounis and Valencia, 2003), with the Needleman-Wunsch algorithm for global alignments (Needleman and Wunsch, 1970) being one of the earliest examples. Local alignments were introduced in 1981 by Smith and Waterman for identifying locally conserved regions between two sequences. Both the Needleman-Wunsch and Smith-Waterman methods are based on dynamic programming and require time proportional to the product of the lengths of the sequences being compared (Altschul, 1998). Thus, due to the rapid increase in database size, heuristic algorithms have been developed to reduce computing time. Fast alignment (FASTA) (Pearson, 1990) and basic local alignment search tool (BLAST) (Altschul et al., 1990, 1997) are the most commonly used algorithms today. Both algorithms determine all instances of k-tuple words that are common to two sequences, and extend the alignments from these matches.
Expression profiling and gene discovery

The sequence alignment approach is inappropriate for short sequence tags. Aligning tags < 20 bp long to databases will produce a multitude of spurious matches due to sequencing errors and polymorphisms. Furthermore, it is difficult if not impossible to distinguish between gene family members.

4.1.3 Characterisation via homology

Closely related to sequence identification is the task of characterising the function of a transcript. Even if a transcript can be identified, for instance as belonging to a given gene index cluster, functional information may be completely lacking (Cirera et al., 2000). In such cases, or when the transcript cannot be unambiguously identified, matches to similar DNA sequences are investigated in searches of informative annotations. Inference of function is based on two sequences being homologous, which in turn can be inferred from high similarity. However, sequence similarity is not always due to homology (Eisen, 1998), and sequence similarity search algorithms perform badly when searching with sequences < 200 bp (Anderson and Brass, 1998). If the DNA sequence contains coding regions it is preferable to search protein databases (Nicholas et al., 2000).

4.2 Expression profiling and gene discovery

EST databases provide an important source of gene expression data (Ewing and Claverie, 2000), and new methods that increase tag throughput (e.g. SAGE and MPSS) seemingly fulfil the scale requirement of functional genomics (Claverie, 1999). Furthermore, since there is no limit to the number of tags that can be sequenced, rare transcripts whose expression levels lie under the detection limit of arrays may be studied. Finally, an additional benefit of tag sequencing is its ability to discover new genes\(^1\), since new tags correspond to previously unidentified genes.

At present the most widely used method for analysing global gene expression patterns is DNA microarray analysis (Schena et al., 1995; Duggan et al., 1999; Lockhart and Winzeler, 2000). In this approach, labelled RNA or cDNA samples are hybridised to arrays (usually glass slides) containing immobilised probes or cDNA clones. Changes in hybridisation intensities indicate changes in gene expression levels. However, gene family members may cross-hybridise, potentially confounding analysis. Furthermore, the use of cDNA microarrays for gene discovery is limited.

\(^1\)I make a distinction between gene discovery, as it is described here, and gene finding/prediction, by which I mean *in silico* gene finding in genomes.
4. ANALYSIS AND APPLICATIONS OF cDNA SEQUENCES

4.2.1 Expression profiling

Tag sequence expression profiles are digital in nature, with tag counts related to gene expression level. In the case of EST sequencing, 3’-directionally cloned cDNA libraries are preferred. Since 3’-UTRs are less conserved than coding regions (Makalowski and Boguski, 1998) and relatively long (Pesole et al., 2002), 3’-ESTs are suitable for fingerprinting transcripts. However, the technical difficulties involved in sequencing through the poly(A)-tail makes 5’-sequencing more viable in practice.

Transcript profiles are obtained by clustering 5’-ESTs and counting the members of each cluster. However, this representation of gene expression is not entirely correct. 5’-reads originating from the same transcript may not overlap due to the partial nature of cDNA clones. SAGE, MPSS and other short tag-sequencing protocols rely on restriction enzymes to define unique 5’ starting points for the transcript. The drawback of this is that a proportion of transcripts will be left out.

Expression profile datasets can be subjected to various forms of analyses. Identification of high-abundance transcripts can reveal genes that are important to the tissue in question (Channeliere et al., 2002). Comparison of multiple datasets allows the study of differentially expressed genes (Bhalerao et al., 2003). Alternatively, comparison of datasets from different organisms may reveal homologous relationships (Kirst et al., 2003). To date, data production has been quicker than expression analysis. Many of the publicly available tag libraries still contain large amounts of data that remain to be mined (Bortoluzzi and Danieli, 1999; Ewing and Claverie, 2000).

Ascertainment of differential expression requires a statistical framework for determining levels of significance. EST programs are unsuitable for detecting differential expression of all but highly expressed genes, due to costs related to the redundancy in sampling (Soares, 1997; Ishii et al., 2000).

Several statistical tests have been developed for pairwise comparisons of tag libraries (Audic and Claverie, 1997; Zhang et al., 1997; Kal et al., 1999), and the performance of the tests has been compared and benchmarked (Man et al., 2000; Ruijter et al., 2002). These tests cannot be applied to multiple library comparisons so various alternatives have been proposed (Greller and Tobin, 1999; Stekel et al., 2000). Software tools that implement these tests and provide features for visualisation and data management are available (Larsson et al., 2000; Margulies and Innis, 2000).

The establishment of expression patterns over different tissues and conditions provides insights into function and may reveal regulatory pathways. In multilibrary analyses, the expression of a transcript can be represented as a vector, with each entry corresponding to the expression level in one library. Hierarchical clustering methods were first applied to microarray expression data (Eisen et al., 1998) to group genes with similar expression patterns, thus revealing functional and regulatory relationships between genes. Similar methods have also been introduced for tag data (Ewing et al., 1999).
4.2.2 Gene discovery

EST gene discovery programs have been applied to various tissues and organisms (Adams et al., 1995; Marra et al., 1999; Dimopoulos et al., 2000; Blackshear et al., 2001; Whitfield et al., 2002). Gene cataloguing approaches generally offer quick means to identify highly expressed genes.

As pointed out in section 2.2.1, since multiple copies of specific mRNAs are present in sources of ESTs, the resulting data inevitably contain redundancies, and consequently, gene discovery rate will diminish as more of a library is sequenced. For instance, in 1996, 10.6% of ESTs represented new genes, but already by 1998, the figure had dropped to only 2.7% (Wang et al., 2000). Normalisation and subtraction techniques (Bonaldo et al., 1996; Soares, 1997) are usually applied to allow further sequencing. However, normalisation only reduces redundancy within libraries. Serial subtraction of normalised libraries has been developed to minimise the presence of ubiquitously expressed mRNAs (Scheetz et al., 2004). Since transcript tissue origin is one of the chief interests in gene discovery, tissue-specific sequence tags have been used to label ESTs (Gavin et al., 2002). The even higher throughput of short-tag techniques have lowered the detection limits of rarely expressed transcripts even further. For instance, as well as EST sequencing the CGAP project also employs SAGE for gene cataloguing (Strausberg et al., 2000). Furthermore, SAGE has become a widely used gene discovery approach (Fizames et al., 2004; Chen et al., 2002; Moreno et al., 2001; Patankar et al., 2001) and it has been put forward as a means to generate a complete catalogue of gene transcripts (Boheler and Stern, 2003; Stern et al., 2003).

Recent studies with MPSS have generated $> 10^7$ signatures for human (Jongeneel et al., 2003) and A. thaliana (Meyers et al., 2004b) cell lines, accomplishing redundant coverage of the transcriptomes. In both cases, several thousand signatures could be mapped to unannotated genomic regions, identifying new transcript variants.

Now that detection limits have been diminished to the point where single-copy transcripts can be observed, new findings have emerged. First, the current estimate of 30,000–40,000 human genes has been questioned by several studies (Kapranov et al., 2002). Second, it has become evident that antisense RNAs (Vanhee-Brossollet and Vaquero, 1998) and non-coding RNAs (ncRNAs) are more common than previously thought (Eddy, 2001; Mattick, 2001). These phenomena are amenable to study as shown by Meyers et al.

4.3 Genome annotation

The annotation of a genome links sequences to the biology of an organism (Stein, 2001). Tremendous effort is therefore being put into the annotation process, either through annotation jamborees where scientists gather and commit a short period of time to genome annotation (Adams et al., 2000; Kawai et al., 2001; Imanishi
4. ANALYSIS AND APPLICATIONS OF cDNA SEQUENCES

et al., 2004), or through dedicated longterm projects such as Ensembl (Birney et al., 2004a).

In early stages of annotation, most attention is paid to mapping genetic markers such as sequence tagged sites (STSs), repetitive elements and genes. Obviously, short cDNA tags are of great help since gene loci are easily identified in the genomic sequence with simple similarity searches (Bailey et al., 1998; Saha et al., 2002). For instance, SAGE tag mapping has been applied to the human genome to pinpoint transcriptionally active regions (Caron et al., 2001). Short tags are also finding use in gene prediction. A trend in gene finding programs has been to corroborate computational predictions with experimental evidence, as seen for instance in GenomeWise (Birney et al., 2004b) and GenomeScan (Yeh et al., 2001).

Another primary issue is assessment of gene function. Of necessity, gene annotation is a largely automated process (see for instance Curwen et al., 2004) that relies on the quality of data in the available databases. Sequence similarity searches quickly become out of date (Wheelan and Boguski, 1998), and many outdated annotations are transitively propagated through the databases (Boguski, 1999). Therefore, most functional annotations must be seen as hypothetical, awaiting manual curation and experimental confirmation (Ashurst and Collins, 2003).

Early comparisons of ESTs with human genomic sequences suggested that ESTs could provide information about gene structure and phenomena such as alternative splicing (Wolfsberg and Landsman, 1997). Several recent studies have employed this approach successfully (Burke et al., 1998; Mironov et al., 1999; Modrek et al., 2001; Eyras et al., 2004). Other features, including alternate polyadenylation (Gautheret et al., 1998; Beaudoin and Gautheret, 2001; Iseli et al., 2002) and variant polyadenylation signal usage (Beaudoin et al., 2000; Legendre and Gautheret, 2003), have been detected. However, the fragmentary nature of short cDNA sequences inevitably gives a partial view of gene structure. Many of the annotation efforts have therefore included large full-length cDNA (FLcDNA) collections (Imanishi et al., 2004; Seki et al., 2002; Kawai et al., 2001) since full-length sequences improve annotation (Haas et al., 2002).

4.4 Analysis of genetic variation

The redundancy in EST data collections can be turned to an advantage in the search for sequence polymorphisms. Different cDNA libraries often originate from different individuals, and even in the case of libraries from single individuals, variations between paternally and maternally inherited alleles can be assessed. Short sequence polymorphisms are used as markers for genetic traits, genome mapping, and tools for understanding genome evolution (Syvänen, 2001).

Recently, much attention has been turned to SNPs. In its simplest form, SNP discovery is a matter of analysing multiple sequence alignments and identifying alignment slices where bases differ. However, the low-quality nature of ESTs however makes it difficult to distinguish between true discrepancies and those intro-
duced by sequencing errors. Furthermore, paralogous ESTs with high sequence similarity are likely to be grouped in the same cluster, which introduces additional false SNP candidate sites. Several SNP-finding methods have been developed, with different solutions to these problems. Given a SNP candidate, most methods exploit base quality values to detect sequencing errors (Picoult-Newberg et al., 1999; Buetow et al., 1999; Marth et al., 1999). The obvious drawback of this is that trace data is missing in many EST data collections. Alternatively, a conservative approach can be applied, in which SNP redundancy combined with cosegregating haplotypes are used to distinguish SNPs from sequencing errors (Barker et al., 2003; Kota et al., 2003).

It is only recently that SNPs have been identified for plants, so other markers have been employed in their analysis. For example, simple sequence repeats (SSRs) are short polymorphic molecular markers that are extremely important in plant breeding (Saha et al., 2004). Early studies used genomic DNA for marker development, but data mining of ESTs has provided a quicker and cheaper option. It is estimated that 1–5% of plant ESTs contain SSRs (Kantety et al., 2002) for marker development.

Comparative genomics holds much promise for unravelling facts about species evolution, species origin and genetic processes (Wei et al., 2002). For species without complete genome sequences, EST collections are the best current alternative (Rudd, 2003). Comparisons of gene indices have revealed genes shared between different plants (Ewing et al., 1999–2000; Ronning et al., 2003). Plant evolution has been studied by aligning species-specific EST data sets to *A. thaliana* sequences (Kirst et al., 2003; Nishiyama et al., 2003). Finally, orthologous comparisons reveal substitution rates in coding regions and UTRs (Makalowski and Boguski, 1998; Fulton et al., 2002).
Part II

PRESENT INVESTIGATION
Objectives

The work presented in this thesis has focused on various methods for analysing cDNA sequence tags. Three major research topics can be distinguished. First, papers I and II describe how short sequence tags can be used to monitor gene expression and the difficulties involved in identifying the expressed genes with the aid of short tags. Second, paper III deals with the bioinformatics of a large EST project and various ways of gleaning information from large data sets. Finally, papers IV and V are mainly concerned with approaches for identifying sequence polymorphisms and sequence differences.

In the following sections, overviews of each project are presented, along with summaries of the most important findings.
Chapter 5

Analysing gene expression with short sequence tags (I, II)

5.1 Using pyrosequencing for gene expression analysis (I)

The benefits of pyrosequencing as compared to Sanger sequencing were alluded to in section 2.1.2. Ease of sample handling and reduced laboratory time promoted pyrosequencing as an option for high-throughput global transcript sequencing. With this in mind, a pilot study was performed to evaluate the potential use of pyrosequencing for gene expression analysis.

A model system for human atherosclerosis was used for validation. Two cDNA libraries, representing foam cells and macrophages, were constructed. mRNA was purified and transformed into cDNA, followed by fragmentation of the cDNA with restriction endonuclease DpnII and directed cloning of the 3’ ends. 1000 clones from each library were sequenced from the 5’ end using both pyrosequencing and Sanger sequencing, producing a pyrosignature and an EST for each clone. A gene identifier was assigned to each sequence tag by performing database searches. The pyrosignature identifier was unique and identical to the corresponding EST identifier for 75% of the clones (85% for pyrosignatures longer than 18 bp). Consequently, consistent results were obtained using both methods.

The average read length of the pyrosignatures in this study was 24–25 bp, which is longer than the tags generated by both SAGE and MPSS. This is advantageous since it reduces problems with ambiguities in gene identification.

5.2 The importance of reliable databases (II)

The feasibility of using pyrosequencing to generate tag sequences for gene expression studies was clearly shown in paper I. Although it was intuitively clear that longer tag sequences enhance in silico transcript identification, no study had pinpointed the number of bases required to ensure tag uniqueness. Furthermore,
issues such as gene index database quality were not taken into account. These potential sources of transcript identification error were investigated using publicly available human sequence data.

The short tag sequencing protocols described in sections 2.3.2-2.3.3 could lead to additional problems since sample preparation requires the use of restriction endonucleases (Velculescu et al., 1995; Spinella et al., 1999; Brenner et al., 2000). Various properties of the data sets vis-à-vis restriction enzyme recognition sites were studied to estimate the effect of enzyme choice.

The studies in paper II can be divided into three distinct parts, as outlined below.

5.2.1 Restriction enzyme analysis

Due to its high quality, the RefSeq gene index database (Pruitt et al., 2003) was chosen for the analysis of characteristics related to recognition sites. At the time of analysis, it was estimated that RefSeq covered 10–15% of the human transcriptome.

The entries were exposed to in silico cleavage by enzymes that utilise 4-bp recognition sites. Generally, recognition sites are palindromic, but for the sake of generality, all 256 possible 4-mer combinations were used.

Any given 4-bp sequence would occur on average once in every 256 bp in a random DNA sequence. Since most mRNAs are longer than 1 kb, the majority of transcripts are expected to contain recognition sites of any given kind (Yamamoto et al., 2001). The distribution of recognition sites in the RefSeq dataset provided an empirical estimate of the fraction of transcripts that lack a given recognition site, thus identifying the enzymes which empirically have the highest cutting frequencies and would be most apt for use in short tag sequencing protocols.

For transcripts that contained recognition sites, 10-bp tags were extracted from the bases following the 3' most recognition site. As pointed out previously, not all transcripts will be unambiguously identified at these sequence lengths. The fraction of unique tags was assessed for each recognition site, giving percentages in the range 89–92% for most sites. Increasing tag length only moderately decreases ambiguous identification, with approximately 94% tags being unique at a tag length of 30 bp. A related study assuming 15,000 genes estimated 94% 10-bp tags to be unique (Stollberg et al., 2000). Recently, Pleasance et al. (2003) showed, using finished genome sequences of D. melanogaster and C. elegans, that 6% and 12% of the genes in the respective organisms produce ambiguous 10-bp tags. They also suggested that analyses such as those presented here allow for the preselection of a restriction enzyme, possibly organism-specific, that generates the best representation of the mRNA population.

Furthermore, the number of bases from the cleavage site to the poly(A)-tail was determined. Given too few intermediary bases, the resulting tag would extend into the poly(A)-tail and thus be more likely to be ambiguous. The number of intermediary bases would also vary due to polyadenylation cleavage site hetero-
geneity (Pauws et al., 2001; Iseli et al., 2002), leading to multiple tags being derived from the same gene.

We finally extracted all tags that derived from upstream cleavage sites, and compared this set to the set of tags derived from the 3’-most cleavage site. The intersection between these sets was non-empty for all recognition sequences, implying that tag counts would be shifted from their true values were upstream tags to be sequenced. This scenario has been reported for SAGE (Welle et al., 1999). One possible explanation for this phenomenon is alternative splicing. Splicing of an exon containing the 3’-most cleavage site would generate a tag from an upstream cleavage site, if present.

5.2.2 Transcript identification using sequence alignments

Large-scale transcript profiling generates a vast amount of sequences that need to be identified in a quick and automated manner. Therefore, we sought to evaluate the possibility of using the BLAST algorithm to identify a set of short virtual tags derived from EST data and sequences in the RefSeq database. The problems of using sequence alignments for short sequences has previously been discussed (section 4.1.2). However, BLAST was chosen partly due to its popularity, and partly because other algorithms would be too slow to meet the time requirements of high-throughput sequencing.

The objective was to investigate the effect increasing tag length has on ambiguity of identification. For each query sequence in the study, tags with lengths of 10 to 50 bp were extracted from the 3’-most recognition site and aligned to the UniGene non-redundant database (see section 3.3.2). Since the goal was to identify transcripts, queries that had alignments with less than 90% bases identical to the hit sequence were removed. The effect of increasing tag length was studied in two ways. First, for each query, the number of equally scoring best hits were recorded as a function of tag length. The average number of hits per query were then plotted against tag length. Second, we defined the decision length as being the tag length from which a query achieved unambiguous identification, and that the same identifier was retained for all increasing tag lengths. The fraction of queries with a given decision length was plotted against tag length. Both plots indicated that a tag length of 16–17 bp was sufficient to unambiguously identify ≈ 90% of the transcripts (Figure 9).

There are several reasons why 100% unambiguous classification is not attained. First, we identified cases where merging distinct UniGene clusters would be appropriate. Consequently, any clustering problems will affect the results and interpretation of the identification process. Second, transcripts from genes belonging to the same families will likely not be distinguishable with short tags.
5. ANALYSING GENE EXPRESSION WITH SHORT SEQUENCE TAGS (I, II)

Figure 9. Transcript classification by BLAST as a function of tag length (excluding recognition site). The left y-axis represents the number of hits per query, and the right y-axis represents the fraction of stable queries. Results are for a pooled EST data set, using the NlaIII restriction enzyme.

5.2.3 Tag mapping

Finally, we compared the difference between identification by full-length sequence alignment and SAGE tag mapping. For this, EST data were aligned to the UniGene non-redundant database, providing a UniGene identifier for each EST query. This identifier was compared to the one obtained by extracting the corresponding SAGE-tag and looking up the tag in the SAGEmap database. The results were in concordance for 80% of the cases.

The majority of the discrepancies could be attributed to the UniGene reference sequence selection. SAGE tags derive from the 3′-most recognition site, whereas reference sequences may lack sequence pertaining to this region. Consequently, different analysis methods, here represented by ESTs and SAGE, may in some cases produce different identifications.
Chapter 6

An EST resource for *Populus* (III, IV, V)

The genus *Populus* has emerged as the model system for forest trees, due in part to its rapid growth rate, small genome size, and ease of cloning (Bradshaw et al., 2001). During 2004, it is expected that the 520 Mbp genome of *Populus trichocarpa* will be released. Although many topics of plant biology can be studied in *Arabidopsis*, many features of trees cannot, since they have long lifespans and develop woody tissue (Brunner et al., 2004).

*Populus* EST sequencing efforts have been ongoing at Umeå University (UmU), Swedish University of Agricultural Sciences (SLU) and KTH since 1997. When sequence assembly was performed in 2003, more than 140,000 sequence reads had been generated for 18 different cDNA libraries. The libraries were derived from three *Populus* taxa, aspen (*Populus tremula*), a hybrid aspen (*P. tremula × tremuloides*) and black cottonwood (*P. trichocarpa*).

Paper III describes the organisation of these data into a public EST resource for functional genomics, along with an analysis of poplar gene composition. A web-based database, PopulusDB (www.populus.db.umu.se/), was designed to provide easy access to the ESTs. The sequences were clustered and assembled with PTA, generating 11,885 clusters and 12,759 singletons. Sequences from different species generally clustered together, indicating that *Populus* DNA sequences are conserved over the included taxa.

Comparison of assembled contigs with the *Arabidopsis* proteome showed that most genes have close homologs in *Arabidopsis*. A TBLASTX search of contigs against the *Arabidopsis* genome also showed that *Populus* is of little use for gene finding in *Arabidopsis*. Consequently, the gene content between species is very similar, and does not support the suggestion that *Arabidopsis* has lost 5–10% of its genes during evolution (Allen, 2002). However, estimating gene content from EST data is precarious, so the validity of these conclusions must await future verification.
Since the libraries were non-normalised, the assembly was used to generate expression profiles over the tissues. These profiles represent averages over individuals and taxa and are thus crude approximations, but still they convey important information. Hierarchical clustering of expression profiles over libraries resulted in libraries of similar origin clustering together. Similar results were obtained when clustering was performed over genes.

In summary, the Populus EST resource will provide an important tool for functional genomics in plant biology. Currently, the data are being used to train gene prediction programs that are to be used in the annotation of the P. trichocarpa genome, hosted by the International Populus Genome Consortium (IPGC) (www.ornl.gov/sci/ipgc/). The data set has been augmented with Populus EST sequences from other sequencing laboratories around the world. I have been responsible for assembly of this data set and preparing the data for the annotation process.

6.1 SNP discovery in EST data (IV)

As the Populus EST resource consists of data originating from several species and individuals, there is an opportunity to answer questions about genetic variation within (polymorphism) and between (divergence) species. Polymorphism discovery in plant EST data sets has been utilised for marker development (Nasu et al., 2002; Grivet et al., 2003; Liewlaksaneeyanawin et al., 2004). SNPs have lately become the marker of choice, promising a wide range of applications to crop genetics (Rafalski, 2002). With this in mind, a project with the aim of developing SNP discovery was initiated.

Traditionally, SNP discovery has been an arduous, time-consuming and expensive procedure (Kwok and Gu, 1999). One of the attempts to reduce the effort and cost of SNP discovery utilises the abundance of cDNA data, which have proven their worth as valuable sources for SNP discovery (Marth et al., 1999; Picoult-Newberg et al., 1999; Buetow et al., 1999; Cox et al., 2001; Barker et al., 2003; Batley et al., 2003; Dantec et al., 2004). Identification of a candidate SNP is a matter of identifying discrepancies in multiple alignments, and discriminating between discrepancies caused by variation as opposed to sequencing error. This constitutes a binary classification problem, where the objective is to classify a data object into one of two classes; in this case, SNP or monomorphic. Current methods rely on Bayesian statistics for calling a SNP when a score exceeds a predefined threshold (Marth et al., 1999; Buetow et al., 1999; Dantec et al., 2004), or simply correlate SNPs with high-quality discrepancies (Altshuler et al., 2000; Barker et al., 2003).

We developed a solution that exploits the ideas of previous methods and draws advantage of the non-linear classification abilities of neural networks (Ripley, 1994). The general idea of our approach was to represent a candidate SNP by a vector \( x \) of \( n \) features, defined on some feature space \( F \), and let two networks classify \( x \). Although there is some flexibility in the choice of features, they should reflect factors that affect the decision process of manual SNP finding. Multiple align-
ments are scanned for columns where base discrepancies occur. As pointed out previously, the main issue is to discriminate between sequencing error and true variation. Thus, some of the most important factors are chromatogram shapes, base quality values, and major allele frequency. We represented a candidate SNP by the two latter characteristics, along with a Bayesian probability that the site be polymorphic.

Since neural network algorithms rely on training data with known class memberships, a carefully selected data set of validated and well-characterised SNPs were acquired from the CGAP web-site. Furthermore, an alignment column set corresponding to known monomorphic positions was constructed, reflecting the appearance of data obtained from typical cDNA projects. The data was partitioned into a training and test set for network training, and a validation set for method evaluation.

The networks consist of a set of reference vectors \( \{ \mathbf{m}_i ; \mathbf{m}_i \in \mathcal{F}, i = 1, 2, \ldots, k \} \), labelled as SNP or monomorphic. During training, the vectors are fine-tuned to optimise the decision border between the classes (Kangas et al., 1990). Classification is a matter of locating the reference vector \( \mathbf{m}_c \) that is closest (based on some distance measure) to the input vector \( \mathbf{x} \), and assigning the label of \( \mathbf{m}_c \) to \( \mathbf{x} \). An input vector was scored as a SNP candidate given that both networks unanimously classified the vector as a SNP.

A verification model was built from the data set in order to benchmark the performance of our method to other commonly used approaches. The results indicated performance characteristics comparable to those of the best method, with a slight improvement in the false negative rate.

The modular architecture of our application enables additional analysis features. For instance, by coupling the SNP discovery engine to UTR and protein databases, SNPs can be correlated with annotated gene regions. Although method development was conducted on cDNA sequence sets, SNP analysis can also be performed on assembled genomic sequence data, as shown in one of the evaluation studies, which was done on sequences from promoter regions of atherosclerosis gene candidates.

### 6.2 Comparison of two poplar EST data sets (V)

In paper V an in-depth analysis of the \( P. \) tremula and \( P. \) trichocarpa EST data sets is presented. Analysis of \( P. \) tremula x tremuloides was excluded since genetic variation studies in hybrids is inefficient. For \( P. \) tremula, 47,767 reads from 6 libraries went into the analysis; the corresponding numbers for \( P. \) trichocarpa was 22,980 reads from 3 libraries.

For one part, the study served to estimate the number of gene observations in either species and to survey the status of our EST sequencing programme. Separate assemblies were performed on libraries, species and the entire data set. The number of gene observations was estimated from the number of clusters and sin-
gletons, resulting in 14,213 genes for the entire data set. Similarly, 11,238 *P. tremula* and 5,911 *P. trichocarpa* genes were observed. Since the sum of the gene observations for the latter assemblies is higher than for the large assembly, identical genes have been observed in both species. This finding indicates the similarity between the species on the DNA level. The library assemblies were used to estimate library redundancy, giving estimates as low as 50%. Comparison with redundancy estimates obtained from the entire assembly showed that all libraries had > 70% redundancy, implying the need for normalisation or subtraction approaches for further gene discovery studies (Bonaldo et al., 1996).

Subsequent analyses focused on the species assemblies. The *Arabidopsis* proteome was used to identify transcripts in both species with entire CDSs. 353 *P. tremula* and 250 *P. trichocarpa* sequences covering entire *Arabidopsis* CDSs were identified. A noteworthy observation was the short average CDS length of ≈ 500 nt for both species. For comparison, the average coding length in *A. thaliana* is 1,900 nt (The *Arabidopsis* Genome Initiative, 2000). This indicates some of the shortcomings of EST sequencing, for instance that the assembly of long full-length sequences requires large sample sizes, cDNA inserts are truncated and that the ESTs might not cover the complete transcript.

![Box plots of sequence conservation by region for aligned ortholog pairs.](image)

**Figure 10.** Box plots of sequence conservation by region for aligned ortholog pairs.

In addition, we applied stringent criteria to identify a set of putative ortholog pairs, an important asset for comparative plant genomics (Fulton et al., 2002). Ortholog comparisons are a key to studying organism evolution (Gogarten and Olendzenski, 1999), but unfortunately, defining orthologs in plants is difficult since...
most plant genes are members of gene families, and their copy numbers vary significantly between species (Clegg et al., 1997). We found 89 putative ortholog pairs, which potentially can be used for comparison against other plants. Preliminary analyses showed extensive sequence similarity over both UTR and CDS regions (Figure 10).

Finally, the SNP discovery method was used to assess SNP characteristics in both species. The ortholog set was used to screen for common polymorphisms. However, only 2 were found, which may indicate that the species share few ancestral polymorphisms. More likely though is that the sample size is too small, especially for *P. trichocarpa*. 
This thesis has demonstrated the utility of computational analysis of cDNA sequences. Within this framework, numerous analysis approaches have been applied to cDNA sequence collections, and new methodologies have been developed.

The pyrosequencing method has been successfully applied to gene expression analysis. At the time of the pilot project (paper I), separation of Sanger sequencing fragments was done on slab-gels, and pyrosequencing included many manual steps. Both methods have undergone several improvements to increase sequencing throughput. However, the parallelisation and miniaturisation of pyrosequencing, as exemplified by 454 Life Sciences, makes this method seem more likely to boost sequencing throughput into a new dimension.

With more increase in sequence data production, databases will likely become even more littered with erroneous and dubious information. Studies like those conducted in paper II are essential for pinpointing problems that lead to ambiguous interpretations of data (Pleasance et al., 2003). Several issues with databases that affect transcript profiling studies have been identified.

Large EST data collections provide a gateway into understanding the biology of eukaryotes. The *Populus* EST resource will provide clues and answers to questions about the unique features of trees. Paper III has mainly been concerned with data organisation, the optimisation of which is important for efficient analyses.

The genetic variation within and between *Populus* species has been investigated, showing the high similarity between *P. trichocarpa* and *P. tremula* on a detailed level (paper V). This finding is of major importance for the *P. trichocarpa* genome sequencing project, since *P. tremula* genes can aid annotation efforts. Moreover, the development of software for identification of sequence polymorphisms (paper IV) has identified SNPs that are of potential use in marker development.

The development of more efficient methods for analysing cDNA sequences is an issue that will continue to be an important part of biotechnological research. Eventually, we will go from the current phase of data collection and data organi-
7. CONCLUDING REMARKS

...ation to one of hypothesis-driven research. The ongoing analysis of cDNAs and its contribution to biological knowledge is one step on the way.
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>adenine</td>
</tr>
<tr>
<td>ADP</td>
<td>adenosine di-phosphate</td>
</tr>
<tr>
<td>ATP</td>
<td>adenosine tri-phosphate</td>
</tr>
<tr>
<td>bp</td>
<td>base pair</td>
</tr>
<tr>
<td>BLAST</td>
<td>basic local alignment search tool</td>
</tr>
<tr>
<td>cDNA</td>
<td>complementary DNA</td>
</tr>
<tr>
<td>CDS</td>
<td>coding sequence</td>
</tr>
<tr>
<td>CGAP</td>
<td>Cancer Genome Anatomy Project</td>
</tr>
<tr>
<td>C</td>
<td>cytosine</td>
</tr>
<tr>
<td>DDBJ</td>
<td>DNA Data Bank of Japan</td>
</tr>
<tr>
<td>ddNTP</td>
<td>dideoxynucleotide</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>dNTP</td>
<td>deoxynucleotide</td>
</tr>
<tr>
<td>EMBL</td>
<td>European Molecular Biology Laboratory</td>
</tr>
<tr>
<td>EST</td>
<td>expressed sequence tag</td>
</tr>
<tr>
<td>FASTA</td>
<td>fast alignment</td>
</tr>
<tr>
<td>FLcDNA</td>
<td>full-length cDNA</td>
</tr>
<tr>
<td>G</td>
<td>guanine</td>
</tr>
<tr>
<td>GOLD</td>
<td>Genome Online Database</td>
</tr>
<tr>
<td>HGP</td>
<td>Human Genome Project</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>-------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>hnRNA</td>
<td>heterogenous nuclear RNA</td>
</tr>
<tr>
<td>IPGC</td>
<td>International Populus Genome Consortium</td>
</tr>
<tr>
<td>kb</td>
<td>kilobases</td>
</tr>
<tr>
<td>Mb</td>
<td>megabases</td>
</tr>
<tr>
<td>MPSS</td>
<td>massively parallel signature sequencing</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>NCBI</td>
<td>The National Center for Biotechnology Information</td>
</tr>
<tr>
<td>ncRNA</td>
<td>non-coding RNA</td>
</tr>
<tr>
<td>nt</td>
<td>nucleotide</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PP&lt;sub&gt;1&lt;/sub&gt;</td>
<td>pyrophosphate</td>
</tr>
<tr>
<td>PTA</td>
<td>Paracel TranscriptAssembler</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal RNA</td>
</tr>
<tr>
<td>SAGE</td>
<td>serial analysis of gene expression</td>
</tr>
<tr>
<td>SCF</td>
<td>standard chromatogram format</td>
</tr>
<tr>
<td>SLU</td>
<td>Swedish University of Agricultural Sciences</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>SSR</td>
<td>simple sequence repeat</td>
</tr>
<tr>
<td>STACK</td>
<td>Sequence Tag Alignment and Consensus Knowledgebase</td>
</tr>
<tr>
<td>STS</td>
<td>sequence tagged site</td>
</tr>
<tr>
<td>TC</td>
<td>tentative consensus</td>
</tr>
<tr>
<td>TGICL</td>
<td>TIGR Gene Indices clustering tools</td>
</tr>
<tr>
<td>TIGR</td>
<td>The Institute for Genome Research</td>
</tr>
<tr>
<td>T</td>
<td>thymine</td>
</tr>
<tr>
<td>UmU</td>
<td>Umeå University</td>
</tr>
<tr>
<td>UTR</td>
<td>untranslated region</td>
</tr>
<tr>
<td>U</td>
<td>uracil</td>
</tr>
</tbody>
</table>

52
Acknowledgements

Så står man där vid målet för de senaste fem årens slit... När jag blickar tillbaka på den tid som varit tycker jag inte det känns som om det är särskilt länge sedan jag klev upp på plan 6 (eller var det 2?) på Teknikringen 34 för att intervjus av Nipe. Tiden går som bekant fort när man har roligt, och det har jag verkligen haft under min tid på KTH. Det har varit roligt och stimulerande att gå till arbetet, och det beror till stor del på alla trevliga och härliga kollegor jag haft nöjet att lära känna genom åren. Ett jättestort tack vill jag därför rikta till:

Joakim – för att du antog mig som doktorand, för ditt uppmuntrande och engagerande sätt, och för dina stimulerande och roliga forskningsförslag. Tack också för att jag fick gå mina egna vägar och inte jobbade med det vi först kom överens om att jag skulle jobba med...


Mathias – kungen av bioteknik vars entusiasm aldrig tycks sina, den smittar av sig på alla som jobbar här. Det är roligt med forskning – lite som att hålla på med motocyrkla faktiskt.

Sophia, Per-Åke och Stefan – för att ni med säker handledning och vägledning bidrar till den goda stämningen på labbet, för att ni delar med er av era kunskaper, och för alla diskussioner som inte har med forskning att göra.

Magnus och Anders – tag-peket roddes så småningom i hamn, efter många möten då ni efter att jag förirrat mig in i svårigheter fick leda mig in på rätt spår igen.

Stefan Jansson – för all hjälp och vägledning i poppelprojektet. Nu vet jag nästan hur ett poppelträd ser ut.

Jag vill också rikta ett tack till övriga kollegor på labbet, gamla som nya:

Holmberg, Michael och Danko – datorgurusarna som vet allt om kisel och som lärt mig allt jag vet om kisel.

Innebandygänget – Olof, Ronny, Erik, Björn, Markus, Tove, Johan, Henrik, Emilie, Jocke, John m fl – många, långa matcher till 10 blev det...

Rumskamraterna – Henrik, för allt NHL-snack tidiga vintermorgnar, Maria & Jocke för ert ständigt goda humör, Malin, Jorge, Johan, Valtteri, Marcus för trevlig samvaro.


Ett tack också till dem utanför labbet, som påminner mig om det verkliga livet utanför matrixen:


Syster och Pentti – för allt stöd, semesterupplevelser, scooterturer, matupplevelser och roliga stunder.

Mor och Far – för att ni alltid trott på mig och uppmanat mig att göra det jag tycker är roligt. Och för att ni alltid ställer upp – att ni orkar! Jag vet inte vad jag skulle ta mig till utan er.

Peace. /Per

Stiftelsen för Strategisk Forskning (SSF) och Knut och Alice Wallenbergs stiftelse tackas för ekonomiskt stöd.
REFERENCES


57
REFERENCES


REFERENCES


Ewing RM, Claverie JM (2000). EST databases as multi-conditional gene expres-
sion datasets. Pac Symp Biocomput p. 0–42.

analyses of rice ESTs reveal correlated patterns of gene expression. Genome


detection platform for single nucleotide polymorphisms. Hum Mutat 19(5): 9–
85.

sis root transcriptome by serial analysis of gene expression. Gene identification

Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.
Science 269(5223): 6–512.

for gene prediction in human: from whole-genome shotgun reads to a global

Fujii S, Amrein H (2002). Genes expressed in the Drosophila head reveal a role for

Fulton TM, der Hoeven RV, Eannetta NT, Tanksley SD (2002). Identification, analy-
sis, and utilization of conserved ortholog set markers for comparative genomics


Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, et al. (2003). The
genome sequence of the filamentous fungus Neurospora crassa. Nature 422(6934):
9–68.

REFERENCES


64


65
REFERENCES


REFERENCES


REFERENCES


REFERENCES


Publications (I-V)