GeneSieve: A Probe Selection Strategy for cDNA Microarrays

Maulik Shukla

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Computer Science and Applications

Lenwood S. Heath, Ph.D., Chairman

Naren Ramakrishnan, Ph.D.

Ruth Grene, Ph.D.

T. M. Murali, Ph.D.

September, 2004
Blacksburg, Virginia

GeneSieve: A Probe Selection Strategy for cDNA Microarrays

by

Maulik Shukla

Committee Chairman: Lenwood S. Heath, Ph.D.

Computer Science

(ABSTRACT)

The DNA microarray is a powerful tool to study expression levels of thousands of genes simultaneously. Often, cDNA libraries representing expressed genes of an organism are available, along with expressed sequence tags (ESTs). ESTs are widely used as the probes for microarrays. Designing custom microarrays, rich in genes relevant to the experimental objectives, requires selection of probes based on their sequence. We have designed a probe selection method, called GeneSieve, to select EST probes for custom microarrays. To assign annotations to the ESTs, we cluster them into contigs using phrap. The larger contig sequences are then used for similarity search against known proteins in model organism such as *Arabidopsis thaliana*. We have designed three different methods to assign annotations to the contigs: bidirectional hits (BH), bidirectional best hits (BBH), and unidirectional best hits (UBH). We apply these methods to pine and potato EST sets. Results show that the UBH method assigns unambiguous annotations to a large fraction of contigs in an organism. Hence, we use UBH to assign annotations to ESTs in GeneSieve. To select a single EST from a contig, GeneSieve assigns a quality score to each EST based on its protein homology (PH), cross hybridization (CH), and relative length (RL). We use this quality score to rank ESTs according to seven different measures: length, $3'$ proximity, $5'$ proximity, protein homology, cross hybridization, relative length, and overall quality score. Results for pine and potato EST sets indicate that EST probes selected by quality score are relatively long and give better values for protein homology and cross hybridization. Results of the GeneSieve protocol are stored in a database and linked with sequence databases and known functional category schemes such as MIPS and GO. The database is made available via a web interface. A biologist is able to select large number of EST probes based on annotations or functional categories in a quick and easy way.

ACKNOWLEDGMENTS

I would like to thank my adviser Dr.Lenwood Heath for being a constant source of inspiration and encouragement. I thank him for being patient enough to listen to and solve the smallest and most trivial problems I came up with and for providing me all the resources I needed for my research. I would like to thank Dr.Ruth Grene for introducing me to the world of biology and for helping me understand the biological concepts essential for this research. This research would not have been possible without Dr.Heath and Dr.Grene's thoughtful guidance. I thank Dr.Naren Ramakrishnan for encouraging me for this research and helping me understand the concepts of database design. I would like to thank Dr.T. M. Murali for reviewing this document for accuracy.

I would like to thank my colleagues Amrita pati and Vibha Singhal for developing GeneSieve web interface; Cecilia Vasquez-Robinet, Jonathan Watkinson, and Shrinivas Rao Mane for evaluating GeneSieve web interface and providing valuable feedback; Allan Sioson and Douglas Slotta for listening to and answering all my stupid questions. I would also like to thank Deept Kumar, Harsha Rajasimha, and all the people in Torgerson 2160 for making my stay at Virginia Tech a mixed balance of work and fun.

Last but not the least, I extend my thanks and owe my gratitude to my family for their unconditional love and encouragement all through my life.

This thesis is dedicated to my parents, Sudha and Pradip Shukla.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 Motivation

In recent years, DNA microarrays have emerged as a powerful technique for the measurement of the expression levels of tens of thousands of genes simultaneously. In some circumstances, a microarray hybridization can yield a transcriptome-wide measurement of RNA levels in a given cell or tissue at a given point in time, or an average characterization of the response of a tissue to experimental manipulation. Information gleaned from these studies can generate working hypotheses for molecular pathways essential to a given biological process or potential drug targets for therapies.

The utility of the large volume of data generated, however, depends upon proper experimental design at many levels. Selection of a set of cDNA probes to be printed on a microarray is an important step that is often not given due consideration. Frequently, selection is based primarily on convenience and availability rather than on importance to the biological process under investigation. While the selection process is relatively simple for an organism with a small genome, such as yeast, it is more complex for an organism with a larger genome, such as mouse, human, or pine. It is costly to print probes for all the genes in a multicellular organism on a single microarray. Typically, a biologist is interested in studying only some specific tissues, biological processes, or biochemical pathways. A focused selection of clones that are relevant to the experimental objectives will be more economical. The selection process should ensure adequate coverage of genes of interest, sufficient sensitivity and specificity, unambiguous annotation, and reproducible and biologically meaningful

results. Also, increasing availability of microarray data creates a need to identify the same gene from different species, so that results can be compared across species.

This study is a part of the development of Expresso, a microarray experiment management system [3, 25, 58] that supports computationally all the stages of a microarray experiment, including microarray design; selecting cDNA probes; retrieving experimental results as scanned images; extracting meaningful information from these images; analyzing the results of an experiment as a whole; and providing the biologists with tools to explore and mine these results. One biological objective motivating Expresso is the functional genomics of stress response in loblolly pine *Pinus taeda*. To meet this objective, it is necessary to design a cDNA microarray rich in genes relevant to stress response in loblolly pine.

The goal of this research is to study factors affecting the probe selection for cDNA microarrays, devise a strategy that can be applied to select a subset of clones relevant to the experimental objectives, and design evaluation criteria that can be applied to measure the effectiveness of any selected clone set.

1.2 Summary of Results

Because of their availability, ESTs are widely used as probes for microarrays. To select EST probes for a custom microarray, it is important to assign annotation to ESTs. To assign reliable annotations to ESTs, we cluster them into contigs. Longer contig sequences are then used to search for similarity with known proteins in model organism. An EST is assigned the same annotation as its contig. We have developed three different methods to assign annotations to contigs: bidirectional hits (BH), bidirectional best hits (BBH), and unidirectional best hits (UBH). We implemented these methods on pine and potato EST datasets, using Arabidopsis as the model organism for comparison. Results show that the BH method assigns numerous proteins ($\sim$18 on average) to each contig, resulting in ambiguous annotations. The BBH method assigns only one protein to each contig and only one contig to each protein, but it covers only a small fraction of the contigs (18% for pine and 23% for potato). The UBH method assigns only one protein to each contig. Thus, it assigns an unambiguous annotation to each contig. It also covers a large fraction of the contigs (45% for pine and 80% for potato). For this reason, we use the UBH method to assign annotations to the contigs in GeneSieve.

Once a contig is selected based on its annotation, the next step is to select one of its component

ESTs to be printed on the microarray. Often, the longest EST or the EST closer to the $3'$ end of the contig is selected. This EST may not have the properties desirable in a microarray probe. We have devised a scoring system to evaluate the quality of an EST probe set. We assign a quality score (Q) to each EST based on its protein homology (PH), cross hybridization (CH), and relative length (RL). We assign quality scores to all of the pine and potato ESTs. We select a single EST from each contig by one of seven measures: maximum length, $3'$ proximity, $5'$ proximity, maximum PH, minimum CH, maximum RL, and maximum Q. we show that the longest ESTs show higher cross hybridization with other contigs and show less protein homology. Thus, the overall quality scores for such ESTs are relatively low. On the other hand, ESTs selected by $3'$ proximity show less cross hybridization, but show poor homology to known proteins. This is because $3'$ regions are less conserved than the protein coding regions. Also, our analysis reveal that only a small fraction of contigs align to the $3'$ end of the Arabidopsis proteins (11% for pine and 25% for potato). This suggests that most of the ESTs selected from the $3'$ end of the contigs actually come from the coding regions of the genes. ESTs selected by quality score give considerably better values of all three quality parameters. They show higher protein homology, less cross hybridization, and greater length. For this reason, we recommend that EST probes be selected based on their quality scores.

We have designed a web interface for quick and easy selection of EST probes for microarrays. We have linked the results obtained from the GeneSieve protocol to sequence databases and well known functional categorization schemes such as MIPS and GO. This allows one to select EST probes based on their annotation or functional categories.

1.3 Organization

The rest of this thesis is structured as follows. Chapter 2 gives biological background on cDNA microarrays. Chapter 3 defines the problem and scope of this research. Chapter 4 gives a brief overview of related work. Probe selection criteria are discussed in detail in Chapter 5. The probe selection strategy is presented in Chapter 6. Results obtained using our probe selection strategy are presented in Chapter 7. Chapter 8 gives details of GeneSieve – a web based tool that helps quick and easy selection of EST probes for cDNA microarrays. Chapter 9 concludes this research and provides direction for future research.

Chapter 2

cDNA Microarray Background

Section 2.1 introduces the biological concepts necessary to understand this study. Section 2.2 describes the construction of cDNA and EST libraries. Finally, Section 2.3 presents the principles of microarray technology.

2.1 Biological Background

All living organisms consist of cells, which contain many kinds of molecules, including nucleic acids and proteins. This section introduces nucleic acids, genes, and proteins. It also describes the process by which information is transferred from genes to proteins. Most of the material presented in this section can be found in greater detail in standard textbooks on biology such as Alberts, *et al.* [2], Cooper [14], and Griffiths, *et al.* [23].

2.1.1 Nucleic Acids: DNA and RNA

Nucleic acids are universal constituents of all living matter and are essential for storage, transmission, and transfer of genetic information. A nucleic acid is either a single- or double-stranded polynucleotide chain. A nucleic acid is either a *deoxyribonucleic acid* (DNA) or a *ribonucleic acid* (RNA).

DNA is a logically linear, spatially double-helical, structure composed of two intertwined chains of building blocks called nucleotides (see Figure 2.1). Each nucleotide consists of a phosphate group,

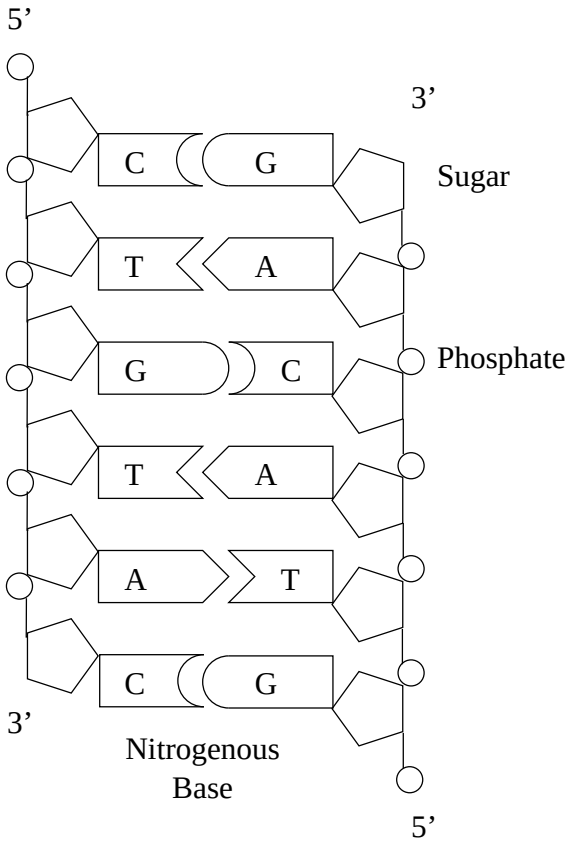

Figure 2.1: Structure of DNA

a deoxyribose sugar molecule, and one of the four different nitrogenous bases — adenine, guanine, cytosine, or thymine. Each of the four nucleotides is abbreviated to the first letter of the base that it contains: A, G, C, or T. The carbons in the deoxyribose sugar group are assigned numbers followed by a prime ($1', 2', 3', 4'$, and $5'$) to distinguish them from the numbering of the atoms in the bases. In DNA, nucleotides are connected to each other at the $3'$ and $5'$ positions; hence each chain is said to have polarity, with one end having a $5'$ phosphate group and the other having a $3'$ OH group. The polarities of two intertwined nucleotide chains are in opposite directions; hence the chains are said to be anti-parallel. Two nucleotide chains are held together by weak hydrogen bonds between complementary bases: A pairs with T, and G pairs with C.

Replication is a process by which a copy of a DNA molecule is made. During replication, two strands of the double helix unwind like a zipper (see Figure 2.2). The two exposed nucleotide chains

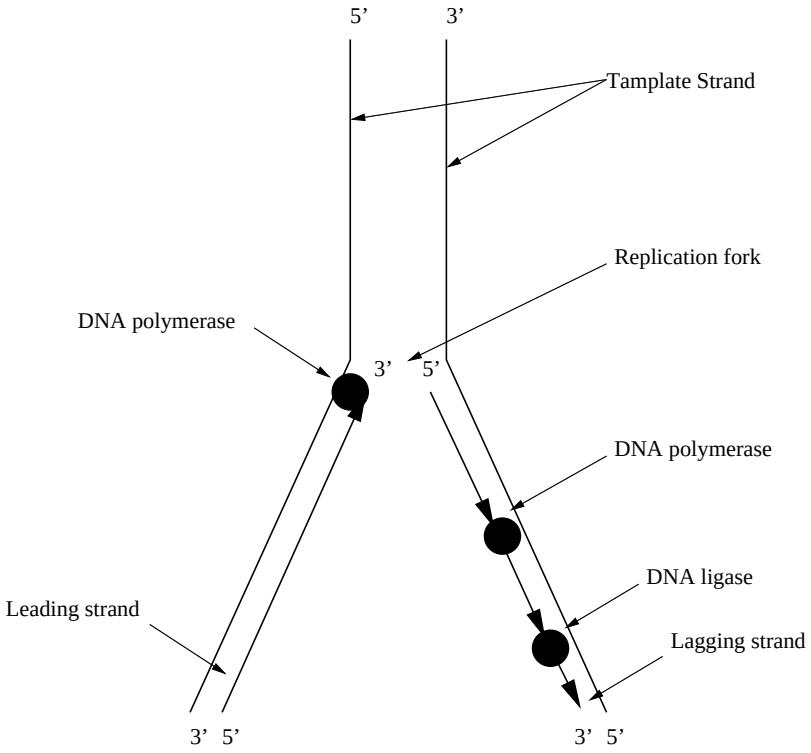

Figure 2.2: DNA replication

then act as *templates* for deposition of free nucleotides. Polymerization of free nucleotides into a new strand is catalyzed by the enzyme *DNA polymerase*. This enzyme initially binds to a double helical DNA at a specific nucleotide sequence called the *origin of replication* and then moves along the DNA, polymerizing new chains. Because of complementarity, the two *daughter DNA* molecules are identical to each other and to the original molecule. However, one strand of each daughter molecule is original, while the other is newly polymerized.

RNA is a single-stranded polynucleotide chain similar to DNA. Like DNA, RNA is composed of nucleotides, but these nucleotides contain the sugar ribose instead of deoxyribose. Furthermore, instead of thymine, RNA utilizes uracil (U), a base that has hydrogen bonding properties identical to those of thymine. Hence, the RNA bases are A, G, C, and U. There are three main types of RNA: messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). In this study, we are interested in mRNA (Section 2.1.3).

2.1.2 Genes and Proteins

Genes are fundamental units of heredity and carry information from one generation to the next. A gene is a functional region of a long DNA molecule. One end of the gene contains a regulatory region to which various proteins may bind, causing the gene to be transcribed at the right time and in the right amount. The other end of the gene contains sequence encoding the termination of transcription (see Section 2.1.3 for more on transcription). Each gene is responsible for coding of at least one specific protein or a part of a protein. In the genes of many eukaryotes, protein-coding sequences are interrupted by segments called *introns*. The split-up coding sequences between the introns are called *exons*.

Proteins are the main macromolecules of an organism. The primary structure of a protein is a linear chain of building blocks called *amino acids* linked together by peptide bonds. This primary chain is coiled, folded and, in some cases, associated with other chains, to form a functional protein. Proteins are important either as structural components, such as the proteins that constitute hair, skin, and muscle, or as active agents in the cellular processes, for example, enzymes and active-transport proteins.

2.1.3 Transcription and Translation

The most important concept in molecular biology is the way information stored in DNA is passed on and expressed. It is sometimes referred to as the flow of genetic information or the *central dogma* of molecular biology. The transfer of information from a gene to a protein is a two step process. These steps are: *transcription* and *translation*.

The first step taken by the cell to make a protein is to copy the information encoded in a gene into an mRNA molecule by a process called transcription. This RNA molecule represents a working copy of the gene and is called a *transcript*. The polymerization of ribonucleotides to form RNA is catalyzed by the enzyme *RNA polymerase*. This enzyme binds to a specific sequence, the transcriptional start site, near the $5'$ end of the gene. It separates two strands of DNA and then moves along the gene, maintaining the separated bubble. As it proceeds, it uses only one of the separated strands as a template, synthesizing a growing tail of polymerized ribonucleotides that eventually become the full length *primary transcript*. The addition of ribonucleotides by RNA polymerase is always at the $3'$ end of the growing chain. In eukaryotes, the introns are then cut out of the primary transcript. The

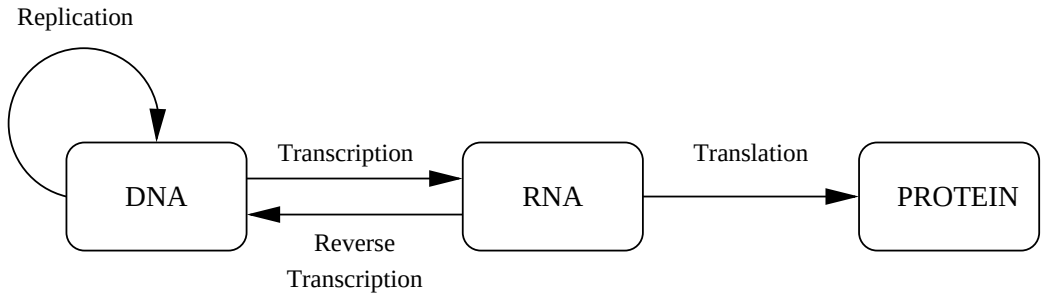

Figure 2.3: Transfer of genetic information

remaining RNA sequence is called messenger RNA (mRNA). The mRNA molecules exit the nucleus through nuclear pores and enter the cytoplasm.

Protein synthesis takes place on cytoplasmic organelles called *ribosomes*. The nucleotide sequence of the mRNA is read from the $5'$ end to the $3'$ end, in groups of three. These groups are called *codons*. A ribosome attaches to the $5'$ end of an mRNA molecule and moves along the mRNA, catalyzing the assembly of the string of amino acids that constitute the primary structure of the protein known as a polypeptide. Each amino acid is brought to the ribosome by a specific tRNA molecule that docks at a codon of the mRNA. Docking is by base pairing between a three-nucleotide tRNA segment, called an *anti-codon*, and the codon. This process of protein synthesis is called *translation*. Figure 2.3 shows the transfer of genetic information from genes to proteins.

2.2 cDNA and EST Libraries

Complementary DNA (cDNA) is synthetic DNA made from mRNA with the use of a special enzyme called *reverse transcriptase*. With the use of an mRNA as a template, reverse transcriptase synthesizes a single-stranded DNA molecule. This process is called *reverse transcription*. This single-stranded DNA molecule can then be used as a template for double-stranded DNA synthesis. Because it is made from mRNA, cDNA is devoid of both upstream and downstream regulatory sequences and introns. Therefore, cDNA from eukaryotes can be translated into functional proteins in bacteria.

Clone refers to a section of DNA or cDNA that has been inserted into a vector molecule, such as a plasmid or a phage chromosome, and then replicated to form many copies. A collection of such cDNA clones is known as a *cDNA library*. To prepare a cDNA library, total mRNA is extracted from

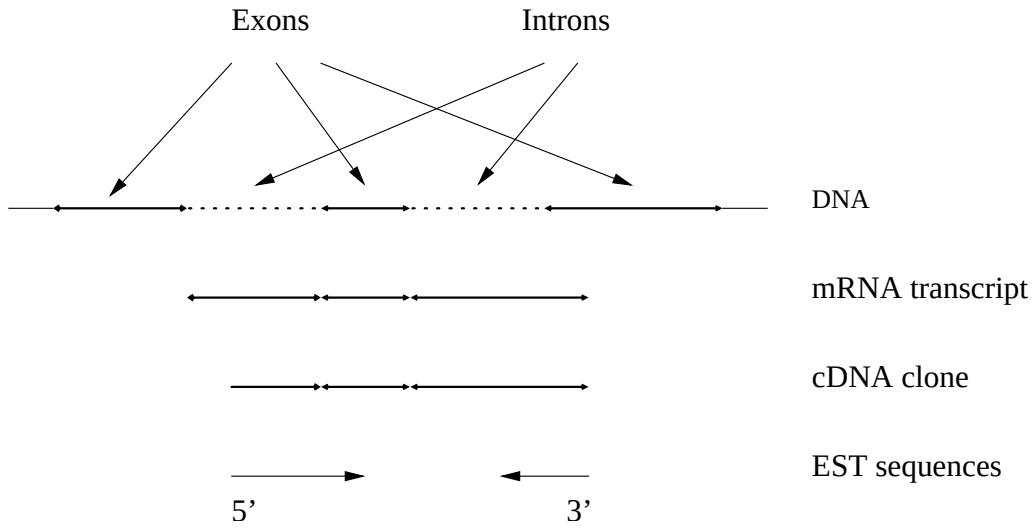

Figure 2.4: Relationship between RNA, cDNA, and ESTs

a particular tissue, and DNA copies (cDNA) of the mRNA molecules are produced by the enzyme reverse transcriptase. A short oligonucleotide, complementary to the poly(A) tail at the $3'$ end of the mRNA, is first hybridized to the mRNA to act as a primer for the reverse transcriptase, which then copies the mRNA into a cDNA chain, thereby forming a DNA/RNA hybrid helix. Treating the DNA/RNA hybrid with alkali selectively degrades the RNA strand into individual nucleotides. The remaining single-stranded cDNA is then copied into double-stranded cDNA by the enzyme DNA polymerase. If a gene is transcribed abundantly in the cells from which a cDNA library was made, it will be represented often in the cDNA library producing redundant clones. For this reason, *normalization* procedures are used to reduce the frequent representation of highly expressed genes, and thus enhance the probability of finding rarer mRNA transcripts [13, 32, 42, 44].

The most important advantage of cDNA clones is that they contain only coding sequences and only those genes that have been transcribed into mRNA in the tissue from which the RNA came. As the cells of different tissues produce distinct sets of mRNA molecules, a different cDNA library will be obtained for each type of tissue. cDNA libraries are also constructed to reflect the genes expressed by cells at different stages in their development.

To ensure that most of the information present in the cDNA library has been extracted requires sequencing a sufficiently large number of clones. Single-pass unverified reads are normally obtained from the $3'$ or $5'$ end of randomly selected cDNA clones. Sequences produced by this strategy are

termed *expressed sequence tags* (ESTs). ESTs are short (typically 400–600 bases) and relatively inaccurate [49, 57]. Typically, each EST sequence represents only a small part of a cDNA clone. Figure 2.4 illustrates the relationship between RNA, cDNA and ESTs.

This approach, known as EST sequencing, has been enormously successful in the framework of many genome projects. Single-pass sequencing is an important aspect of making the approach cost effective. In most cases, no initial attempt is made to identify or characterize the cDNA clones [49]. A clone is annotated by comparing its EST sequence to the sequences of known genes. It is fully expected that many clones will be redundant and that a smaller number will represent various sorts of contaminants or cloning artifacts. There is little point in incurring the expense of high-quality sequencing until later in the process, when clones are validated and a non-redundant set is selected. Despite their fragmentary and imprecise nature, ESTs are an invaluable resource for the discovery of new genes whose functions can be tentatively deduced from their sequence and experimentally verified.

More details on construction of cDNA libraries can be found in textbooks biology such as Cooper [14] and Griffiths, *et al.* [23]. Kohchi, Fujishige, and Ohyama [32] and Patanjali, Parimoo, and Weissman [44] discuss construction of normalized cDNA libraries. Adams, *et al.* [1] and Jongeneel [30] provide information on EST sequencing and EST databases.

2.3 cDNA Microarrays

The cDNA microarray is a hybridization-based experimental technique that allows one to study expression levels of tens of thousands of genes simultaneously. The principle behind the microarray technique is that two complementary strands of nucleic acid can hybridize. This method provides a high degree of accuracy of detection as a consequence of exquisite, mutual selectivity between complementary strands of nucleic acids.

The process of a microarray experiment starts with a biological hypothesis and selection of a set of genes of interest. DNA templates for these genes are then obtained and amplified by the *polymerase chain reaction* (PCR). Following purification and quality control, clones are deposited on coated glass slides using a computer controlled robot. Tens of thousands of clones can be deposited on a single glass slide. These clones on the glass slide are known as *probes*. To determine the genes that are differentially expressed when cells are exposed to experimental conditions, such as drought,

stress, or toxic chemicals, total mRNA is extracted from both test and reference cells and reverse transcribed to form cDNA molecules. cDNA molecules are then labeled with either Cye3 or Cye5 fluorescent dye. These fluorescently labeled cDNA clones are known as *targets*. Targets are pooled and allowed to hybridize to the probes on the glass slide. Laser excitation of the incorporated targets yields an emission with a characteristic spectra, which is measured using confocal scanners. The result is two images, one for each dye, in which each pixel is a measured intensity value. The two images are analyzed using image processing software. Information such as clone identifier, intensity values and intensity ratios, is attached to each target. Ratios of Cye3 and Cye5 signal intensities are analyzed for significant deviations from 1 (no change) which indicate either increased (> 1) or decreased (< 1) levels of gene expression relative to the reference sample.

The reader can refer to [9, 16, 24, 29, 35, 48] for more information on cDNA microarrays and their applications.

Chapter 3

Problem Definition

An important step in microarray design is the selection of a set of cDNA clones to amplify and print on the microarray. Selection is strongly influenced by the availability of the cDNA clones. An alternative is to use prefabricated arrays, either commercially available in the market or available from collaborators. Most commercially available off-the-shelf arrays contain a series of well characterized genes expressed across many cell types. The primary advantages of commercial arrays are their wide availability and general applicability. The efficiency, sensitivity, and reproducibility of the commercially available arrays are good and improving.

There are three main disadvantages to using prefabricated arrays. First, a group interested in studying some specific tissue or process, may find that a significant number of the genes being assayed are not relevant to the experiment. Time, effort, and expense is consumed in analyzing these irrelevant genes. Second, until whole genome arrays become available, it is quite likely that the given array used in the experiment will be lacking genes that may be important for the tissue or process being studied. This results in an incomplete picture of the process being studied. Third, the number of species for which prefabricated arrays are available is small, and their cost is often prohibitive. Many companies offer custom services for producing arrays, but the process is still expensive and cumbersome and does not lend itself well to the commonly changing needs of the researcher.

For these reasons, it is useful to design custom microarrays containing only those genes relevant to the tissue or biological process being studied. For example, a researcher interested in liver will be better served by an array representing only those genes expressed in liver cells, and a researcher

interested in metabolic pathways will find more value in an array rich in the genes involved in metabolic pathways. In the Expresso project [3, 25, 58], plant biologists are interested in studying genes that are responsible for resistance to stress in loblolly pine trees. The goal is to identify genes whose expressions are influenced by the application of stress. These genes serve as candidates for further experimentation and to direct the construction of more specific hypothesis.

A microarray experiment starts with some biological hypothesis. A list of genes relevant to this hypothesis can be selected by reviewing literature and results from previous biological experiments on the same or other related species. Often a biologist is interested in studying all the genes belonging to a certain functional category or biochemical pathway. Then the need is to identify a set of cDNA clones or ESTs representing those genes in the species being studied.

The ESTs representing a gene of interest can be identified by a sequence similarity search. There can be a large number of ESTs showing similarity to the gene of interest. ESTs from other closely related genes also show high sequence similarity. Then the problem is to select an EST which best represents the gene of interest. The factors that affect the performance of a microarrays are as follows:

- *Specificity*: Kane, *et al.* [31] define specificity as the sequence similarity between hybridizing probe and target sequences. It can also be viewed as the ability of a probe to distinguish the target from other similar sequences.

- *Cross Hybridization*: Xu, *et al.* [71] define cross hybridization as the binding of the mRNA of one gene to the probe of another gene. Cross hybridization results from the regions of high sequence similarity found among repetitive elements or between the members of a multi-gene family. Cross hybridization contributes to the overall signal intensity and leads to misinterpretation of the expression data. To increase reliability of microarray results, cross hybridization needs to be minimized.

- *Coverage*: Coverage is the fraction of all genes relevant to the experiment included on the microarray. High coverage is desirable so that a complete picture of the process being studies is obtained.

- *Redundancy*: According to Tomiuk and Hafmann [64], redundancy is the selection of more than one clone to represent the same gene. Redundancy reduces the number of genes that can be represented on a microarray.

The results obtained from microarray experiments are variable. Some of the factors contributing to this variability include: low sequence fidelity, inconsistency in spotting and hybridization, low specificity of the EST probes, cross hybridization, inaccurate signal intensities obtained from image processing, and discrepancies in fold-change calculations by different statistical methods. Because of these factors, it is important to independently confirm microarray results with non-array based methods such as northern blots, in situ hybridization, and RT-PCR. Making biological inferences from microarray data requires correct and unambiguous annotation of the EST probes. Typically, an EST represent only a small part of a gene. For this reason, it is often difficult to assign reliable annotation to ESTs based on homology to the known proteins. Annotation helps in organizing ESTs in functional categories and comparing results with literature databases.

Increased use of microarrays makes it possible to compare microarray results across different species. In many situations, results obtained in one species can be verified or studied in greater detail in another species. For example, genes identified by a microarray study on human cancer can be used to form a hypothesis of the disease model. This model can be verified by studying those genes in mouse or rat, as mouse and rat can be subjected to physiological or genetic manipulations. In some situations, genes identified by experiments on mouse and rat can be investigated in humans. Such a study requires identification of equivalent probes representing the same genes in different species.

A selection process can be devised to select a subset of clones from a much larger initial set and from a variety of sources. The selection process should strive for high coverage of relevant genes, sufficient sensitivity and specificity of the array, reproducibility of the results to ensure statistical significance, and correct annotation to provide unambiguous link to the corresponding entries in gene and literature database.

The goal of this research is to study factors affecting the probe selection process, to design a probe selection strategy and to implement the strategy in an efficient computational framework. The strategy includes a set of evaluation criteria that can be used to measure the quality of any selected probe set. Considering the diverse nature of animal and plant species, both biologically and the amount of information available, the scope of this research is restricted to the plant species *Arabidopsis thaliana*, *Pinus taeda* (pine), and *Solanum tuberosum* (potato). *Arabidopsis thaliana* is chosen as the model organism while *Pinus taeda* and *Solanum tuberosum* are selected for cross-species comparison.

Chapter 4

Related Work

This chapter discusses prior work on clone selection, custom microarray design, cross hybridization between closely related genes, and finding equivalent clones in different species.

4.1 Biological Considerations

Tomiuk and Hofmann [64] discuss various probe selection criteria in detail. Physical properties of a probe, such as its length, can affect its hybridization properties. Quality of an array depends on the reliability of the cDNA or EST library used. Whenever possible, probes should be selected from normalized cDNA libraries to avoid frequent representation of highly expressed genes. Another method to control redundancy is clustering of EST data. In clustering, ESTs from the same gene are assigned to the same cluster based on their sequence similarity. Gene region ($3'$ untranslated region, $5'$ untranslated region, or coding region) from which a probe is selected can greatly affect specificity and cross hybridization. Coding regions are more conserved and show high degree of similarity with other closely related genes. Hence, probes selected from coding region are the most susceptible to cross-hybridization events. The choice of selection strategy depends on the application and the problem to be solved. In general, high-throughput arrays containing all the genes in an organism do not require an intensive scrutiny of the spotted cDNA probes. Targeted arrays, on the other hand, do not aim at identification of new genes but allow monitoring of complex expression patterns. Targeted arrays require careful annotation and quality control of cDNA probes. Often, a probe selection strategy is a compromise between experimental objectives and practicability.

4.2 Custom Arrays

Many research groups build custom microarrays to accommodate specific requirements of their focused research. Loftus, *et al.* [36] discuss selection of a human neural crest-melanocyte (NC-M) cDNA set for microarray analysis. Using 21 NC-M expressed genes and NCBI dbEST, they identify one library (library 198) rich in NC-M ESTs. The dbEST is a division of GenBank that contains sequence data and other information on single pass cDNA sequences, or ESTs, from a number of organisms. After clustering 22,889 ESTs in library 198, they select 852 ESTs that have a library distribution profile similar to that of the 21 neural crest-expressed genes.

Barrett, *et al.* [8] design a similar neural cDNA microarray containing 1,152 human cDNAs. These cDNAs represent all the major cellular types of the brain including neurons, astrocytes, microglia, and oligodendrocytes. Rockett, *et al.* [53] develop a 950-gene cDNA array for examining gene expression patterns in mouse testis. They assemble two lists of genes, one of 950 mouse genes and the other of 960 human genes, expressed in either mouse testis or human testis or both. All of these genes are available as sequence-verified clones from public sources. 764 of these genes are homologs in human and mice, making it possible to compare gene expression between mouse models and human clinical testicular samples. Carlisle, *et al.* [11] design a microarray comprising 5,184 cDNA clones for prostate gene expression studies in human. A method similar to the NCBI UniGene [49] is used to cluster sequences derived from prostate cDNA libraries sequenced within the context of the Cancer Genome Anatomy Project. A single cDNA is selected to represent each cluster and placed on the array. Takemasa, *et al.* [61] construct "Colonochip", a cDNA microarray specialized for human colorectal carcinoma. It contains 4,608 non-redundant cDNA clones from over 30,000 cDNA clones derived from three types of human cDNA libraries, as well as clones from 170 genes suspected to be involved in colorectal carcinogenesis.

Pennie [47] discusses the construction of "ToxBlot II", a custom microarray containing cDNAs representing 12,564 human genes chosen on the basis of their potential relevance to a broad range of toxicities. ToxBlot II allows the simultaneous expression profiling of genes representing cellular pathways, facilitating a detailed investigation of potential mechanisms of toxicity. Such microarrays can contribute significantly to our understanding of basic toxicology mechanisms, and may in the future contribute to the faster development of drugs and agrochemicals products. Lorenz, *et al.* [37] construct "ImmunoChip", a microarray for mouse immunology research. They select immunologically relevant clusters based on the expression of ESTs in the immunological organs and cells. They

categorize these clusters into one of the three modules: gene module (GM), homologous gene module (HM), or EST module (EM). Based on the presence of a cluster in a specific module, they rank each constituent EST based on its alignment and BLAST score with the reference gene (GM), length (HM), or its presence in the immunological libraries (EM). For each cluster, they choose the clone with the highest rank as the best representative for that cluster.

4.3 EST Annotation

NCBI UniGene [49, 68] is a system for automatically partitioning a set of GenBank sequences into a non-redundant set of gene-oriented clusters. UniGene takes ESTs of an organism from GenBank and creates clusters of sequences based on their similarity. This is done by converting similarity scores between sequences to boolean links between them. Two sequences are considered linked if their sequence similarity exceeds a threshold. To reduce the frequency of multiple clusters being identified for a single gene, any cluster that does not contain at least one sequence with a polyadenylation signal or two labeled $3'$ ESTs is discarded. In eukaryotes, the end of the transcription process is marked by cleavage of the primary transcript at the polyadenylation signal followed by the addition of a poly-A tail. Thus, each UniGene cluster is likely to contain sequences that represent a unique gene in that organism. Each cluster is linked to related information, such as a list of accession numbers of ESTs and known mRNAs or gene transcripts belonging to the cluster, tissue types in which the gene is expressed, and location of the gene on a chromosome map. Each sequence in the UniGene is compared to the database of known proteins in the model organisms. A protein is assigned to a cluster if their sequence similarity exceeds a threshold. Figure4.1 shows a sample output for one of the UniGene clusters for Arabidopsis:

The UniGene collection has been used as a source of unique sequences for the fabrication of microarrays for large-scale gene expression studies [17]. While the UniGene collection is very useful, it has four disadvantages. First, UniGene is likely to put all splice variants of a gene in the same cluster [49]. Hence, it does not help in finding probes to distinguish among different splice variants of the same gene. Second, UniGene does not produce contigs or consensus sequences for EST clusters. A *contig* is defined as a continuous sequence of DNA that has been assembled from overlapping DNA fragments. Also, UniGene does not give an alignment of the constituent ESTs within a cluster. Third, out of many ESTs constituting a cluster, UniGene does not indicate which one is

```
UniGene Cluster At.47584 Arabidopsis thaliana
L-ascorbate peroxidase
LINKS: HomoloGene
SELECTED MODEL ORGANISM PROTEIN SIMILARITIES

organism, protein and percent identity and length of aligned region
A.thaliana: pir:S20866 - S20866 L-ascorbate peroxidase (EC 1.11.1.11) precursor -
Arabidopsis thaliana (fragment) 100.00 % / 263 aa (see ProtEST )
E. coli: ref:NP_052682.1 - Catalase-peroxidase [Escherichia coli] 24.61 % / 210 aa
(see ProtEST )
S. cerevisiae: pdb:1JDR - A Chain A, Crystal Structure Of A Proximal Domain Potassium
Binding Variant Of Cytochrome C Peroxida 39.92 % / 237 aa (see ProtEST )

EXPRESSION INFORMATION
cDNA sources: roots ; Leaf ; aboveground organs ; seedlings leaf and root ; whole plant ;
green siliques ; seed ; inflorescence ; root ; seedling hypocotyl ; flower buds

SEQUENCE INFORMATION

mRNA SEQUENCES (7)
X59600.1 A.thaliana mRNA for ascorbate peroxidase PA
AY039879.1 Arabidopsis thaliana At1g07890/F24B9_2 mRNA, complete cds P
AY056395.1 Arabidopsis thaliana At1g07890/F24B9_2 mRNA, complete cds P
NM_100663.2 Arabidopsis thaliana ascorbate peroxidase, putative (APX) (At1g07890) mRNA,
complete cds P
AY094002.1 Arabidopsis thaliana At1g07890/F24B9_2 mRNA, complete cds P
AY086425.1 Arabidopsis thaliana clone 25057 mRNA, complete sequence P
NM_179276.1 Arabidopsis thaliana ascorbate peroxidase, putative (APX) (At1g07890) mRNA,
complete cds P

EST SEQUENCES (10 of 145)[Show all ESTs]
  BE662951.1 cDNA clone AtAOzE28 Leaf 1.0 kb P
  CF652474.1 cDNA clone MPIZp2001A011Q root 5' read 0.9 kb
  CF651760.1 cDNA clone MPIZp2001A011Q root 5' read 0.9 kb
  CB257555.1 cDNA clone MPIZp771F153Q whole plant 5' read 0.7 kb
  CB264317.1 cDNA clone MPIZp2000O173Q inflorescence 5' read 0.7 kb
  CB257556.1 cDNA clone MPIZp771F163Q whole plant 5' read 0.6 kb
  CB259445.1 cDNA clone MPIZp770P119Q whole plant 5' read 0.6 kb
  BE662835.1 cDNA clone AtCOzJE1 Leaf 0.6 kb P
  BE662830.1 cDNA clone AtCOzJB8 Leaf 0.6 kb P
  CB261482.1 cDNA clone MPIZp769A163Q whole plant 5' read 0.6 kb P

Key to Symbols

P Has similarity to known Proteins (after translation)
A Contains a poly-Adenylation signal
M Clone is putatively CDS-complete by MGC criteria
```

Figure 4.1: Example of an NCBI UniGene cluster

the best candidate as a microarray probe. It does identify the longest EST in each cluster, which certainly could be selected as a microarray probe. Fourth, proteins are assigned to a cluster based on sequence similarity between sequences constituting a cluster and known proteins in the model organisms. Typically, any EST sequence in a cluster represents only a small part of a gene. This might result in an inaccurate assignment of a protein to a cluster. Use of a consensus sequence, which is usually larger than the constituent EST sequences, for protein similarity search can assign more reliable annotations to clusters.

The TIGR Gene Indices [50] are a collection of species- specific databases that uses a highly refined protocol to analyze EST sequences and identify genes represented by them. For each species, it first obtains EST sequences from dbEST. It trims these EST sequences to remove vector, polyA/T tails, and contaminating bacterial sequences. It also obtains gene sequences (NP sequences) from GenBank and expressed transcripts (ET) sequences from the TIGR EGAD database. Then, it compares EST and gene sequences using FLAST, a rapid sequence comparison program, in which query sequences are first concatenated and then searched against a nucleotide database. Sequences sharing a minimum of 95% identity over a 40 nt or longer region with less than 20 bases of mismatched sequence at either end are grouped into clusters. For each cluster, component EST, NP, and ET sequences are obtained and these sequences are then assembled using CAP3 [28] to produce TCs. Each cluster is assembled separately. A TC containing a known gene is assigned the function of that gene; TCs without assigned functions are searched against a non-redundant protein database. High scoring hits are assigned a putative function.

The Sputnik database at MIPS [55] is a similar system for automatically clustering and annotating large ESTs datasets. It assembles ESTs into contigs using the HarvESTer software and then uses consensus sequences for assigning annotations based on homology to the known proteins. Sputnik also links ESTs to the MIPS functional categories. Software tools, such as ESTAnnotator [27] and PipeOnline [7], are available for high throughput EST annotation. GOblet [26] is a software package for automated Gene Ontology annotation for anonymous cDNA or protein sequences.

4.4 Cross-Species Comparison

Wang, *et al.* [65] discuss ProbeMatchDB — a web based database to facilitate search for ESTs that can be used to represent the same gene across different microarray platforms and species. It

integrates the UniGene and the HomoloGene databases of NCBI [68] as well as probe information provided by Affymetrix, Research Genetics, and Operon. It can be used to find equivalent EST clones in the Research Genetics sequence verified clone set based on results from Affymetrix GeneChips. The accession numbers of the oligo probes in the GeneChips are available, which can be used to identify UniGene clusters represented by these probes. A list of ESTs constituting any cluster is obtained and then these ESTs are searched for in the Research Genetics database. ProbeMatchDB can also be used to identify probes representing homologs across human, mouse, and rat on different microarray platforms. This is done by using HomoloGene database of NCBI, which identifies homologs in several organisms by sequence comparison between all UniGene clusters for each pair of organisms. ProbeMatchDB essentially uses the UniGene database for all its searches, and hence, inherits the disadvantages of the UniGene database, as discussed in Section 4.3.

4.5 Cross Hybridization

If there is high sequence similarity between two closely related genes, then the mRNA of one gene may sometimes hybridize to the probe of another gene. This phenomenon is known as cross hybridization. Xu, *et al.* [71] analyze gene expression in the cytochrome P540 gene super-family of *Arabidopsis thaliana*. P450 genes are classified according to the degree of amino acid sequence identity, with P450s of the same family defined as having greater than 40% identity, and P450s of the same subfamily having greater than 55% identity. They design experiments to evaluate the specificity of P540 microarrays. Results show that sequences with less than 80% identity with probe display less than 20% cross hybridization. Sequences with greater than 80% identity with probes show higher cross hybridization. Averts, *et al.* [18] describe hybridization experiments with a model array representing four distinct gene families: chemokinases, cytochrome P450s, G proteins, and proteases. The cDNA clones selected for this array exhibit pairwise sequence identities ranging from 55% to 100%. Results reveal that sequences containing less than 80% sequence identity to the probe sequences show cross hybridization ranging from 0.6% to 12%. The mRNA sequences containing greater than 80% sequence identity to the probes show higher cross hybridization. Wren, *et al.* [69] discuss observations of cross hybridization on microarrays made by others and argue that overall similarity across large regions is not as much of a predictor of cross hybridization effects as the existence of small, highly similar regions, such as repetitive elements.

4.6 PCR Primer Design

Due to the problem of potential cross hybridization, using full-length genes for microarray construction is not appropriate in some situations. To avoid cross hybridization, researchers sometimes do not use full-length genes but rather use gene-specific fragments as probes on a microarray. For this purpose, it is necessary to identify a fragment of a gene that does not have high sequence similarity to any other sequence in a given organism and then to design forward and reverse primers based on the selected gene-specific fragment to allow amplification by PCR.

Several software products are available to design PCR primers for amplifying microarray probes. Nielsen and Knudsen [43] describe PROBEWIZ, an automated approach to design PCR primers for amplifying probes for cDNA microarrays. PROBEWIZ designs PCR primers for amplifying probe sequences that have minimal sequence homology with the other expressed sequences from a given organism, and at the same time places the probe as near to the $3'$ end of the sequence as possible. The primer selection is based on user-defined penalties for homology, primer quality, and proximity to the $3'$ end. Xu, *et al.* [70] develop PRIMEGENS, a bioinformatics tool for the automatic design of PCR primers using DNA fragments that are specific to individual open reading frames (ORFs). An ORF is a section of a sequenced piece of DNA that begins with a start codon and ends with a stop codon. It is presumed to be the protein coding sequence of a gene. PRIMEGENS first carries out a BLAST search for each target ORF against all other ORFs of the genome to identify possible homologous sequences. Then it performs optimal sequence alignment between the target ORF and each of its homologous ORFs using dynamic programming. PRIMEGENS uses these sequence alignments to select gene specific fragments, and then feeds these fragments to the primer3 program to design primer pairs for PCR amplification.

Talla, *et al.* [62] design a *Saccharomyces serevisiae* microarray with an aim of reducing cross hybridization between related sequences. They design probes of similar lengths, preferably located near the $3'$ end of the open reading frames. They compare the sequence of each gene against the entire yeast genome to identify non cross hybridizing regions. They submit these non cross-hybridizing regions to the primer3 program. They compare each candidate primer sequence to the 16 yeast chromosome sequences using BLASTN program. Primers are considered unique if no match is found. They select the best primer pair for each ORF based on the following criteria: an optimal probe length of 500 base pairs, at least one of the two primers is unique, and the position of the reverse primer is as close as possible to the $3'$ end of the ORF. They are able to design primers for

more than 97% of yeast genes using this approach.

4.7 Oligonucleotide Probe Design

High density synthetic oligonucleotide microarrays are widely used in biomedical research. These microarrays are made from either short (20-25 mers) or long (50-70 mers) oligonucleotide probes. Kane, *et al.* [31] demonstrate that oligonucleotide microarrays compare well with cDNA microarrays and that a single oligonucleotide probe per gene is sufficient to monitor gene expression. The oligonucleotide approach allows researchers to design a probe specific to each gene to avoid regions that are repetitive or very similar to other known genes.

Several softwares are available for the selection of oligonucleotide probes for microarrays. Rouillard, Herbert, and Zuker [54] develop OligoArray, a program to design gene specific and secondary structure free oligonucleotide probes for genome-scale oligonucleotide microarrays. For each sequence in a given organism, OligoArray reads from the $3'$ end using a moving window of length equal to the length of the oligonucleotide. It checks compares each oligo sequence to all other sequences in a given organism using BLAST. Sequences that pass the specificity criteria are examined for the presence of strong secondary structure that could interfere with hybridization. Oligo sequences that are free of secondary structures are then selected as microarray probes. Li and Stormo [34] design ProbeSelect, an algorithm for the selection of short or long oligo probes for each gene in the entire genome based on sequence information and hybridization free energy. Wang and Seed [66] present a strategy for selecting oligonucleotide probes for protein coding sequences. They compare each candidate probe with all other sequences in a given organism and reject it if it contains 15 bases of perfect identity with any other sequence. Next, they remove sequences with low complexity and predicted poor probe accessibility. Oligos free of secondary structure are used as microarray probes.

Chapter 5

Selection Criteria

When monitoring expression levels of a large number of genes, sufficient sensitivity and specificity of an array, as well as broad coverage of relevant genes, are of crucial importance. In addition, the quality of an array should guarantee reproducible results to ensure their statistical significance. This chapter describes some of the criteria to be considered while selecting probes for cDNA microarrays.

5.1 General Considerations

Physical properties of probes, such as probe length, can influence hybridization kinetics and sensitivity of a microarray [59, 60]. Longer cDNA probes have more reliable hybridization properties but increased viscosity can complicate the array manufacturing process. In addition, increasing the probe length raises the danger of non-specific cross-hybridization events. If probes of substantially different lengths are used, they may exhibit different hybridization kinetics, making it difficult to compare results across different genes on the same array.

The choice of a probe selection strategy depends on the objective of a biological experiment. Microarrays can be designed to be either broad coverage 'discovery' style arrays, which include an unbiased selection of gene sequences (which may include probes for genes of uncharacterized functions), or 'hypothesis driven' where the arrays are designed to focus on genes relevant to a particular biological problem. In situations, where little prior information is available, or where the prime motivation is an unbiased overview of global changes in gene expression patterns, high density arrays are appropriate choices. Typically, probes are selected from a pre-existing collection of cDNA

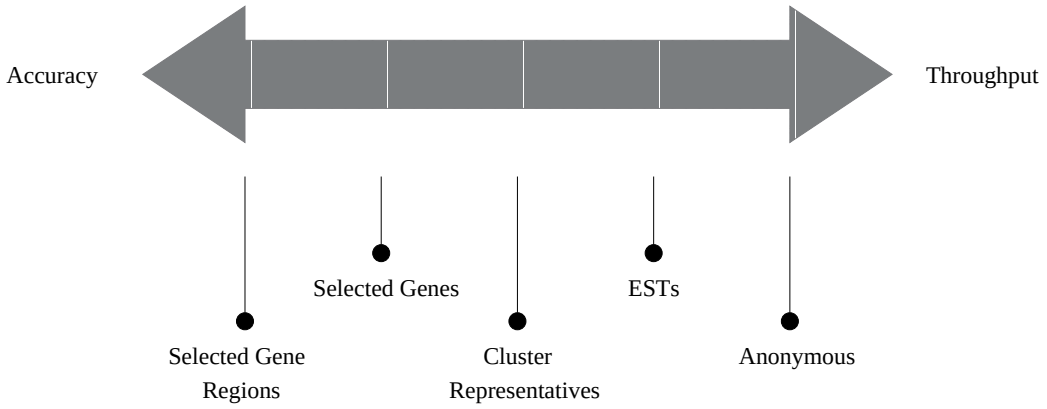

Figure 5.1: Probe selection approaches

clones. A disadvantage of this approach is the lack of reliable clone annotations, shifting work to the post-hybridization phase. On the other hand, small but specialized arrays are designed with a focus on defined biological problems such as genes relevant to a particular metabolic pathway or a particular tissue type. The limited number of probes on these arrays allow a more thorough selection and annotation protocol. Figure 5.1 illustrates a range of probe selection approaches between high-density and high-accuracy arrays. These probe selection approaches are discussed in the following sections.

5.2 Sequence Libraries

The easiest and the least expensive option is to use clones from a cDNA library without prior sequencing. Only those clones that show differential expression after hybridization are subjected to sequencing and further analysis. This strategy is useful for high density arrays, since only a small fraction of presumably interesting genes must be annotated. Typical applications include high-throughput arrays for potential new drug targets or analysis of biological systems without any available sequence information. Often, highly expressed genes are represented more frequently in the cDNA libraries. Normalization procedures can be used to reduce the frequent representation of such highly expressed genes [44, 13].

A more reliable choice is to use sequenced cDNA clones as microarray probes. NCBI's dbEST [10] provides sequence data and other information on single pass cDNA sequences, or ESTs, from a

number of organisms. The IMAGE consortium [33] and several other distributors provide access to physical clones. ESTs are a valuable and widely used source for microarray probes. One common problem when dealing with ESTs is their sometimes poor reliability. Three different types of errors are often observed: a sequence in a database is different from the actual clone; a sequence is correct, but corresponding gene annotation is wrong; and the predicted $3'$ or $5'$ orientation is wrong [64].

Whenever prior information on relevant genes is available, a pre-selection can be made by first assembling a list of genes of interest, and then selecting EST clones based on this list. The ESTs obtained can be used directly for the production of a microarray. A more reliable, but also more costly and time consuming strategy is to amplify the suitable region of an EST by polymerase chain reaction (PCR). This can help in controlling important properties of a probe such as its length, its $3'$ or $5'$ orientation, and its position within the gene.

5.3 Clustering

When using cDNA or EST clones as microarray probes, an obvious problem is redundancy. Highly expressed genes are often represented by multiple clones. As a result, several selected probes might correspond to a single gene, and hence, do not enhance the coverage of the array. The standard method for controlling redundancy in microarray probe selection is clustering of EST data. In the clustering process, two cDNA clones belonging to the single gene are recognized by their overlapping sequences and assigned to the same cluster. For most of the large scale EST sequencing projects, there are also publicly available EST cluster databases. The most notable in this respect is the NCBI UniGene database (Section 4.3), which provides non-redundant gene-oriented clusters for the GenBank sequences. UniGene clusters are the basis for several sets of cDNA array probes [17]. TIGR [50, 51] and SANBI [39] also provides similar databases of clustered EST sequences. Several software packages are available for sequence clustering, such as phrap [22] and TIGR Assembler [19]. These tools allow clustering of any EST sequence collection that is not addressed by the databases mentioned above.

Sequence clustering is not a perfect process, and there are two reasons for failure. Tomiuk and Hafmann [64] discuss these modes of failures. The first reason for failure is *under-clustering*, when all the ESTs corresponding to the same gene are not placed in a single cluster (see Figure 5.2). This results in more clusters than the number of actual genes. Often, two clusters per gene are observed,

one being formed by 5′ ESTs, and the other consisting of 3′ ESTs. For large genes, the number of clusters per gene can be greater than two. Under-clustering is frequent and hard to avoid. As a consequence, multiple probes per gene may be selected, resulting in unwanted redundancy.

The second reason for failure is *over-clustering*. Here, ESTs that do not correspond to the same gene are found erroneously combined within a single cluster. The major cause of this type of error is the presence of *chimeric clones*, containing sequence coming from more than one genes. Figure 5.3 illustrates over-clustering. There are multiple dangers for probe selection in an unnoticed over-clustering situation. First, only a single probe is selected for two or more genes. Second, depending on the fragment selection, a probe can represent either a single gene or a combination of more than one genes. The orientation of the cDNA may change at the chimeric boundary. Third, annotation of such a cluster will be erratic unless a very good annotation protocol is used. Despite the problems mentioned above, sequence clustering is the most useful method to avoid redundancy in microarray probe selection.

5.4 Genes and Gene Regions

One important issue in probe selection is: how many genes, and which ones, to represent on a microarray? A straight forward approach is to select DNA fragments corresponding to all the genes of a given species. This approach has been used for organisms with smaller genomes such as *Saccharomyces cerevisiae*. The genes of higher eukaryotes are more complex and subject to alternative splicing, resulting in multiple proteins per gene. Identification of all splice variants on the one hand, and selection of corresponding DNA fragments on the other hand, complicates generation of an 'all inclusive' microarray suitable for analysis of a complete transcriptome. Often, investigation of defined biological questions or subject areas is the objective of a microarray experiment. Examples include toxicology and tissue-specific arrays. Besides general selection criteria, experimental objectives either dictate the choice of relevant genes or, at least, offer guidelines for gene selection.

After selecting cDNAs corresponding to genes of interest, one has to determine whether the complete cDNA clone or a pre-selected part of it should be deposited on the array. The use of complete cDNA sequence bears the danger of cross hybridization to other closely related genes. A probe can be selected from the 3′ UTR, 5′ UTR, or the coding region of a gene (see Figure 5.4). The choice of a probe from the 3′-untranslated region of a gene reduces the probability of cross-

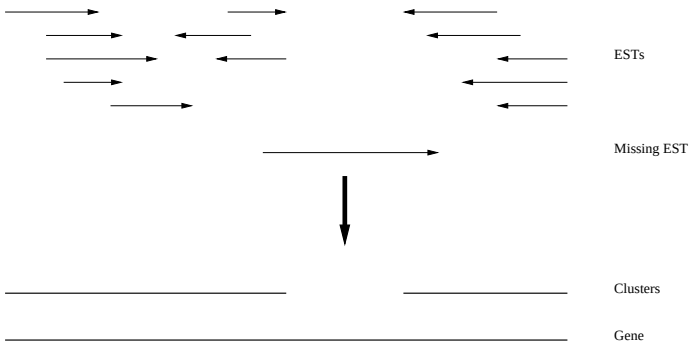

Figure 5.2: Under-clustering

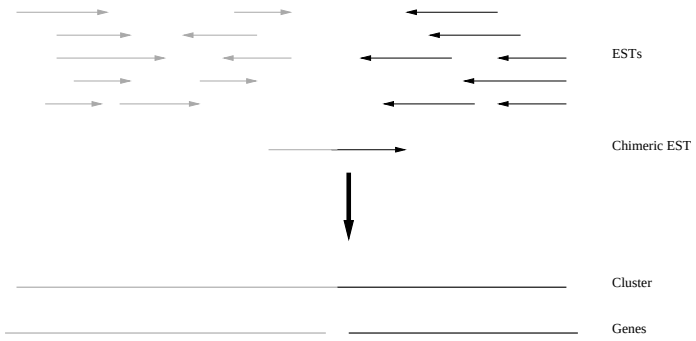

Figure 5.3: Over-clustering

hybridization because sequence divergence is typically greater in this region [66]. Additionally, this region is rarely affected by alternative splicing events. However, potential existence of alternative polyadenylation signals and elevated propensity for repetitive elements require a careful examination of probes coming from this region. The $3'$ regions of genes are not conserved. Hence, a probe selected from the $3'$ region of a gene may not show homology to a known gene, especially when the $3'$ region is very long. For this reason, it is often difficult to assign reliable annotation to these probes. Figure 5.5 illustrates the effect of alternative polyadenylation. Microarray probes localized in the $5'$-untranslated regions are closely linked to promoters. However, these regions are often missing in cDNA clones generated by reverse transcription. Moreover, a $5'$-untranslated region bears the danger of alternative promoter usage. Figure 5.6 illustrates the effect of alternative promoter usage. Selecting a probe from the coding region of a gene, i.e. the region that is translated into the corresponding protein sequence, enables the most reliable annotation. However, it is the coding region that shows the highest degree of similarity with related sequences and therefore is the most susceptible to cross-hybridization events. Also, coding regions of eukaryotic genes suffer from alternative splicing events (see Figure 5.7). For these reasons, ESTs from coding region have to be thoroughly investigated before they are used as microarray probes. If the $3'$-untranslated region of a gene is extremely long, it is possible that the coding region may not be present in the cDNAs generated by reverse transcription process.

5.5 An EST Selection Strategy

Considering the above mentioned criteria, a recommended microarray probe selection strategy includes the following steps (see Figure 5.8): The first step is to select a list of genes relevant to the biological experiment. Then, if available, obtain the complete genomic DNA sequence for each of these genes of interest. Mask repetitive elements and the region beyond the first polyadenylation signal. Remove introns and untranslated regions from the remaining gene sequence. Use the remainder of the gene sequence to predict its protein sequence. Compare this predicted protein sequence with all other sequences in a given organism to identify homologous sequences. Use this information to mask regions of a gene that could cross hybridize with other closely related genes. Finally, select EST probes using gene segments with no apparent risk of cross hybridization.

Choice of an actual probe selection strategy depends on the application and the problem to be

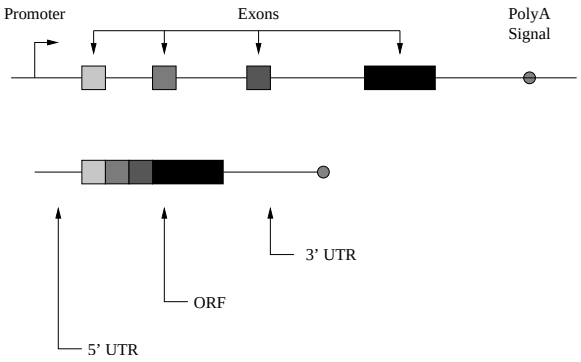

Figure 5.4: Selection of gene regions

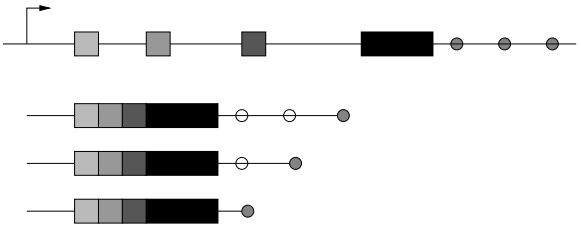

Figure 5.5: Alternative polyadenylation

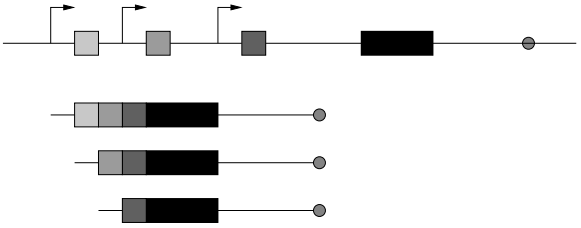

Figure 5.6: Alternative promoter usage

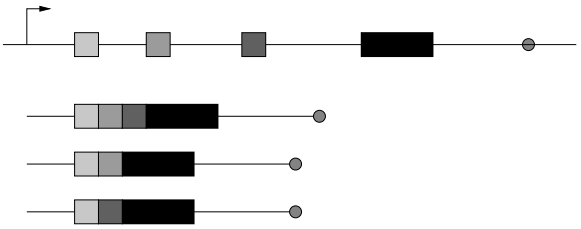

Figure 5.7: Alternative splicing

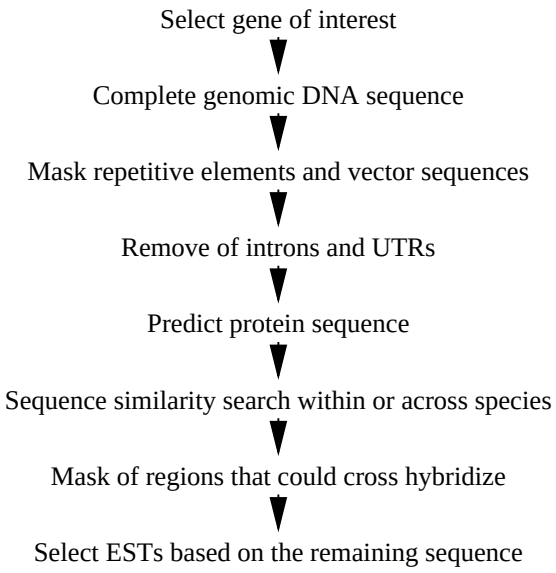

Figure 5.8: An EST selection strategy

solved. In general, high-throughput procedures do not require an intensive scrutiny of the spotted DNA fragments before the hybridization. On the down side, substantial analysis work has to be performed for each spot showing an interesting expression behavior. Targeted arrays, on the other hand, do not aim at identification of new genes, but allow monitoring of complex expression patterns. This objective requires careful annotation and quality control of DNA fragments to allow a reliable interpretation of results. Often, the choice of selection strategy is a compromise between experimental objectives and practicability. It is the decision of the user as to which strategy is the most suitable for the application.

Chapter 6

GeneSieve – Selection Strategy

This chapter describes the GeneSieve probe selection strategy. The GeneSieve selection process starts with gathering sequence information for a *model organism* and for an *organism of interest*. A model organism is one whose genome has been fully sequenced, that has been studied in great detail in the past years, and that is close to the organism being studied, in terms of evolution history. For example, the higher plant *Arabidopsis thaliana* can serve as a model organism for plant species such as pine and potato. For such a model organism, a complete list of genes, gene sequences, their protein products, and associated functional categories are available through various public databases, such as TIGR [50, 51], MIPS [38, 56], and TAIR [52, 20]. The organism of interest is the one being studied in the experiment. Often, only EST sequences, without proper annotation and functional categorization, are available for this organism. NCBI and TIGR maintain EST sequence databases for many such organisms. Section 6.1 provides an overview of the selection strategy. The following sections explain various steps involved in the selection process in detail.

6.1 Overview

The GeneSieve probe selection strategy consists of the following steps:

1. Protein sequences and functional categories: Obtain a complete set of proteins, their sequences, functional annotations, and the functional categories associated with them for a suitable model organism such as *Arabidopsis thaliana*.

2. EST sequences: Obtain a complete set of EST sequences and corresponding base quality scores for the organism of interest such as pine or potato.

3. Clustering: Assemble ESTs into contigs and singletons using a clustering program such as phrap. For each resulting contig, obtain its consensus sequence, list of constituent ESTs, and its alignments with these ESTs.

4. Sequence similarity search: Set up a stand-alone BLAST server for sequence similarity search. Configure local BLAST databases to incorporate protein sequences, EST sequences, and contig sequences.

5. Contig selection: Obtain a list of proteins that are relevant to the biological experiment. Select contigs showing homology with these proteins.

6. EST selection: Choose a single EST from each of the selected contigs based on length, proximity to the $3'$ or $5'$ end, protein homology, or the possibility of cross hybridization.

Figure 6.1 illustrates the GeneSieve probe selection process as a system. The remaining sections describe the steps involved in the selection process in detail.

6.2 Protein Sequences and Functional Categories

The first step in the GeneSieve selection process is to obtain a complete set of protein sequences for the model organism, their annotations, and their functional categories. We use protein sequences of the model organism for cross-species homology search because protein sequences are free of introns, untranslated regions, and nonsense repeats. Also, protein sequences are more conserved across different species than gene sequences. For model organisms, such as *Arabidopsis thaliana* or *Saccharomyces cerevisiae*, a complete set of proteins, their sequences, and associated functional annotations can be downloaded in FASTA format from the MIPS or TIGR FTP site [40, 63].

Functional categories help to speed up selection of a large number of genes that are similar in function or involved in the same biological process or biochemical pathway. Functional categorization also helps in better interpretation of microarray results. For model organisms, such as *Arabidopsis thaliana* or *Saccharomyces cerevisiae*, many of the genes have been assigned to functional categories. MIPS [38, 56, 40] and the Gene Ontology (GO) Consortium [6, 21] maintain functional category databases for *Arabidopsis thaliana*.

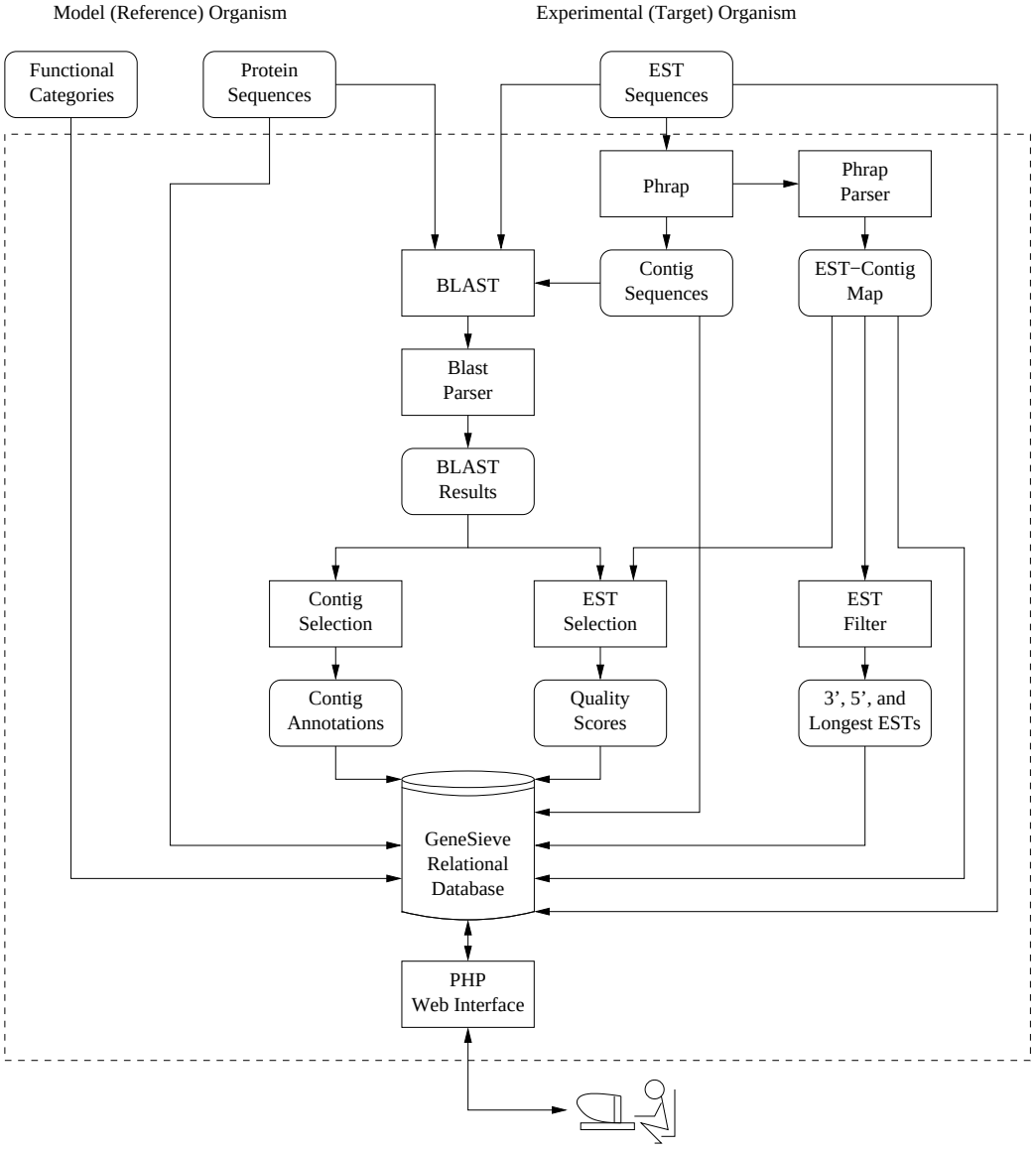

Figure 6.1: GeneSieve system architecture

MIPS

MIPS maintains a database for manually assigned functional categories for *Arabidopsis thaliana*. Arabidopsis proteins are assigned to functional categories based on experimental evidence or sequence similarity to other manually classified proteins. Proteins and associated functional categories can be downloaded from the MIPS web site [40]. Below is an example of the MIPS functional hierarchy:

```
01 METABOLISM
01.01 amino-acid metabolism
01.01.01 amino-acid biosynthesis
01.01.01.01 assimilation of ammonia, biosynthesis of the glutamate family
   At5g38710 proline oxidase, putative
   At5g04140 glutamate synthase (GLU1)
```

GO

The goal of Gene Ontology Consortium is to produce a dynamic, controlled vocabulary that can be applied to all eukaryotes even as knowledge of gene and protein roles in cells is accumulating and changing. GO provides three structured networks of defined terms to describe gene product attributes. These three ontologies or categories are as follows:

1. *Biological Process Ontology*: Biological process refers to a biological objective to which a gene or gene product contributes. An example of the biological process hierarchy is as follows:

   ```
   GO:0008150 : biological_process
      GO:0007610 : behavior
      GO:0000004 : biological_process unknown
      GO:0009987 : cellular process
         GO:0007154 : cell communication
            GO:0007155 : cell adhesion
                At5g03170 fasciclin-like arabinogalactan-protein (FLA11)
   ```

2. *Molecular Function Ontology*: Molecular function is defined as the biochemical activity (including specific binding to ligands or structures) of a gene product. An example of the molecular function hierarchy is as follows:

```
GO:0003674 : molecular_function
   GO:0016209 : antioxidant activity
      GO:0045174 : glutathione dehydrogenase (ascorbate) activity
      GO:0004362 : glutathione-disulfide reductase activity
      GO:0004601 : peroxidase activity
         At1g05260 peroxidase 3 (PER3) (P3)
```

3. *Cellular Component Ontology*: Cellular component refers to the place in the cell where a gene product is active. An example of the cellular component hierarchy is as follows:

```
GO:0005575 : cellular_component
   GO:0005623 : cell
      GO:0005622 : intracellular
         GO:0005737 : cytoplasm
            GO:0005829 : cytosol
               At1g07890 L-ascorbate peroxidase 1
```

6.3 EST sequences

The next step in GeneSieve is to obtain a set of EST sequences for the organism of interest. EST sequencing is the method of choice for many large scale sequencing projects because the transcriptome of most organisms is much smaller than their genomes, making this approach cost effective. Several public databases, such as TIGR [50, 51] and NCBI dbEST [10], maintain collections of EST sequences for many organisms. A collection of EST sequences for an organism can be downloaded from these web sites. We obtain pine and potato EST sequences from the Center for Computational Genomics and Bioinformatics (CCGB) at University of Minnesota [12] and TIGR [63] web sites respectively. For an organism, the number of sequences in the EST database depends on the complexity of the sequencing project and that of the organism itself.

Often, the corresponding base quality files are available along with the EST sequence files. The format of the quality file is similar to that of the sequence file, but it gives the quality of each base in EST sequences. Base quality scores are integers between 0 and 99 and reflect probability of errors in base calls at each position in an EST. Base quality score q is calculated using the transformation

$q = -10log_{10}(p)$. Here, p is the probability of error in base calling. Thus, a base quality score of 30 corresponds to an error probability of 1/1000.

6.4 Clustering

As discussed in Chapter 5, depending on the complexity of the cDNA library, a single gene might be represented by a large number of ESTs. Selecting more than one clones from the same gene adds redundancy to the microarray. If coverage and efficiency is an issue, redundancy has to be minimized. Clustering is the standard method to control redundancy in microarray probe selection. Many software packages are available for sequence clustering, such as phrap [22] and TIGR Assembler [19], allowing clustering of any EST sequence collection.

6.4.1 Phrap

We use phrap [22] for clustering EST sequences. Phrap is the most widely used program for automated contig assembly in genome projects. A contig is defined as a continuous sequence of DNA that has been assembled from overlapping DNA fragments. Phrap can handle very large sequence data sets, allows the use of entire EST sequence (not just the trimmed, high quality part), and uses a combination of user-supplied and internally computed data quality information to improve accuracy of the assembly in the presence of repeats. It constructs a consensus sequence for each contig as a summary of the highest quality parts of ESTs. It also provides extensive information about assembly, such as base quality for each contig sequence, a list of constituent ESTs for each contig, and the start and the end positions for contig-EST alignments. This information is useful in finding the longest EST, or the EST closest to the $3'$ or the $5'$ end of a contig, for example.

The phrap assembly algorithm consists of the following steps [15]:

1. Read in sequence and quality data, trim off any near-homopolymer runs at ends of reads, construct read complements.

2. Find pairs of reads with matching words. Eliminate exact duplicate reads. Do swat comparisons of pairs of reads which have matching words, compute (complexity-adjusted) swat score.

3. Find probable vector matches and mark so they aren't used in assembly.

4. Find near duplicate reads.
5. Find reads with self-matches.
6. Find matching read pairs that are "node-rejected" i.e. do not have "solid" matching segments.
7. Use pairwise matches to identify confirmed parts of reads; use these to compute revised quality values.
8. Compute LLR scores for each match (based on qualities of discrepant and matching bases). (Iterate above two steps).
9. Find best alignment for each matching pair of reads that have more than one significant alignment in a given region (highest LLR-scores among several overlapping).
10. Identify probable chimeric and deletion reads (the latter are withheld from assembly).
11. Construct contig layouts, using consistent pairwise matches in decreasing score order (greedy algorithm). Consistency of layout is checked at pairwise comparison level.
12. Construct contig sequence as a mosaic of the highest quality parts of the reads.
13. Align reads to contig; tabulate inconsistencies (read / contig discrepancies) and possible sites of misassembly. Adjust LLR-scores of contig sequence.

Phrap relies heavily on quality values assigned to each base by its companion program, phred. The use of base quality information, whenever available, along with EST sequence data is strongly recommended. It greatly improves the accuracy of assembly and the quality of consensus sequences. Phrap considers sequences for which there is only one EST as unreliable. It trims off ends of a contig if there are not at least two ESTs confirming each other's sequence.

6.5 Sequence Similarity Search

Once clustering is done and contig sequences are available, the next step is to search for sequence similarities between these contigs and known proteins in the model organism. This process associates a functional annotation with each contig. EST sequences can also be compared to contig sequences to estimate the probability of cross hybridization. Several algorithms, notably BLAST [4, 5] and FASTA [46, 45], are available to compare a query sequence against a sequence database.

6.5.1 BLAST

BLAST (Basic Local Alignment Search Tools) [4, 5] is a set of sequence comparison algorithms that can be used to search sequence databases for optimal local alignments to a query. BLAST improves the overall speed of searches, while retaining good sensitivity, by breaking the query and database sequences into small fragments (words) and initially seeking matches between words. The initial search is done for a short word with some minimum score when compared to the query using a given substitution matrix. Word hits are then extended in either direction in an attempt to generate an alignment with a score exceeding a threshold. The pair-wise alignments with scores exceeding the threshold are ranked by score.

BLAST programs are available as executables or as source code from the NCBI website [41]. A similar set of programs, called WU-BLAST [67], is also available from Washington University. We set up a BLAST server using the NCBI BLAST programs. This allows us to create custom searchable databases for contigs and ESTs in different organisms. Of course, querying is faster than submitting sequences to the BLAST server at NCBI. After installing the BLAST server, we have configured custom databases for protein sequences from the model organism, and for EST and contig sequences from several organisms of interest. Consequently, we can readily compare EST and contig sequences to themselves or to the protein sequences of the model organism.

6.6 Selection of Contigs

Selection of a contig requires finding its similarity with known proteins of the model organism and assigning an unambiguous annotation to it. For this, we need to establish "reasonable links" between contigs and known proteins. Based on these links, we can transfer annotation and functional categories of a protein to a contig.

We propose three approaches to establish links between contigs constructed from organism of interest and proteins of a model organism: bidirectional hits (BH), bidirectional best hits (BBH), and unidirectional best hits (UBH). There is a *similarity link* between two sequences if the E value from BLAST is less than 10^{-06}.

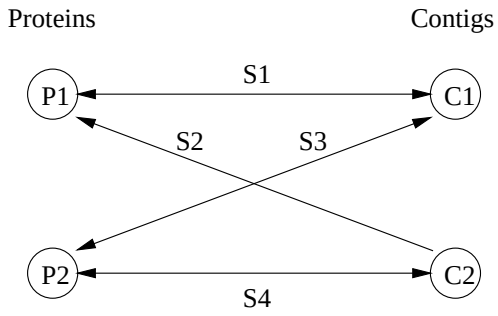

Figure 6.2: Bidirectional hits (BH)

6.6.1 Bidirectional Hits (BH)

A bidirectional hit is defined as follows: A protein P in the protein-dataset D_p and a contig C in the contig-dataset D_c are bidirectional hits if there are similarity links between them when P is searched against D_c and when C is searched against D_p. Figure 6.2 illustrates the concept of bidirectional hits. Here, $P1$ and $P2$ are proteins and $C1$ and $C2$ are contigs. There are bidirectional similarity links between $(P1, C1)$, $(P2, C1)$, and $(P2, C2)$. There is a similarity link between $(P1, C2)$, when $C2$ is searched against the protein database. But there is no similarity link between $(P1, C2)$ when $P1$ is searched against the contig database. $S1, S2, S3, andS4$ are the BLAST scores for the corresponding similarity links. So, by definition, $(P1, C1)$, $(P2, C1)$, and $(P2, C2)$ are bidirectional hits.

Often, BLAST does not give symmetric results, i.e., the BLAST score and the E value between a protein P and a contig C when searched using P as a query against the contig database can be different from the BLAST score and the E value when searched using C as a query against the protein database. As we apply a threshold for the maximum E value, it is possible that some similarity links are present in just one direction and not the other. The BH approach eliminates such asymmetric similarity links from consideration.

One drawback of the BH approach is that, often, it assigns more than one annotations to a contig. This causes problem while selecting contigs by functional annotations or functional categories. The same contig can be selected more than once, each time for a different annotation or a different functional category.

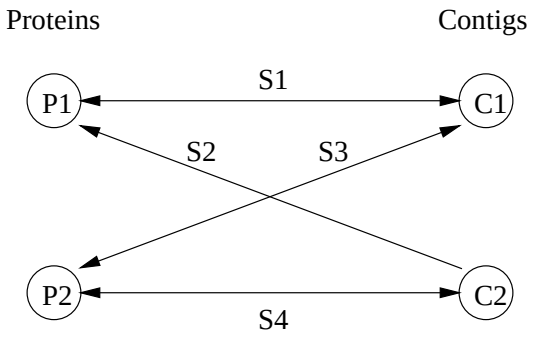

Figure 6.3: Bidirectional best hits (BBH)

6.6.2 Bidirectional Best Hits (BBH)

The bidirectional best hit (BBH) is a more restrictive approach than the BH. It overcomes the drawback of the BH approach. A bidirectional best hit is defined as follows: A protein P in the protein-dataset D_p and a contig C in the contig-dataset D_c are called bidirectional best hits if and only if P and C are bidirectional hits, there is no other contig C' in D_c that is more similar to P than C is to P, and there is no other protein P' in D_p that is more similar to C than P is to C. Figure 6.3 illustrates the concept of BBH. Here, $P1$ and $P2$ are proteins and $C1$ and $C2$ are contigs. There are bidirectional similarity links between $(P1, C1)$, $(P2, C1)$, and $(P2, C2)$. There is a similarity link between $(P1, C2)$, when $C2$ is searched against the protein database. But there is no similarity link between $(P1, C2)$ when $P1$ is searched against the contig database. $S1, S2, S3, and S4$ are the BLAST scores for the corresponding similarity links and $S1 > S2 > S3 >> S4$. So, by definition, only $(P1, C1)$ are bidirectional best hits.

The BBH criterion is stringent and suffers from low coverage. For example, if there are five contigs showing the highest similarity with a protein P, when searched using contig sequences as queries against the protein database, the one showing the highest similarity with P, when searched using protein sequence as query against the contig database, will be assigned the annotation of P. The remaining four will not be assigned any annotation, even though they show significant similarity with the protein P.

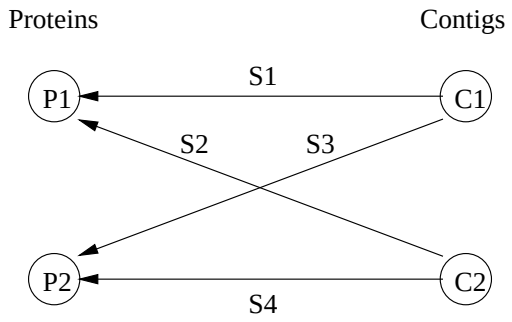

Figure 6.4: Unidirectional best hits (UBH)

6.6.3 Unidirectional Best Hits (UBH)

A unidirectional best hit is defined as follows: A protein P in the protein-dataset D_p is called the unidirectional best hit of a contig C in the contig-dataset D_c if and only if there is a similarity link between them when C is searched against S_p, and there is no other protein P' in D_p that is more similar to C than P is to C. Figure 6.4 illustrates the concept of UBH. Here, $P1$ and $P2$ are proteins and $C1$ and $C2$ are contigs. There are similarity links between $(P1, C1)$, $(P1, C2)$, $(P2, C1)$, and $(P2, C2)$, when contigs $C1$ and $C2$ are searched against the protein databases. $S1, S2, S3, and S4$ are the corresponding BLAST scores and $S1 > S2 > S3 >> S4$. So, by definition, $(P1, C1)$, $(P1, C2)$ are unidirectional best hits.

The UBH approach is a compromise between the liberal BH approach and the stringent BBH approach. It does not assign more than one annotation to any contig. At the same time, it ensures that a contig is assigned an annotation if it shows a similarity link to at least one known protein. For this reason, we use the UBH method in the GeneSieve probe selection process.

We also calculate a *protein coverage score* (PC) for each contig showing a similarity link to a protein in the model organism. Whenever there are multiple contigs available for the selection for a single protein, we select one with the highest protein coverage score. Protein coverage score PC is defined as follows: Let C be a contig showing a similarity link to a protein P. Let L_P be the length of the protein P and L_{PC} be the length of the protein P that aligns to the contig C. Then, the protein coverage score is

$$PC = \frac{L_{PC}}{L_P} \tag{6.1}$$

6.6.4 Evaluation Criteria

We compare BH, BBH, and UBH based on the following criteria:

1. *Assignment of unambiguous annotations*: We compare the three methods based on their ability to assign unambiguous annotations to the contigs, i.e., no contig should be assigned more than one annotation. Otherwise, a contig can be selected more than once for a microarray, each time representing a different gene or a different functional category.

2. *Coverage*: Coverage is defined as the number of contigs that are annotated using a given method. The higher the coverage the greater number of contigs that will be available for the selection.

6.7 Selection of EST in Contig

Once GeneSieve selects a contig to represent a gene or functional category of interest, the next step is to select a single EST from this contig. There are several factors that may be considered to select an EST, such as its similarity to a known protein, its potential to cross hybridize with other genes, its length, and its proximity to the $3'$ or $5'$ end of the contig. These factors can be considered alone or in conjugation with other factors to assess the quality of an EST.

6.7.1 Quality Function

We assign an overall quality score Q to each EST derived from three other scores: protein homology (PH), cross hybridization (CH), and relative length (RL), which we define now.

Let E be an EST that belongs to a contig C. Let protein P be the unidirectional best hit of contig C with BLAST score S_{CP}. If protein P is also the best hit for EST E, when E is searched against the protein database and if S_{EP} is the BLAST score between them, then the *protein homology score* for EST E is

$$PH = \frac{S_{EP}}{S_{CP}} \tag{6.2}$$

GeneSieve BLASTs each EST and each contig against the protein database. BLAST scores for the best hits are considered for calculating PH. The best hit for a contig and that for a constituent ESTs should be the same protein, otherwise the PH score for that EST is 0. The PH score indicates how well a contig-protein alignment is represented by the constituent EST. The typical range for PH is 0 to 1 (inclusive), but in a few cases it exceeds 1.

Let contig C_1 be the best hit with BLAST score S_{EC_1} and contig C_2 be the second best hit with BLAST score S_{EC_2}, when an EST E is searched against the contig database. If E is a constituent EST of contig C_1, then the cross hybridization score for EST E is defined by the following equation:

$$CH = \frac{S_{EC_2}}{S_{EC_1}} \tag{6.3}$$

Here, we BLAST each EST to the contig database for the same organism. The best hit contig should be the one E belongs to; otherwise CH = 1. The CH score estimates the probability of an EST hybridizing with other contigs. The range for the CH is 0 to 1 (inclusive). The higher CH is, the greater the probability that the given EST will cross hybridize with non-target mRNA sequences.

Let E be an EST which belongs to contig C. Let L_C be the length of C and L_{CE} be the length of C that aligns to E. Then, the relative length score of EST E is

$$RL = \frac{L_{CE}}{L_C} \tag{6.4}$$

RL refers to the length of a contig covered by its constituent ESTs. In EST clustering, phrap trims off ends of contigs if there are not at least two ESTs confirming each other's sequences. As a result, the whole length of an EST may not contribute to the consensus sequence of a contig. The typical range for the RL score is 0 to 1 (inclusive).

Quality Score (Q)

Based on these three quality parameters, we assign a quality score to each EST. The quality score Q for an EST is defined as follows:

$$Q = PH - CH + RL \tag{6.5}$$

The typical range for Q is -1 to 2.

6.7.2 EST Selection Methods

We calculate protein homology score, cross hybridization score, relative length score, and over all quality score for each EST. Once we have selected a contig based on its similarity to the protein of interest, a single EST can be chosen from this contig using one of the following methods:

1. *Maximum Length*: Among all the ESTs in a contig, select the one that is the longest. If there is more than one EST with the same maximum length, select any one.
2. $5'$ *Proximity*: Among all the ESTs in a contig, select one that aligns to the $5'$ end of the contig. If there is more than one such EST, select one that is the longest among them.
3. $3'$ *Proximity*: Among all the ESTs in a contig, select one that aligns to the $3'$ end of the contig. If there is more than one such EST, select one that is the longest.
4. Maximum PH Score: Among all the ESTs in a contig, select one with maximum protein homology score.
5. Minimum CH Score: Among all the ESTs in a contig, select one with minimum cross hybridization score.
6. Maximum RL Score: Among all the ESTs in a contig, select one with maximum relative length score.
7. Maximum Quality Score: Among all the ESTs in a contig, select one with maximum overall quality score.

6.7.3 Evaluation Criteria

We compare these seven EST selection methods based on the following criteria:

1. *Coverage*: Coverage is defined as the fraction of the contigs in an organism that are available for selection by each method.
2. *Average PH*: Average protein homology score for the ESTs selected by each method.
3. *Average CH*: Average cross hybridization score for the ESTs selected by each method.
4. *Average RL*: Average relative length score for the selected ESTs.

5. *Average Q*: Average quality score for the ESTs selected by each method.

Singletons contain only one EST. So, no matter which selection method is used, the selected EST will be the same. For this reason, we exclude singletons when evaluating results of different EST selection methods.

Chapter 7

Results

We apply the GeneSieve probe selection strategy described in the previous chapter to the loblolly pine (*Pinus taeda*) and potato (*Solanum tuberosum*) EST data sets. We use *Arabidopsis thaliana* as the model organism for pine and potato. Results for pine and potato datasets are presented in this chapter. We obtain a total of 27,288 protein sequences for *Arabidopsis thaliana*, along with their functional annotations, from TIGR [63]. Functional categories for these proteins are obtained from MIPS [40] and Gene Ontology [21].

7.1 Pine

We obtained the EST data set for pine from the ftp site of Center for Computational Genomics and Bioinformatics (CCGB) at University of Minnesota [12]. That data set contains 75,047 EST sequences, derived from six pine xylem libraries. We exclude ESTs that are shorter than 100 base pairs from the contig assembly. After removal of nearly duplicate sequences, phrap assembles 59,525 ESTs into 7,141 contigs and 13,331 singletons. We use the following assembly parameters in phrap: minmatch=50 and minscore=100. Table 7.1 summarizes the results of clustering of the pine ESTs.

7.1.1 Contig Selection

We compare all 20,472 pine contigs (7,141 contigs + 13,331 singletons) for sequence similarity against 27,288 Arabidopsis proteins, in both directions, using BLAST. We then use these BLAST results to

Number of ESTs	59,525
Number of contigs	7,141
Number of singletons	13,331
Total (contigs+singletons)	20,472

Table 7.1: Clustering of pine ESTs

assign annotations to the pine contigs using three different methods: BH, BBH, and UBH. Table 7.2 shows the results of the contig selection methods for pine.

	BH	BBH	UBH
Number of contigs and singletons covered	8,833	3,743	9,065
Number of Arabidopsis proteins covered	15,803	3,743	5,563
Avg number of proteins assigned to each contig	17.09	1	1
Avg number of contig assigned to each protein	9.55	1	1.63

Table 7.2: Comparison of contig selection methods for pine

As shown in the table, the BH method covers a large number of Arabidopsis proteins and pine contigs (15,805 and 8,833 respectively), but it suffers from redundancy and assigns ambiguous annotations to the contigs. The average number of proteins linked to each pine contig is 17, i.e. 17 different annotations can be assigned to the same contig, or that the same contig can be selected for 17 different proteins or annotations. Also, the average number of contigs linked to each protein is 9, i.e. 9 contigs are available for selection for each protein of interest. The BBH method assigns at the most one annotation to each protein and at the most one contig to each protein, but it suffers from very low coverage. Only 3,743 pine contigs and the same number of Arabidopsis proteins are available for the selection. The UBH method is a compromise between the BH and the BBH. The number of contigs covered by the UBH is slightly higher than that by BH, though the number of proteins covered by the UBH is much lower than that by the BH. The UBH method assigns unambiguous annotations to pine contigs. Also, the average number of contigs available for selection for each protein is also very low (1.63 contigs/protein).

Annotation Analysis

Once annotations are assigned to the contigs based on their similarity with the known proteins in Arabidopsis, we analyze these annotations to find proteins that are present in both Arabidopsis and pine. This analysis also helps identify proteins that are present in Arabidopsis but not in pine. Results of the annotation analysis are summarized in Table 7.3. Out of a total of 27,288 Arabidopsis proteins, 16,221 (59%) proteins have similarity a link to at least one pine contig. Out of 11,067 Arabidopsis proteins that do not have a similarity link to any pine contig, 8,187 (74%) proteins have uninformative annotations (i.e., 'putative protein', 'unknown protein', 'hypothetical protein', or 'expressed protein'). Out of 12,717 Arabidopsis proteins with informative annotations, only 2,880 (22%) do not have a similarity link to any pine contig. Though only 20.4% of the Arabidopsis proteins are covered by the UBH method, they account for 37% of the unique annotations in Arabidopsis.

	Arabidopsis Proteins	Proteins showing similarity to pine contigs ($p < 10^{-06}$)	Proteins showing similarity to pine contigs by UBH (best hit, $p < 10^{-06}$)
All Proteins	27,288	16,221	5,563
Putative Proteins (P)	4,958	2,520	826
Unknown Proteins (U)	2,108	916	378
Expressed Proteins (E)	2,672	1,441	730
Hypothetical Proteins (H)	4,833	1,507	385
Total (P+U+E+H)	14,571	6,384	2,319
Informative Annotations	12,717	9,837	3,244
Unique Annotations	6,788	4,999	2,515

Table 7.3: Annotation analysis for pine contigs

Annotation analysis can give the following information:

- Proteins present in both Arabidopsis and pine.
- Proteins present in Arabidopsis but not in pine: These might be proteins responsible for some specialized function in Arabidopsis, and hence, not present in pine. Or these are the missing genes as the EST libraries of pine may not represent all the genes in pine. Inclusion of EST derived from other pine tissues can be useful in the later case.
- A list of unique annotations that can be helpful while searching for probes by gene/protein

names.

7.1.2 EST Selection

Once annotations are assigned to the pine contigs using the UBH method, we select a single EST from each contig using seven different methods: the longest EST, $3'$ proximity, $5'$ proximity, maximum PH, maximum CH, maximum RL, and maximum quality score. Results of these methods are summarized in Table 7.4. For each selection method, the table gives the average values of the three quality parameters and the overall quality score for the selected ESTs. We exclude singletons from consideration as there is no choice for selection of ESTs from singletons.

Method of Selection /Average Score	Protein Homology (PH)	Cross Hybridization (CH)	Relative Length (RL)	Quality (Q)	ESTs Selected
All ESTs	0.400	0.304	0.470	0.566	34,739
Maximum Length	0.811	0.187	0.851	1.475	4,496
$5'$ Proximity	0.714	0.189	0.776	1.300	4,496
$3'$ Proximity	0.682	0.178	0.781	1.284	4,496
Maximum PH	0.932	0.196	0.808	1.543	4,496
Minimum CH	0.508	0.112	0.642	1.037	4,496
Maximum RL	0.806	0.182	0.858	1.481	4,496
Maximum Q	0.919	0.180	0.837	1.576	4,496
Maximum Q ($CH <= 0.5$)	0.933	0.128	0.831	1.635	4,027

Table 7.4: Quality analysis for pine ESTs

Maximum average PH (0.932) and maximum average RL (0.858) are obtained when ESTs are selected by PH and RL respectively. Minimum average CH (0.112) is obtained when selection is based on CH. Selection of ESTs by Q gives maximum achievable quality score. General trend is that the CH increases with the PH and the RL. The ESTs selected by $3'$ proximity show lower cross hybridization (0.178), but it is far from minimum that can be achieved (0.112) for the same population of ESTs. Also, the average quality of selected ESTs is low (1.284). Analysis of the pine contigs reveal that only a small fraction of pine contigs align to the $5'$ or the $3'$ end of an Arabidopsis protein (9% and 11% respectively). This suggest that most of the ESTs selected from the $5'$ or the $3'$ end of the contigs actually come from the coding regions of the genes. Selection of ESTs by PH, CH, or RL gives the best value for the parameter used for selection, but gives poor values for the

other parameters.

Selection by Q gives averages for PH, CH, and RL that are close to the best achievable averages for those parameters. An additional filter, such as $CH <= 0.5$, can greatly reduce cross hybridization and improve the overall quality of the selected ESTs. This filter first excludes all the ESTs with $CH > 0.5$, and then selects ESTs with the highest quality score from each contig. Similar filters can also be applied to PH and RL. A combination of the filters, such as ($PH >= 0.2$ & $CH <= 0.5$ & $RL >= 0.2$), can be used to find a probe set of desired quality to meet the experimental requirements.

7.2 Potato

We retrieved a total of 87,677 potato EST sequences and corresponding base quality scores from the TIGR ftp site [63]. All the EST sequences are longer than 99 base pairs, and so, we use them all for the contig assembly. GeneSieve assembles a total of 87,677 potato ESTs into 21,050 contigs and 11,350 singletons using phrap. Parameters used for phrap assembly are: minmatch=200 and minscore =100. Here, we use higher value for phrap parameter minmatch to avoid exceptionally long runtime. The higher number of contigs in potato can be because of the use of base quality scores as input to phrap, the use of higher value of assembly parameter mimmatch, or the quality of EST libraries. Table 7.5 summarizes the results of clustering of the potato ESTs.

Number of ESTs	87,677
Number of contigs	21,050
Number of singletons	11,350
Total (contigs+singletons)	32,400

Table 7.5: Clustering of potato ESTs

7.2.1 Contig Selection

We compare all 32,400 potato contigs and singletons to 27,288 Arabidopsis proteins using BLAST. We use these BLAST results to assign annotations to the potato contigs using the BH, the BBH, and the UBH methods. Results of these contig selection methods are presented in Table 7.6:

Results for different contig annotation methods for potato are consistent with those for pine. Again, the number of contigs covered by the UBH method is slightly higher than that by the BH

	BH	BBH	UBH
Number of contigs and singletons covered	25,602	7,458	26,175
Number of Arabidopsis proteins covered	20,912	7,458	10,338
Avg number of proteins assigned to each contig	19.63	1	1
Avg number of contigs assigned to each protein	24.03	1	2.53

Table 7.6: Comparison of contig selection methods for potato

method. The UBH assigns unambiguous annotations to the potato contigs. The UBH method covers 81% of the potato contigs and 38% of the Arabidopsis proteins, while the respective numbers for pine are 44% and 20% only. With the BBH method, 27% of Arabidopsis proteins are linked to potato contigs, while only 14% of Arabidopsis proteins are linked to pine contigs.

Annotation Analysis

Results of annotation analysis are summarized in Table 7.7. Annotation analysis reveals that proteins covered by the UBH method in potato represent 57% of unique annotations present in Arabidopsis, while the corresponding number for pine is 37%. Comparison of pine and potato results suggest that potato and Arabidopsis have more proteins in common than pine and Arabidopsis have. This may be due to the poor coverage of the pine EST libraries or because of the fact that Arabidopsis is genetically closer to potato than it is to pine.

	Arabidopsis Proteins	Proteins showing similarity to potato contigs ($p < 10^{-06}$)	Proteins showing similarity to potato contigs by UBH (best hit, $p < 10^{-06}$)
All Proteins	27,288	21,185	10,338
Putative Proteins (P)	4,958	3,637	1,761
Unknown Proteins (U)	2,108	1,452	808
Expressed Proteins (E)	2,672	1,989	1,347
Hypothetical Proteins (H)	4,833	2,556	927
Total (P+U+E+H)	14,570	9,634	4,843
Informative Annotations	12,718	11,551	5,495
Unique Annotations	6,788	6,223	3,934

Table 7.7: Annotation analysis for potato contigs

7.2.2 EST Selection

Table 7.8 shows results of the seven EST selection methods for potato. Overall results are in agreement with those obtained for pine. Maximum average PH and RL are higher for potato than those for pine (0.97 vs. 0.93 and 0.81 vs. 0.89). Minimum average cross hybridization is also higher for potato than that for pine (0.54 and 0.29 respectively). Higher value for CH may be the result of a large number of contigs with high similarity as more stringent value of parameter minmatch is used in phrap assembly. A less stringent assembly process may merge some of these highly similar contigs and bring the average CH to a lower value. The best possible average of parameter is obtained when selection is based on that parameter. Maximum average value for PH and RL are higher for potato than for pine.

Method of Selection /Average Score	Protein Homology (PH)	Cross Hybridization (CH)	Relative Length (RL)	Quality (Q)	ESTs Selected
All ESTs	0.696	0.565	0.739	0.869	66,883
Maximum Length	0.762	0.479	0.828	1.110	17,725
$5'$ Proximity	0.802	0.470	0.857	1.188	17,725
$3'$ Proximity	0.829	0.457	0.866	1.238	17,725
Maximum PH	0.986	0.477	0.848	1.357	17,725
Minimum CH	0.769	0.407	0.846	1.207	17,725
Maximum RL	0.832	0.459	0.894	1.266	17,725
Maximum Q	0.971	0.448	0.873	1.395	17,725
Maximum Q ($CH <= 0.5$)	0.965	0.218	0.861	1.607	10,166

Table 7.8: Quality analysis for potato ESTs

As shown in Table 7.8, ESTs selected by $3'$ and $5'$ proximity show less protein homology. Analysis of the potato contigs reveals that only a small fraction of the pine contigs align to the $5'$ end or the $3'$ end of a Arabidopsis protein (26% and 25% respectively). This suggest that most of the ESTs selected from the $5'$ or the $3'$ end of the contigs actually come from the coding regions of the genes. Selection of ESTs by PH, CH, or RL gives better average for the parameter used for selection, but it gives poor averages for the other parameters and the overall quality.

Close to the best averages for PH, CH and RL can be obtained when ESTs are selected based on overall quality. Also, an additional filter, such as ($CH <= 0.5$) greatly reduces average cross hybridization and improves overall quality of the selected ESTs. Similar filters can be applied to

PH and RL.

Correlation between quality parameters

Table 7.9 gives correlation between the parameters used to calculate overall quality. The table shows that there is a high degree of correlation between PH and RL. But CH is not strongly correlated with either PH or RL.

	PH	CH	RL	Q
PH	1.0000	0.0961	0.8110	0.8397
CH	0.0961	1.0000	0.2237	0.5621
RL	0.8110	0.2237	1.0000	0.8917
Q	0.8397	0.5621	0.8917	1.0000

Table 7.9: Correlation between quality parameters

7.3 Comparison of Pine3 Protocol and GeneSieve

The Pine3 protocol is used to select EST probes that represent the genes related to the stress response in loblolly pine for the third round of the microarray experiments in Expresso [58]. In this section we compare the Pine3 EST selection protocol with GeneSieve. The ESTs selected by the Pine3 are compared with the ESTs selected by GeneSieve based on the average values for the quality parameters and the overall quality. The Pine3 protocol consists of the following steps:

1. Prepare a list genes of interest based on previous results or literature survey. These are candidate genes likely to respond to drought stress conditions in loblolly pine.

2. Search for these gene names in the *Arabidopsis thaliana* protein database, and identify Arabidopsis proteins with these names.

3. Obtain amino acid sequences for these proteins.

4. Using BLAST, search the EST database of *Pinus taeda* using the Arabidopsis proteins as query sequences. Obtain ESTs that have a similarity link with Arabidopsis proteins. Here, a similarity link exist if the E value for BLAST score is less than $1e^{-04}$.

Method of Selection /Average Score	Protein Homology (PH)	Cross Hybridization (CH)	Relative Length (RL)	Quality (Q)	ESTs Selected
All ESTs	0.357	0.369	0.440	0.429	7,412
Maximum Length	0.874	0.238	0.906	1.542	1,193
$5'$ Proximity	0.827	0.239	0.870	1.450	1,193
$3'$ Proximity	0.812	0.226	0.871	1.456	1,193
Maximum PH	0.954	0.254	0.885	1.587	1,193
Minimum CH	0.723	0.196	0.800	1.330	1,193
Maximum RL	0.866	0.241	0.912	1.536	1,193
Maximum Q	0.945	0.239	0.899	1.604	1,193
Maximum Q ($CH <= 0.5$)	0.921	0.135	0.882	1.669	1,010
Pine3	0.689	0.184	0.712	1.216	1,682

Table 7.10: Quality analysis for stress responsive genes in pine: Pine3 protocol vs. GeneSieve

5. Select ESTs closer to the $3'$ end of the proteins.

A total of 280 keywords and gene names were provided by the biologists. These genes are believed to participate in the stress response mechanism in pine and are selected based on the results of the previous experiments and literature survey. A total of 1,874 ESTs are selected using the Pine3 protocol representing 1,092 distinct Arabidopsis proteins. Here, we analyze quality of these ESTs using the quality parameters defined in GeneSieve. Out of 1,874 ESTs present in Pine3, 1,682 are covered in the GeneSieve results. These ESTs belong to 1,082 distinct contigs, i.e., on average 1.55 ESTs are selected from the same contig. This indicates some redundancy in the probes selected by the Pine3 protocol. GeneSieve approach ensures that, when desired, only one EST be selected from each contig. Using the GeneSieve annotations, we select the contigs matching these keywords. Then, we select a single EST from each of these contigs by each of the seven methods described in Section 6.7.2. The average quality of the ESTs selected by the quality score is much higher that for the ESTs selected by the Pine3 protocol (1.604 and 1.216 respectively). Table 7.10 summarizes the results of different EST selection methods for the 280 keywords/gene names.

Chapter 8

GeneSieve – A Web-based EST Probe Selection Tool

We have designed a PostgreSQL database to store the results obtained using the GeneSieve protocol. We have also designed a web interface to query this database in a user-friendly manner. The database schema is described in Section 8.1. The sample queries that can be performed on this database are presented in Section 8.2. Important features of the web interface are presented in Section 8.3 using a working search example.

8.1 Database Schema

We have designed a database to store the sequence information, functional categories, and the results obtained using the GeneSieve protocol on pine and potato EST datasets. The database consists of the following tables:

1. Arabidopsis_Sequences (AT_no, Annotation, Gene_Sequence, Coding_Sequence, Protein_Sequence): This table contains information about the Arabidopsis genes such as their ids, annotations, genomic sequences, protein-coding nucleotide sequences, and amino acid sequences. This information is obtained from the TIGR ftp site.

2. MIPS_Categories (AT_no, Category_id, Category_Name): This table stores MIPS functional

categories and Arabidopsis genes belonging to those categories.

3. GO_Categories (AT_no, GO_id, GO_Annotation, GO_Type): This table contains Arabidopsis genes and corresponding GO annotations as obtained from the Gene Ontology Consortium.

4. Contig (Organism, Contig_id, Contig_Sequence): This table stores the contig ids and their consensus sequences obtained as a result of contig assembly procedure using phrap.

5. EST (Organism, EST_id, EST_Sequence): This table stores EST sequences for different organisms. These sequences are obtained from various public sources such as TIGR and CCGB.

6. EST_Contig_Mapping (Organism, Contig_id, EST_id, EST_length, Sr_no, LLR_Score, Start_pos_EST, End_pos_EST, Start_pos_Contig, End_pos_Contig): This table stores results of the contig assembly procedure using phrap. For each contig, it stores its constituent ESTs, their length, and start and end positions of their alignments with the contig.

7. 3p_5p_EST (Organism, Contig_id, Contig_length, Longest_EST_id, 5p_EST_id, 3p_EST_id): This table gives the longest EST, the $3'$ EST, and the $5'$ EST for each contig as defined in Chapter 6.

8. Quality_Table (Organism, EST_id, Contig_id, AT_no, Annotation, Contig2protein_Score, Protein_Coverage, Protein_Homology, Cross_Hybridization, Relative_Length, Quality_Score, Sequence_Identity): This table stores results of the GeneSieve protocol. For each EST, it gives the corresponding contig, Arabidopsis protein, annotation, BLAST score for contig-protein alignment, values of the three quality parameters, overall quality score, and its sequence identity with the second best contig hit.

The relationships among these tables are shown in Figure8.1. The Arabidopsis sequences are linked to the MIPS and GO functional categories by the AT numbers. The quality table is connected to the Arabidopsis sequences and the MIPS and GO functional categories by the AT numbers. The quality table is also connected to the EST and the contig sequences by EST ids and contig ids respectively.

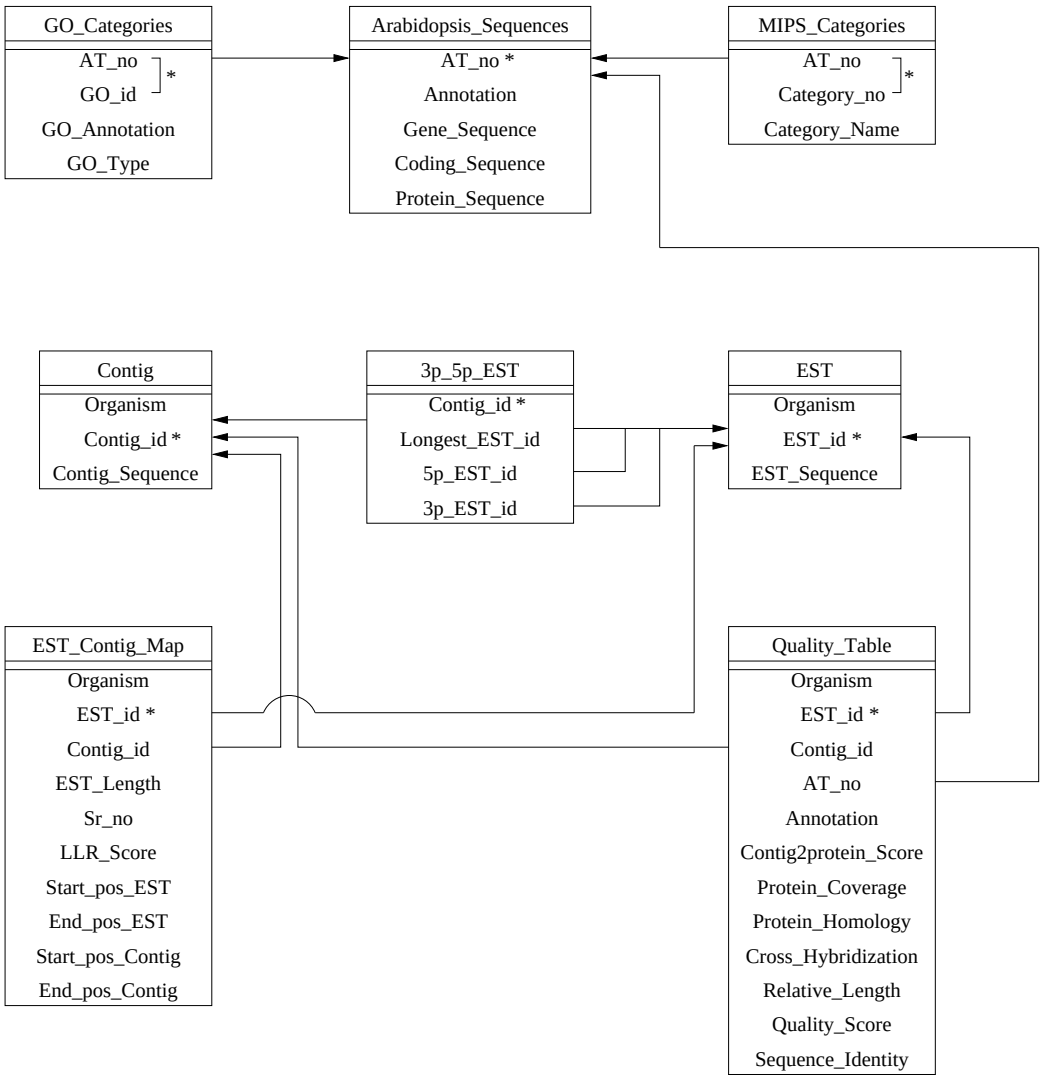

Figure 8.1: GeneSieve: database schema

8.2 Database Search

The GeneSieve database supports a wide variety of searches because of the inter-connectivity of various tables. Some of the searches that can be performed on the GeneSieve database are as follows:

1. Search by annotation: ESTs are associated to the Arabidopsis proteins by sequence homology. This allows one to search the database for ESTs based on protein name or key words, such as ascorbate peroxidase, hexokinase, or isoflavon. A key word, such as kinase, will match to all the kinases present in the database. Annotation search can also be used to identify possible homologs of a gene in different organisms. For example, search with keyword "ascorbate peroxidase" reveals that there are 8 Arabidopsis proteins matching this annotation. Pine has 8 contigs matching the same annotation while potato has 34.

2. Search by EST id: The database can also be searched by EST ids, i.e. NXCI_004_A05_F or NXCI_006_C05_F. This can be helpful in assigning putative annotation and functional categories to the EST of interest. It can also be used to get the sequence of an EST.

3. Search by contig id: The database can be searched by contig or singleton ids, i.e. Contig7730 or Singleton5314. This can be used to get the consensus sequence of a contig or to get the list of its constituent ESTs. It can also be used to assign annotation to a contig.

4. Search by Arabidopsis protein id (AT number): Search by an AT number, such as At1g07890.1, can be used retrieve all the contigs and ESTs in a organism that match the protein of interest. It can also be used to find homologs of a known protein in different organisms.

5. Search by MIPS or GO categories: The database can be searched by MIPS or GO functional category name or number, such as stress response (11.01), amino-acid metabolism (01.01), or pentose-phosphate pathway (02.07). This type of search is helpful in selecting a large number of ESTs based on their functional similarity.

6. EST selection: Above mentioned searches can be further refined to select a single EST from a contig or a protein of interest. A single EST can be selected from a contig of interest based on its length, proximity to the $3'$ or the $5'$ end of the contig, protein homology, cross hybridization, relative length, or overall quality score. A single EST can also be selected to represent a protein of interest based on one of the quality parameters or the overall quality score.

8.3 Web Interface

We have designed a web interface and connected it to the database using PHP. This allows one to query the database in a user-friendly manner. In this section, a series of screen shots are presented to show the important features and functionality of the GeneSieve web interface.

1. **Main Search page**: Figure 8.2 shows the main search page of GeneSieve. Each organism available in GeneSieve is presented by an icon. An organism of interest can be selected by clicking on one of these icons. Below the icons are the links to initiate an EST search by annotation, MIPS or GO functional category, EST id, contig id, or Arabidopsis protein id. These links lead to the relevant search pages. In this example, pine is selected as the organism of interest.

2. **Query Page**: Figure 8.3 shows the query page for annotation search. Multiple annotations can be searched simultaneously, but each of them should be entered as a separate line. Filters can be applied on the upper bounds and lower bounds of the PH, CH, RL, overall quality score, and protein coverage score. In this example, search is initiated for the annotation 'ascorbate peroxidase'. Filters are applied on the lower bounds of PH (0.5) and RL (0.2).

3. **Result Page**: Figure 8.4 shows the search results for 'ascorbate peroxidase' in pine. Results can be sorted by any attribute by clicking on that attribute name. Query can be further refined to select a single EST for each contig Arabidopsis protein based on its length, $3'$ or $5'$ proximity, one of the quality parameters, or the overall quality score. Selected ESTs can be added to the list of microarray probes.

4. **Microarray Probe List**: Figure 8.5 shows the pine ESTs selected as microarray probes. Similarly, ESTs can be searched using different annotations or functional categories and selected ESTs can be added to the list of microarray probes. ESTs once selected are highlighted in gray and can not be selected again. ESTs can also be deleted from the microarray probe list.

5. **EST/Contig/Protein Details**: Details of an EST, a contig, or an Arabidopsis protein can be viewed by clicking on their ids Figure 8.6 shows the sequence and the list of constituent ESTs for a pine contig 5940. For an Arabidopsis protein, a similar page shows its genomic

sequence, protein coding sequence, and amino acid sequence, and MIPS and GO functional categories. Links to the corresponding BLAST reports are provided at the bottom of the page.

6. **Functional Categories**: Figure 8.7 shows the MIPS and GO functional categories for Arabidopsis protein 'At4g29130'.

7. **BLAST Reports**: Figure 8.8 shows the BLAST report for pine contig 5940.

Figure 8.2: The main search page of GeneSieve. An organism can be selected by clicking on the respective icon. Search can be performed by using an annotation, a MIPS or GO functional category, an EST id, a contig id, a protein id, or a combination of all.

Figure 8.3: GeneSieve query page for annotation search. More than one annotations or keywords can be entered simultaneously in the search box. Filters can be applied on the quality parameters to control the quality of the microarray. Here, we search for the annotation "ascorbate peroxidase" in pine.

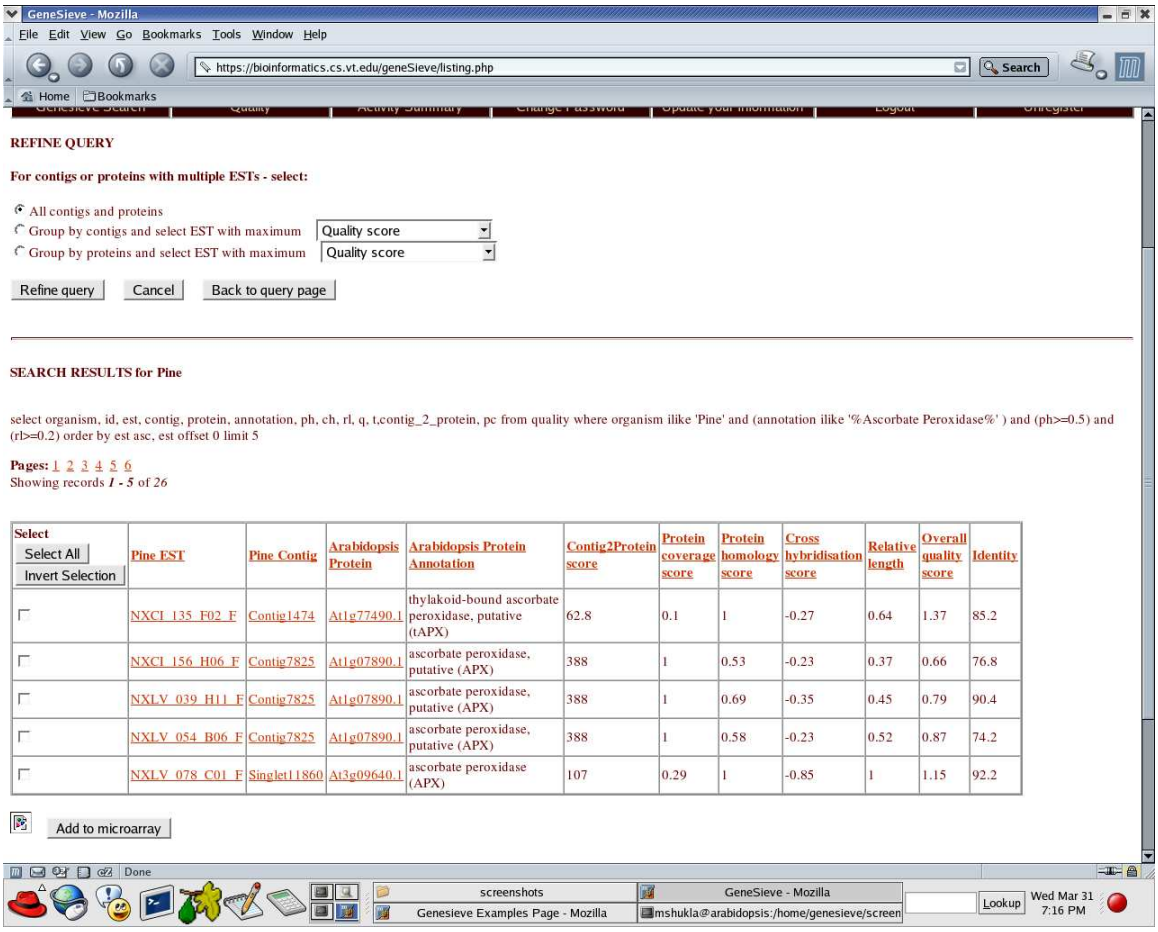

Figure 8.4: Results for the annotation search with the keyword "Ascorbate Peroxidase". Search can be refined to select a single EST from each contig or protein based on its length, $3'$ or $5'$ proximity, one of the quality parameters, or the overall quality score.

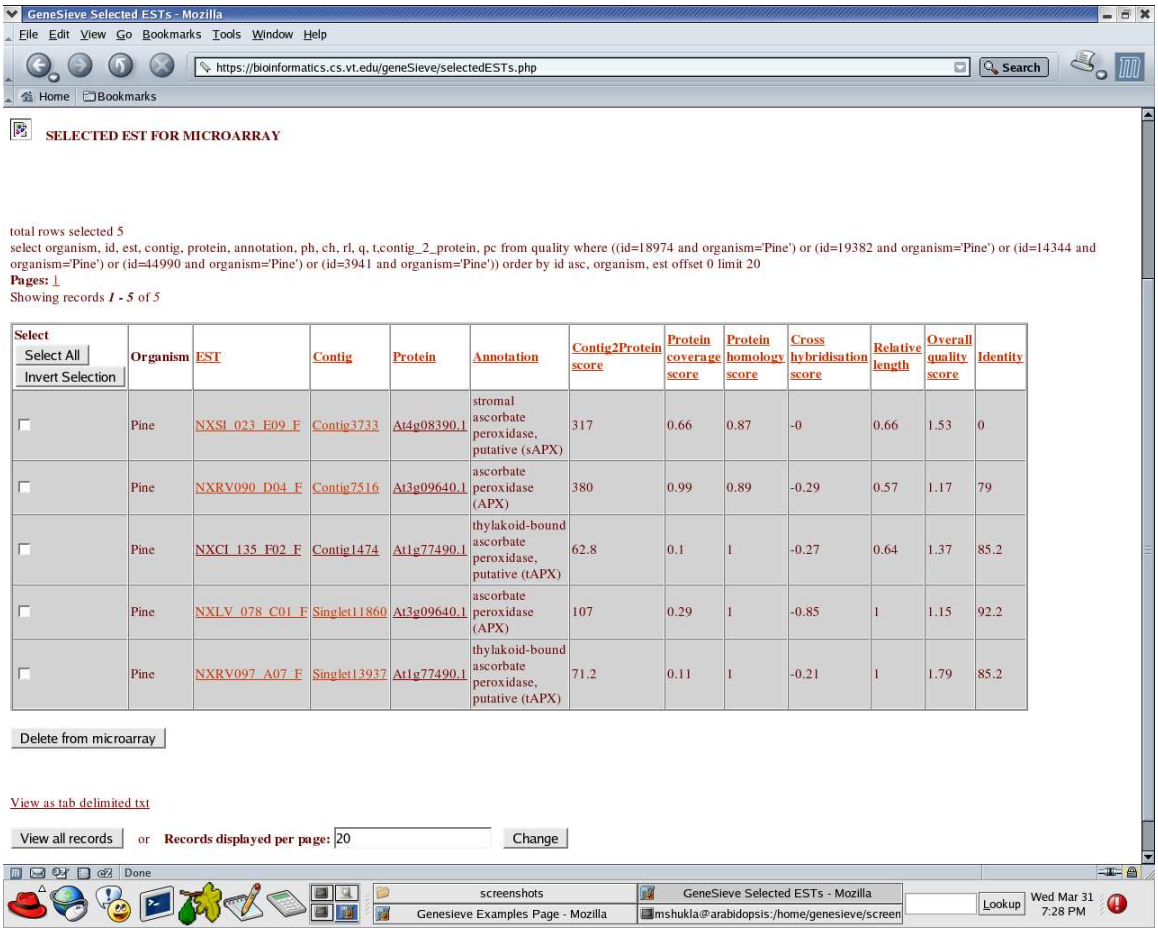

Figure 8.5: ESTs selected as microarray probes. The ESTs obtained using different search criteria can be added to the microarray. At the end of the session, a list selected ESTs can be downloaded as a text file.

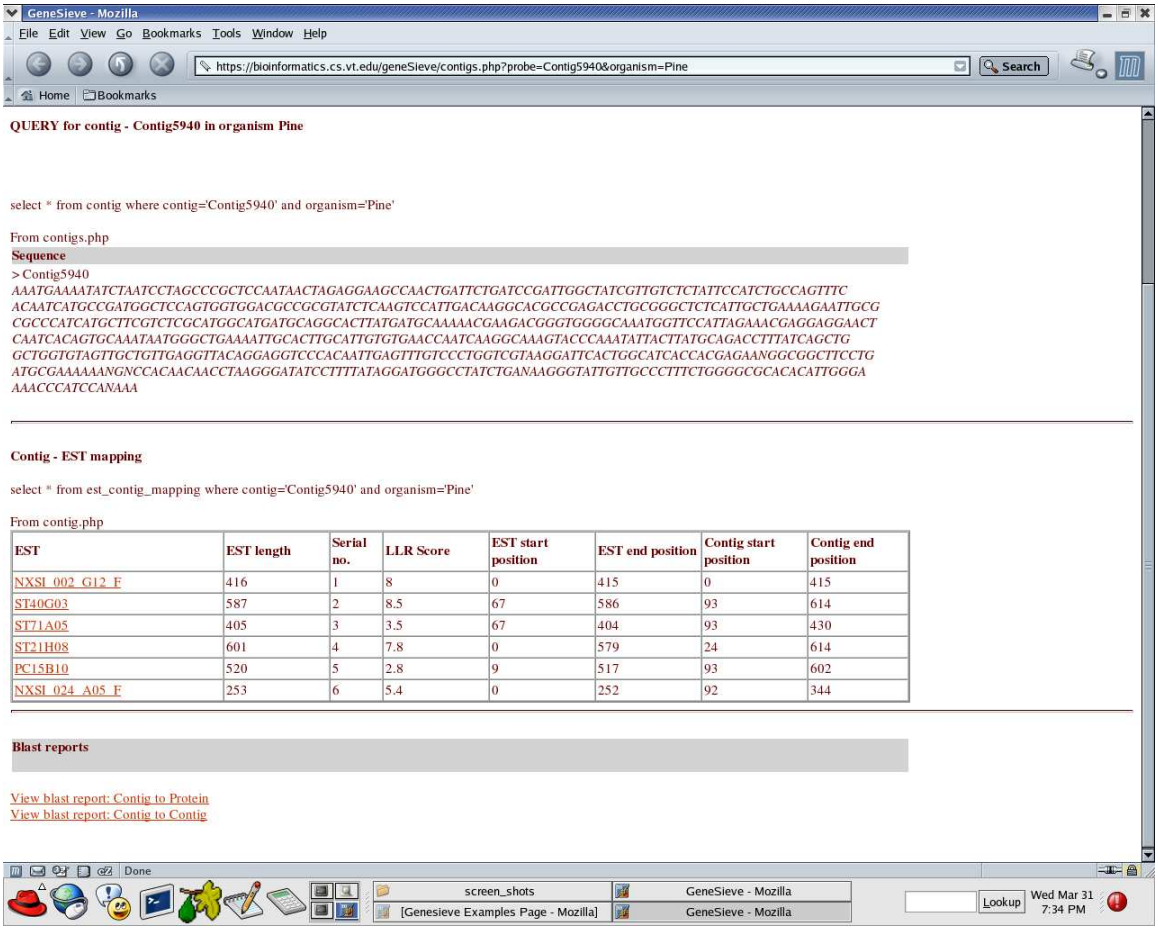

Figure 8.6: Details of pine contig5940. Sequence of the contig and a list of its constituent ESTs can be viewed by clicking on the contig id. BLAST reports for this contig against the contig and the protein databases can be viewed by following corresponding links provided at the bottom of the page.

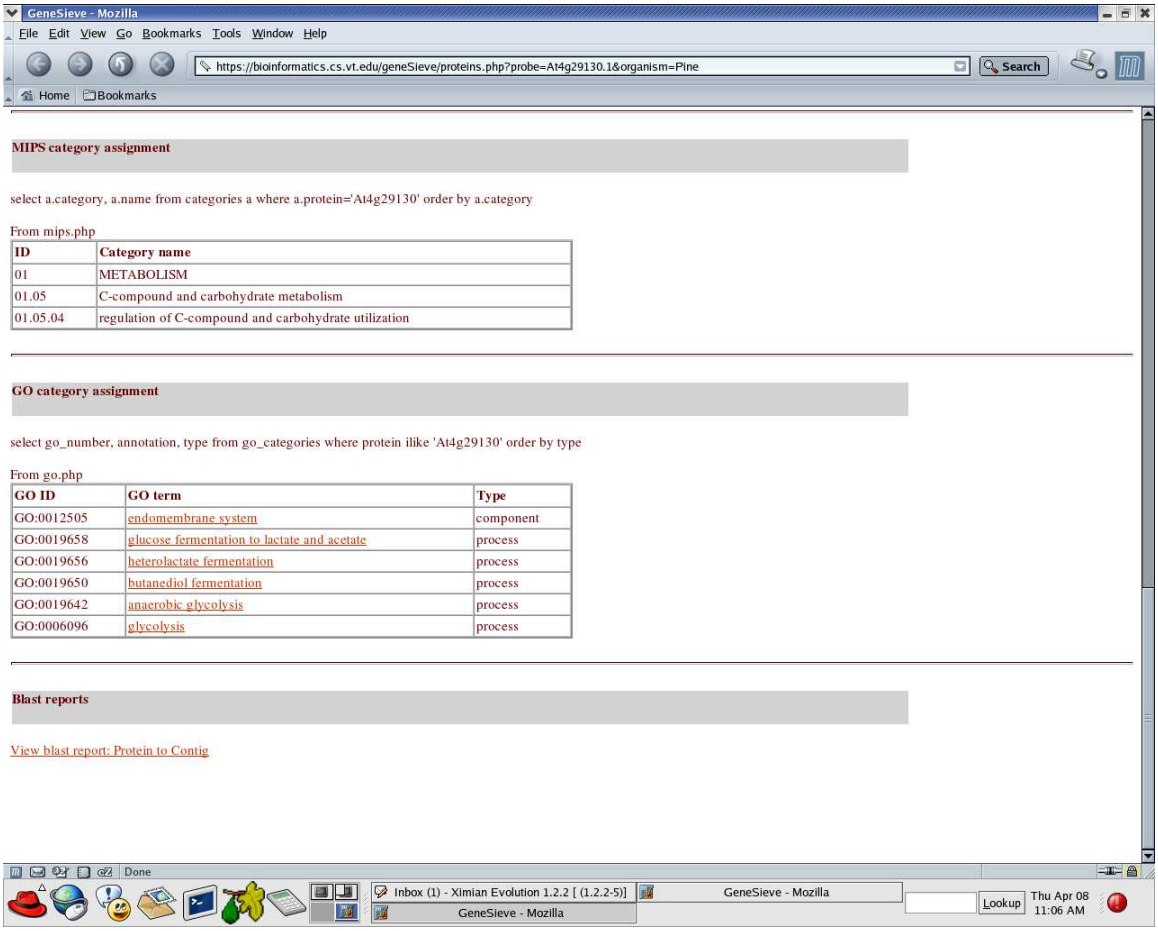

Figure 8.7: MIPS and GO functional categories for Arabidopsis protein At4g29130.

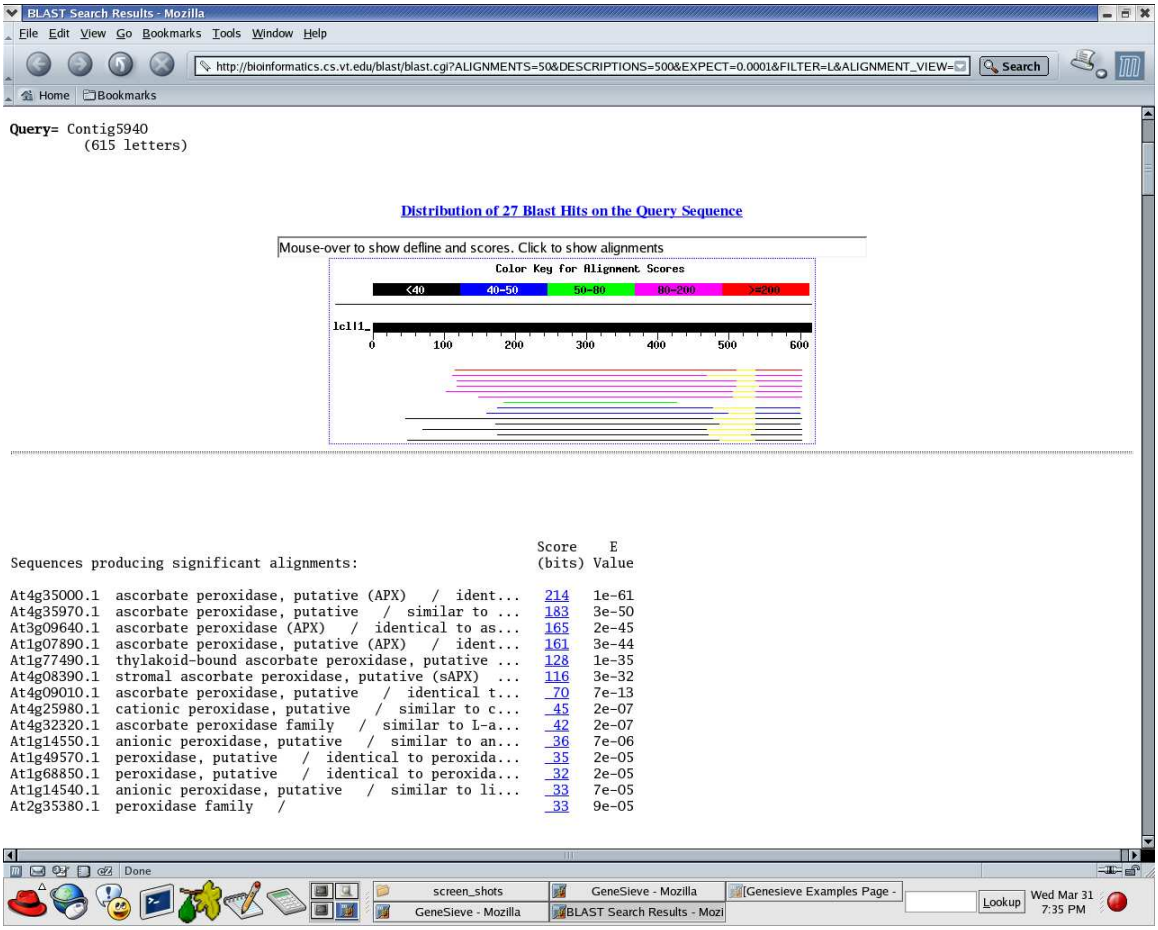

Figure 8.8: BLAST report for pine contig5940. The contig is BLASTed against the Arabidopsis protein database.

Chapter 9

Conclusion and Future Work

In recent years, cDNA microarrays have emerged as a powerful technique for the measurement of the expression levels of tens of thousands of genes simultaneously. Often, cDNA libraries representing expressed genes of an organism are available, along with expressed sequence tags (ESTs). ESTs are widely used as probes for cDNA microarrays. Custom microarrays containing only genes relevant to the experimental objectives are very useful. To build a custom microarray requires selection of ESTs based on their sequence.

It is important to assign unambiguous annotations to ESTs. To assign reliable annotations to ESTs from a given organism, we cluster them into contigs using phrap. The larger contig sequences are then used to search for similarity with the known proteins in the model organism such as *Arabidopsis thaliana*. An EST is assigned the same annotation as the contig it is part of. We have developed three different methods to assign annotations to contigs: bidirectional hits (BH), bidirectional best hits (BBH), and unidirectional best hits (UBH). We implemented these methods on pine and potato EST datasets, using Arabidopsis as the model organism for comparison. Results show that the BH method assigns a large number of proteins (18) to each contig, resulting in ambiguous annotations. The BBH method assigns at the most one protein to each contig and at the most one contig to each protein, but it covers only a small fraction of contigs (18% for pine and 23% for potato). The UBH method assigns at most one protein to each contig. Thus, it assigns unambiguous annotations to the contigs. It also covers a large fraction of contigs in an organism (45% for pine and 80% for potato). For this reason, we use the UBH method to assign annotations

to the contigs in GeneSieve.

Once a contig is selected based on its annotation, the next step is to select one of its constituent ESTs to deposit on a microarray. Typically, the longest EST or the EST closer to the $3'$ end of the contig is selected. This EST may not have the properties desirable in a microarray probe. We have devised a scoring system to evaluate the quality of an EST probe set. We assign a quality score (Q) to each EST based on its protein homology (PH), cross hybridization (CH), and relative length (RL). We assign quality scores to all of the pine and potato ESTs. We select a single EST from each contig by one of seven values: maximum length, $3'$ proximity, $5'$ proximity, maximum PH, minimum CH, maximum RL, and maximum Q. We show that the longest ESTs show higher cross hybridization with other contigs and show less protein homology. Thus, overall quality scores for such ESTs are low. On the other hand, ESTs selected by $3'$ proximity show less cross hybridization, but show poor protein homology. Also, our analysis reveal that only a small fraction of contigs align to the $3'$ end of a protein (11% for pine and 25% for potato). This suggests that most of the ESTs selected from the $3'$ end of the contigs actually come from the coding regions of the genes. ESTs selected by quality the score give considerably better values of all three quality parameters. They show higher protein homology, less cross hybridization, and greater length. For this reason, we recommend that EST probes be selected based on their quality scores.

Our methodology improves on NCBI UniGene as GeneSieve assembles ESTs into contigs and uses more reliable consensus sequences to assign them annotations. GeneSieve is similar to TIGR Gene Indices and Sputnik in assigning annotations to ESTs. However, GeneSieve also links ESTs to MIPS and GO functional categories. In addition, GeneSieve assigns quality scores to the ESTs to aid biologists in selecting EST probes for custom microarrays. To our knowledge, no other system provides this capability to select EST probes for custom microarrays.

We have designed a web interface for quick and easy selection of EST probes for microarrays. We have linked results obtained from the GeneSieve protocol to the sequence databases and well known functional categorization schemes such as MIPS and GO. This allows one to select EST probes based on their annotations or functional categories. Selection of EST probes by GeneSieve ensures representation of all genes of interest, less redundancy, sufficient specificity, and less cross hybridization. It also assigns unambiguous annotations and functional categories to the microarray probes.. GeneSieve can also be used to assign annotations and functional categories to the ESTs of interest. Recently, we have added other plant species, such as tomato, maize, barley, and rice to the

GeneSieve database. GeneSieve can also be used to find homologs of a gene in different species and selecting equivalent probes in multiple species.

At present *Arabidopsis thaliana* is the model organism for all the plant species included in GeneSieve. ESTs from pine, potato, tomato, rice, maize, and barley are compared with the Arabidopsis proteins to assign annotations. Once the completed rice transcriptome is available, we will add rice proteins to the GeneSieve. Rice will serve as the model organism for plant species of the Gramineae family, such as barley and maize. ESTs from these species will be compared to the rice proteins. We will also add other members of the Gramineae family, such as wheat, oat, sorghum, and sugarcane when sufficient number of ESTs become available for them. In the future, GeneSieve will be enhanced to be a flexible repository and analysis tool for biological sequences of additional kinds. BAC sequences and promoter sequences will be incorporated in GeneSieve to support promoter analysis.

To summarize, we have developed the GeneSieve methodology that assembles EST sequences from multiple target organisms into contigs; derives annotation for ESTs and contigs by comparing them with annotated sequences of the model organism; links ESTs and contigs with well known functional categories such as MIPS and GO; and evaluates ESTs according to various criteria and scores to select an EST to best represent a contig and hence a gene in a target organism. We have implemented the methodology with a system that maintains essential information in a relational database and that supports user-guided selection through a web interface. The user has access to protein annotation and functional categories (MIPS and GO) as means to initiate experiment-specific selection.

We hope GeneSieve will become a single stop for selecting EST probes for custom microarrays.

REFERENCES

[1] M D Adams, J M Kelley, J D Gocayne, M Dubnick, M H Polymeropoulos, H Xiao, C R Merril A Wu, B Olde, and R F Moreno. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252(5013):1651–6, June 1991.

[2] B Alberts, D Bray, A Johnson, J Lewis, M Raff, K Roberts, and P Walter. *Essential Cell Biology*. Garland, 1998.

[3] R G Alscher, B I Chevone, L S Heath, and N Ramakrishnan. Expresso – a problem solving environment for bioinformatics: Finding answers with microarray technology. *Proceedings of the High Performance Computing Symposium, Advanced Simulation Technologies Conference (HPC 2001)*, pages 64–9, 2001.

[4] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–10, October 1990.

[5] S F Altschul, T L Madden, A A Schaffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–402, September 1997.

[6] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–9, May 2000.

[7] P Ayoubi, X Jin, S Leite, X Liu, J Martajaja, A Abduraham, Q Wan, W Yan, E Misawa, and R A Prade. PipeOnline 2.0: automated EST processing and functional data sorting. *Nucleic Acids Research*, 30(21):4761–4769, 2002.

[8] T Barrett, C Cheadle, W B Wood, D Teichberg, D M Donovan, W J Freed, K G Becker, and M P Vawter. Assembly and use of a broadly applicable neural cDNA microarray. *Restorative Neurology and Neuroscience*, 18(2):127–35, 2001.

[9] N L Van Berkum and F C Holstege. DNA microarrays: raising the profile. *Current Opinion in Biotechnology*, 12(1):48–52, February 2001.

[10] M S Boguski, T M Lowe, and C M Tolstoshev. dbEST–database for "expressed sequence tags". *Nature Genetics*, 4(4):332–3, August 1993.

[11] A J Carlisle, V V Prabhu, A Elkahloun, J Hudson, J M Trent, W M Linehan, E D Williams, M R Emmert-Buck, L A Liotta, P J Munson, and D B Krizman. Development of a prostate cDNA microarray and statistical gene expression analysis package. *Molecular Carcinogenesis*, 28(1):12–22, May 2000.

[12] Center for Computational Genomics and Bioinformatics, University of Minnesota. http://pinetree.ccgb.umn.edu/.

[13] Y G Chen. Construction of a normalized cDNA library by mRNA-cDNA hybridization and subtraction. *Methods in Molecular Biology*, 221:33–40, 2003.

[14] G M Cooper. *The cell : A Molecular Approach*. ASM Press, 2000.

[15] Phrap Documentation. http://www.phrap.org/phredphrap/phrap.html.

[16] D J Duggan, M Bittner, Y Chen, P Meltzer, and J M Trent. Expression profiling using cDNA microarrays. *Nature Genetics*, 21(1 Suppl):10–4, January 1999.

[17] O Ermolaeva, M Rastogi, K D Pruitt, G D Schuler, M L Bittner ML, Y Chen, R Simon, P Meltzer, J M Trent, and M S Boguski. Data management and analysis for gene expression arrays. *Nature Genetics*, 20(1):19–23, September 1998.

[18] E M Evertsz, J Au-Young, M V Ruvolo, A C Lim, and M A Reynolds. Hybridization cross-reactivity within homologous gene families on glass cDNA microarrays. *Biotechniques*, 31(5):1182,4,6, November 2001.

[19] R D Fleischmann, M D Adams, O White, R A Clayton, E F Kirkness, A R, A R Kerlavage, C J Bult, J F Tomb, B A Dougherty, and J M Merrick. Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*. *Science*, 269(5223):496–512, July 1995.

[20] M Garcia-Hernandez, T Z Berardini, G Chen, D Crist, A Doyle, E Huala, E Knee, M Lambrecht, N Miller, L A Mueller, S Mundodi, R Reiser, S Y Rhee, R Scholl, J Tacklind, D C Weems, Y Wu, I Xu, D Yoo, J Yoon, and P Zhang. TAIR: a resource for integrated Arabidopsis data. *Functional and Integrative Genomics*, 2(6):239–53, November 2002.

[21] Gene Ontology Consortium. http://www.geneontology.org/.

[22] D Gordon, C Abajian, and P Green. Consed: a graphical tool for sequence finishing. *Genome Research*, 8(3):195–202, March 1998.

[23] A Griffiths, J Miller, D Suzuki, R Lewontin, and W Gelbart. *An Introduction to Genetic Analysis*. Freeman, 2000.

[24] C A Harrington, C Rosenow, and J Retief. Monitoring gene expression using DNA microarrays. *Current Opinion in Microbiology*, 3(3):285–91, June 2000.

[25] L S Heath, N Ramakrishnan, R R Sederoff, R W Whetten, B I Chevone, C A Struble, V Y Jouenne, D Chen, L M van Zyl, and R Grene. Studying the functional genomics of stress responses in loblolly pine using the Expresso microarray management system. *Comparative and Functional Genomics*, pages 226–243, 2002.

[26] S Hennig, D Groth, and H Lehrach. Automated Gene Ontology annotation for anonymous sequence data. *Nucleic Acids Research*, 31(13):3712–3715, 2003.

[27] A. Hotz-Wagenblatt, T. Hankeln, P. Ernst, K. H. Glatting, E. R. Schmidt, and S. Suhai. Estannotator: a tool for high throughput est annotation. *Nucleic Acids Research*, 31(13):3716–3719, 2003.

[28] X Huang and A Madan. CAP3: A DNA sequence assembly program. *Genome Research*, 9(9):868–877, 1999.

[29] T R Hughes and D D Shoemaker. DNA microarrays for expression profiling. *Current Opinion in Chemical Biology*, 5(1):21–5, February 2001.

[30] C V Jongeneel. Searching the expressed sequence tag (EST) databases: panning for genes. *Briefings in Bioinformatics*, 1(1):76–92, February 2000.

[31] M D Kane, T A Jatkoe, C R Stumpf, J Lu, J D Thomas, and S D Madore. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Research*, 28(22):4552–7, November 2000.

[32] T Kohchi, K Fujishige, and K Ohyama. Construction of an equalized cDNA library from *Arabidopsis thaliana*. *The Plant Journal*, 8(5):771–6, November 1995.

[33] G Lennon, C Auffray, M Polymeropoulos, and M B Soares. The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics. 1996 Apr 1;33(1):151-2*, 33(1):151–2, 1996.

[34] F Li and G D Stormo. Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, 17(11):1067–76, November 2001.

[35] D J Lockhart and E A Winzeler. Genomics, gene expression and DNA arrays. *Nature*, 405(6788):827–36, June 2000.

[36] S K Loftus, Y Chen, G Gooden, J F Ryan, G Birznieks, M Hilliard, A D Baxevanis, M Bittner, P Meltzer, J Trent, and W Pavan. Informatic selection of a neural crest-melanocyte cDNA set for microarray analysis. *Proceedings of the National Academy of Sciences of the USA*, 96(16):9277–80, August 1999.

[37] M G Lorenz, L M Cortes, J J Lorenz, and E T Liu. Strategy for the design of custom cDNA microarrays. *Biotechniques*, 34(6):1264–70, June 2003.

[38] H W Mewes, D Frishman, U Guldener, G Mannhaupt, K Mayer, M Mokrejs, B Morgenstern, M Munsterkotter, S Rudd, and B Weil. MIPS: a database for genomes and protein sequences. *Nucleic Acids Research*, 30(1):31–4, January 2002.

[39] R T Miller, A G Christoffels, C Gopalakrishnan, J Burke, A A Ptitsyn, T R Broveak, and W A Hide. A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Research*, 9(11):1143–55, November 1999.

[40] Munich Information Center for Protein Sequences. http://mips.gsf.de/.

[41] National Center for Biothechnology Information. http://www.ncbi.nlm.nih.gov/.

[42] P S Nelson, V Hawkins, M Schummer, R Bumgarner, W L Ng WL, T Ideker, C Ferguson, and L Hood. Negative selection: a method for obtaining low-abundance cDNAs using high-density cDNA clone arrays. *Genetic Analysis : Biomolecular Engineering*, 15(6):209–15, December 1999.

[43] H B Nielsen and S Knudsen. Avoiding cross hybridization by choosing nonredundant targets on cDNA arrays. *Bioinformatics*, 18(2):321–2, February 2002.

[44] S R Patanjali, S Parimoo, and S M Weissman. Construction of a uniform-abundance (normalized) cDNA library. *Proceedings of the National Academy of Sciences of the USA*, 88(5):1943–7, March 1991.

[45] W R Pearson. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods in Enzymology*, 183:63–98, 1990.

[46] W R Pearson, T Wood, Z Zhang, and W Miller. Comparison of DNA sequences with protein sequences. *Genomics*, 46(1):24–36, November 1997.

[47] W D Pennie. Custom cDNA microarrays; technologies and applications. *Toxicology*, 181-182:551–4, December 2002.

[48] J R Pollack, C M Perou, A A Alizadeh, M B Eisen, A Pergamenschikov, C F Williams, S S Jeffrey, D Botstein, and P O Brown. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*, 23(1):41–6, September 1999.

[49] J U Pontius, L Wagner, and G D Schuler. UniGene: a unified view of the transcriptome. *The NCBI Handbook*, Bethesda(MD):National Center for Biotechnology Information, 2003.

[50] J Quackenbush, J Cho, D Lee, F Liang, I Holt, S Karamycheva, B Parvizi, G Pertea, R Sultana, and J White. The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research*, 29(1):159–64, January 2001.

[51] J Quackenbush, F Liang, I Holt, G Pertea, and J Upton. The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Research*, 28(1):141–5, January 2000.

[52] S Y Rhee, W Beavis, T Z Berardini, G Chen, D Dixon, A Doyle, M Garcia-Hernandez, E Huala, G Lander, M Montoya, N Miller, L A Mueller, S Mundodi, L Reiser, J Tacklind, D C Weems, Y Wu, I Xu, D Yoo, J Yoon, and P Zhang. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Research*, 31(1):224–8, January 2003.

[53] J C Rockett, Luft J Christopher, G J Brian, S A Krawetz, M R Hughes, KIRN K Hee, A J Oudes, and D J Dix. Development of a 950-gene DNA array for examining gene expression patterns in mouse testis. *Genome Biology*, 2(4):RESEARCH0014, 2001.

[54] J M Rouillard, C J Herbert, and M Zuker. Oligoarray: genome-scale oligonucleotide design for microarrays. *Bioinformatics*, 18(3):486–7, March 2002.

[55] S Rudd, H W Mewes, and K F Mayer. Sputnik: a database platform for comparative plant genomics. *Nucleic Acids Research*, 31(1):128–132, 2003.

[56] H Schoof, P Zaccaria, H Gundlach, K Lemcke, S Rudd, G Kolesov, R Arnold, H W Mewes, and K F Mayer. MIPS *Arabidopsis thaliana* database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome. *Nucleic Acids Research*, 30(1):91–3, January 2002.

[57] G D Schuler. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *Journal of Molecular Medicine*, 75(10):694–8, October 1997.

[58] A Sioson, J I Watkinson, C Vasquez-Robinet, M Ellis, M Shukla, D Kumar, N Ramakrishnan, L S Heath, R Grene, B I Chevone, K Kadafar, and L T Watson. Expresso and chips: Creating a next generation microarray experiment management system. *Proceedings of the Next Generation Software Systems Workshop, 17th International Parallel and Distributed Processing Symposium (IPDPS'02)*, April 2003.

[59] E Southern, K Mir, and M Shchepinov. Molecular interactions on microarrays. *Nature Genetics*, 21(1 suppl):5–9, January 1999.

[60] E M Southern, S C Case-Green, J K Elder, M Johnson, K U Mir, L Wang, and J C Williams. Arrays of complementary oligonucleotides for analysing the hybridisation behaviour of nucleic acids. *Nucleic Acids Research*, 22(8):1368–73, April 1994.

[61] I Takemasa, H Higuchi, H Yamamoto, M Sekimoto, N Tomita, S Nakamori, R Matoba, M Monden, and K Matsubara. Construction of preferential cDNA microarray specialized for human colorectal carcinoma: molecular sketch of colorectal cancer. *Biochemical and Biophysical Research Communications*, 285(5):1244–9, August 2001.

[62] E Talla, F Tekaia, L Brino, and B Dujon. A novel design of whole-genome microarray probes for *Saccharomyces cerevisiae* which minimizes cross-hybridization. *BMC Genomics*, 4(1):38, September 2003.

[63] The Institute for Genomic Research. http://www.tigr.org/.

[64] S Tomiuk and K Hofmann. Microarray probe selection strategies. *Briefings in Bioinformatics*, 2(4):329–40, December 2001.

[65] P Wang, F Ding, H Chiang, R C Thompson, S J Watson, and F Meng. ProbeMatchDB — a web database for finding equivalent probes across microarray platforms and species. *Bioinformatics*, 18(3):488–9, March 2002.

[66] X Wang and B Seed. Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics*, 19(7):796–802, May 2003.

[67] Washington University BLAST. http://blast.wustl.edu/.

[68] D L Wheeler, D M Church, S Federhen, A E Lash, T L Madden, J U Pontius, G D Schuler, L M Schriml, E Sequeira, T A Tatusova, and L Wagner. Database Resources of the National Center for Biotechnology. *Nucleic Acids Research*, 31(1):28–33, January 2003.

[69] J D Wren, A Kulkarni, J Joslin, R A Butow, and H R Garner. Cross-hybridization on PCR-spotted microarrays. *IEEE Engineering in Medicine and Biology Magazine*, 21(2):71–5, April 2002.

[70] D Xu, G Li, L Wu, and J Zhou J Y Xu. Primegens: robust and efficient design of gene-specific probes for microarray analysis. *Bioinformatics*, 18(11):1432–7, November 2002.

[71] W Xu, S Bak, A Decker, S M Paquette, R Feyereisen, and D W Galbraith. Microarray-based analysis of gene expression in very large gene families: the cytochrome P450 gene superfamily of *Arabidopsis thaliana*. *Gene*, 272(1-2):61–74, July 2001.

VITA

Maulik Shukla was born on June 14, 1978 in Ahmedabad, India. He received his primary and secondary education from Diwan Ballubhai School, Ahmedabad. He received his Bachelor's Degree in Chemical Engineering from Nirma Institute of Technology, Ahmedabad, in June, 1999. He worked as a Systems Engineer in the Indian Petrochemicals Corporation Limited, India. In January 2001, he begun Masters in Computer Science at Virginia Tech. The work reported in this thesis was performed between May 2002 and May 2004.