ABSTRACT

AN APPLICATION IN BIOINFORMATICS: A COMPARISON OF AFFYMETRIX AND COMPUGEN HUMAN GENOME MICROARRAYS

The human genome microarrays from Compugen® and Affymetrix® were compared in the context of the emerging field of computational biology. The two premier database servers for genomic sequence data, the National Center for Biotechnology Information and the European Bioinformatics Institute, were described in detail. The various databases and data mining tools available through these data servers were also discussed. Microarrays were examined from a historical perspective and their main current applications—expression analysis, mutation analysis, and comparative genomic hybridization—were discussed. The two main types of microarrays, cDNA spotted microarrays and high-density spotted microarrays were analyzed by exploring the human genome microarray from Compugen® and the HG-U133 Set from Affymetrix® respectively. Array design issues, sequence collection and analysis, and probe selection processes for the two representative types of arrays were described. The respective chip design of the two types of microarrays was also analyzed. It was found that the human genome microarray from Compugen® contains probes that interrogate 1,119,840 bases corresponding to 18,664 genes, while the HG-U133 Set from Affymetrix® contains probes that interrogate only 825,000 bases corresponding to 33,000 genes. Based on this, the efficiency of the 25-mer probes of the HG-U133 Set from Affymetrix® compared to the 60-mer probes of the microarray from Compugen® was questioned.

AN APPLICATION IN BIOINFORMATICS: A COMPARISON OF AFFYMETRIX AND COMPUGEN HUMAN GENOME MICROARRAYS

by
Milind Misra

A Master's Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computational Biology

Federated Biological Sciences Department

May 2003

APPROVAL PAGE

AN APPLICATION IN BIOINFORMATICS: A COMPARISON OF AFFYMETRIX AND COMPUGEN HUMAN GENOME MICROARRAYS

Milind Misra

Dr. Michael L. Recce, Thesis Advisor — Date
Associate Professor of Information Systems, NJIT

Dr. Sanjay Malhotra, Committee Member — Date
Assistant Professor of Chemistry, NJIT

Dr. William J. Skawinski, Committee Member — Date
Research Professor of Chemistry, NJIT

BIOGRAPHICAL SKETCH

Author: Milind Misra

Degree: Master of Science in Computational Biology

Date: May 2003

Undergraduate and Graduate Education:

- Master of Science in Computational Biology,
  New Jersey Institute of Technology, Newark, NJ, 2003

- Master of Science in Applied Chemistry,
  New Jersey Institute of Technology, Newark, NJ, 1999

- Bachelor of Science in Engineering Science,
  New Jersey Institute of Technology, Newark, NJ, 1998

Major: Computational Biology

Presentations and Publications:

D. Pandit, M. Misra, K. M. Gilbert, and C. A. Venanzi, "Analysis of Conformational Families of Analogs of GBR 12909", 15th Annual Molecular Biophysics Symposium, Rutgers University and UMDNJ, Piscataway, NJ, May, 2003.

M. Misra, K. M. Gilbert, R. A. Buono, M. M. Schweri, Q. Shi, H. M. Deutsch, and C. A. Venanzi, "Preliminary Pharmacophore Model for the Methylphenidate Class of Dopamine Reuptake Inhibitors", National Meeting of the American Chemical Society, New Orleans, LA, March, 2003.

K. M. Gilbert, M. Misra, R. A. Buono, C. A. Venanzi, M. M. Schweri, Q. Shi, and H. M. Deutsch, "Comparative Molecular Field Analysis of Methylphenidate Analogues", 17th Chemistry Conference, Santiago de Cuba, Cuba, December, 2002.

K. M. Gilbert, M. Misra, R. A. Buono, C. A. Venanzi, M. M. Schweri, Q. Shi, and H. M. Deutsch, "CoMFA Study of Protonated Methylphenidate Phenyl-Substituted Analogues", National Meeting of the American Chemical Society, Boston, MA, August, 2002.

To Sudha, my mother, whose gift has been Belief

and

To Santosh, my father, from whom I learnt Skepticism

ACKNOWLEDGMENT

I would like to thank Michael Recce, my advisor, for providing the initial spark that helped me get over my fear of learning new and challenging topics in the vast field of Computational Biology. His enthusiastic optimism was frequently contagious and his ability to see a positive outcome from tough situations has been a good lesson in management. I also thank Dr. Recce for making the Computational Biology Program a successful reality at NJIT.

I am grateful to the Federated Biological Sciences Department and the Office of Graduate Studies, NJIT, for providing financial assistance to me in the last semester of this program.

Special thanks to Dr. Sanjay Malhotra and Dr. William Skawinski for agreeing to be members of my thesis committee and providing important suggestions.

I wish to thank Karen Gansner, the Computational Biology Program Coordinator, for ensuring as smooth a functioning of the fledgling program as was possible.

Finally, I acknowledge the support of a number of individuals over the course of this master's degree. They include Dr. Carol Venanzi for providing encouragement and for letting me use her laboratory facilities for my thesis and class work; Dr. David Schoenhaut, my first employer in industry; Morty Kwestel for input toward the general content of this thesis; and various friends who proved to be stimulating colleagues.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

CHAPTER 1

INTRODUCTION

1.1 Perspective: Changing Scope of Modern Life Sciences

Claude Shannon, the mathematician and an architect of classical information theory, defined information as that which reduces uncertainty. Gregory Bateson, the social scientist who sought to re-introduce 'the mind' back into scientific equations, defined information as that which changes us. The two men gave us two schools of thought that wrestle with the same reality. Modern molecular biology, with its main focus on the flow of information in a cell, can perhaps offer a unique arena for reconciling the two views. Our current view of genetic information transfer in cells is set forth by the so-called 'central dogma' of molecular biology (Figure 1.1). Adapted suitably, the genetic flow of information can be seen from the point of view of classical information theory (Figure 1.2[1]) in which the polynucleotidic sequence of DNA is the source or carrier of genetic information, various forms of RNA are the transmitter and receiver of this information, mutations correspond to noise introduction, and the polypeptidic sequence of proteins is the destination. In this model, Shannon's uncertainty would be related to the confidence

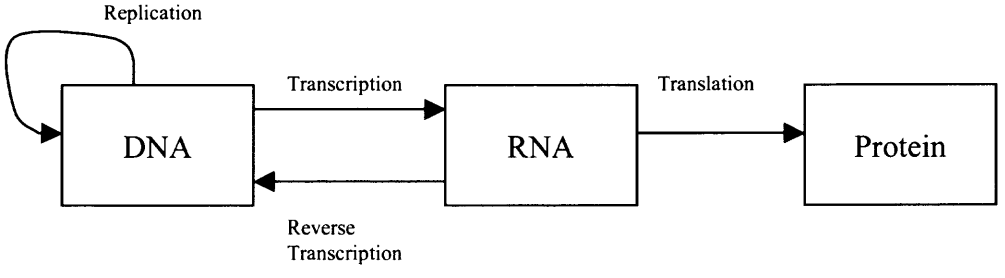

Figure 1.1 Flow of genetic information in a cell

level with which one could say that a given nucleotide sequence would correspond to a particular protein, i.e., uncertainty would be related to gene expression.

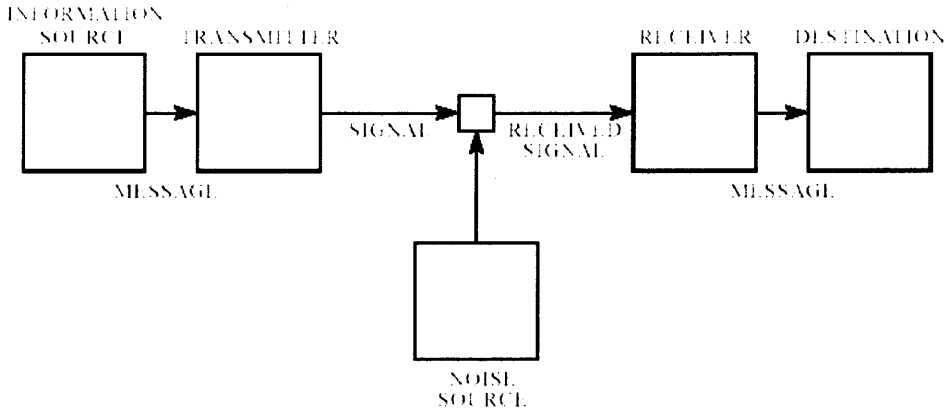

Figure 1.2 Shannon's diagram of a general communication system [1]

In contrast, the application of Bateson's definition of information to genetic information is relatively straightforward: The genetic information encoded in our DNA represents our evolutionary inheritance and controls our growth, reproduction, and self-repair, the three primary activities that represent change.

The preceding discussion is an example of the evolving and ever broadening scope of modern life sciences. The relatively young field of computational biology, in particular, is a unique mix of various shades of biology and computer science. Indeed, the need for the field's development and the demand for qualified professionals in it has led to the establishment in the last several years of many organizations that seek to give a direction to the fledgling field. For example, the International Society for Computational Biology (ISCB)[2] was set up in 1997 and is a fast growing professional organization that

aims to further the advancement of scientific understanding of living systems through computation.

The availability of massive amounts of sequence data has also led to the development of various methods to extract useful information from this data. The last decade saw the rapid development of one such method, microarray technology, and its increased use as a diagnostic and analysis tool at the molecular level. Several business corporations have introduced their own microarrays that frequently contain whole-genome representations. The purpose of this thesis is to explore some of the differences between the human genome microarrays introduced by Affymetrix® and Compugen®.

1.2 Outline

The next chapter provides a background for the magnitude of biological data that has been the reason for the development of specialized databases. It describes, in some detail, two of the most important database servers that are used by scientists and researchers from all over the world and briefly mentions some related databases. It also illustrates some of the sequencing techniques that are used for collecting the genomic information that digitally populates these databases.

Chapter 3 introduces microarrays in the context of their Southern blot ancestry. It lists some key and popular applications of microarrays and briefly describes three of these applications: comparative genomic hybridization, expression analysis, and mutation analysis. It also describes the two main types of microarrays, the spotted arrays and the high-density oligonucleotide arrays and introduces microarrays from Compugen® and Affymetrix® as the representative arrays of the two types.

Chapter 4 describes and performs a comparison of the human genome microarrays from Affymetrix® (the HG-U133 Set) and Compugen®. Finally, Chapter 5 provides the conclusions of this research.

CHAPTER 2

DATA AND DATABASES

2.1 Dealing with Data Explosion: Data Management and Dissemination

The growth of organizations like the ISCB around the world reflects the explosive growth in biological data that has emerged from laboratories worldwide over the last couple of decades. Figure 2.1 illustrates the exponential growth during 1982-2002 in the number of sequences and base pairs of DNA submitted to the GenBank database. According to the latest release notes of GenBank (February 15, 2003) the database contains over 29 billion

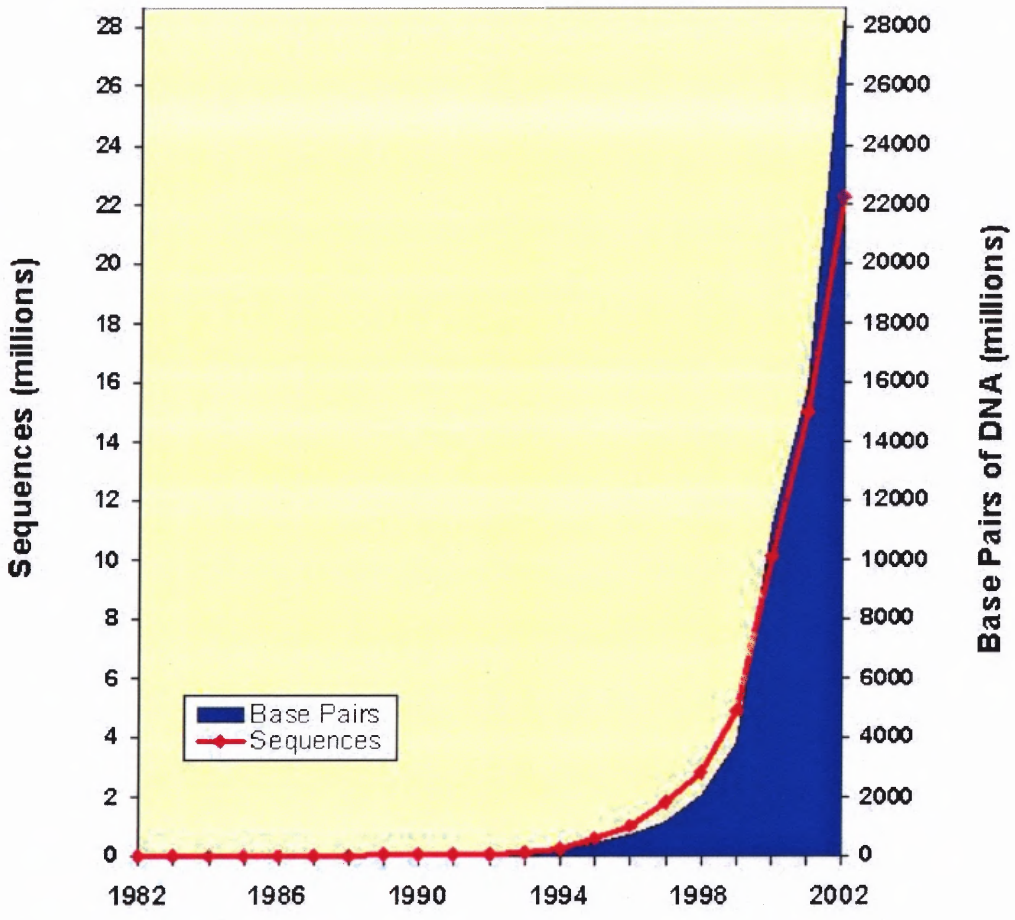

Figure 2.1 Growth of GenBank, 1982-2002 [3]

base pairs of DNA from over 23 million reported sequences.[3] Figure 2.2 shows a similar graph for the number of annotated protein sequences submitted to the Protein Data Bank database (PDB). As of April 14, 2003, the PDB contained over 20,400 protein structures.[4]

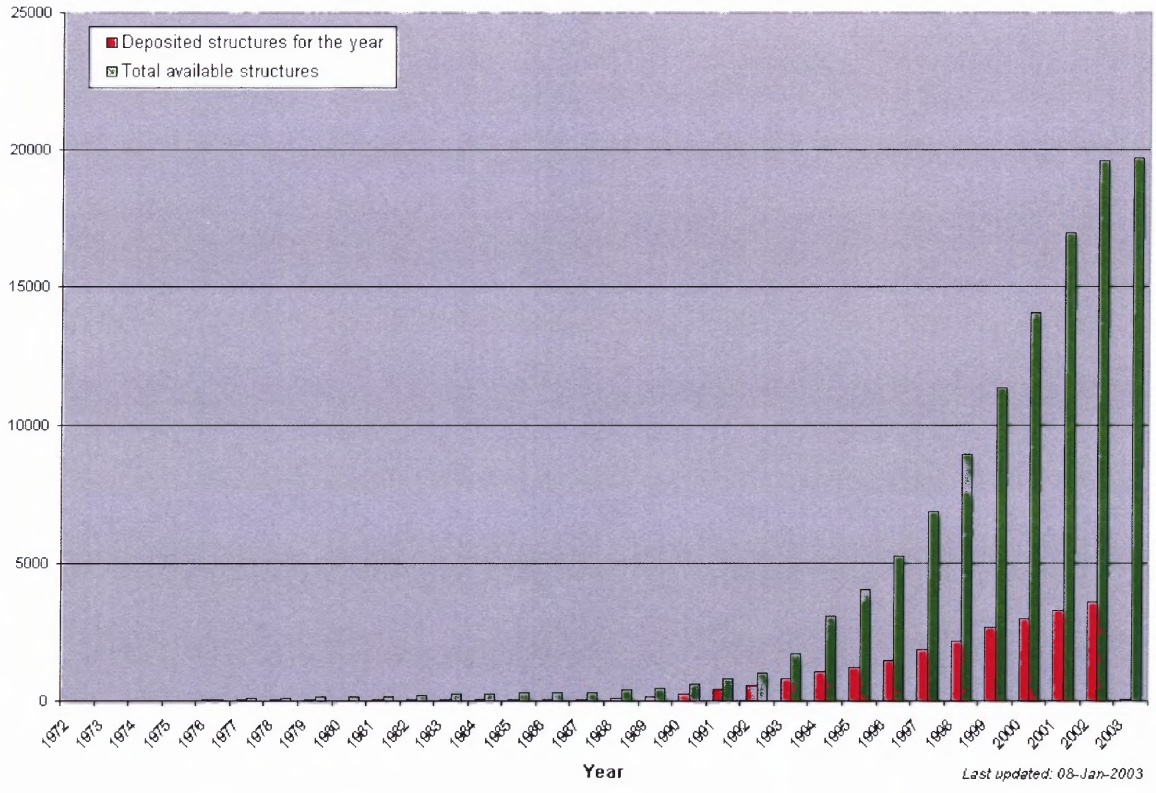

Figure 2.2 Growth of PDB, 1972-2003 [4]

One explanation for the staggering growth of biological data can be found in the motivation for the pharmaceutical industry to screen progressively larger number of drug candidates. Rising costs of research and development, of introducing a new drug into the market, and of lengthy clinical trials have ensured that the drug companies face elevated competition for a share of the market. This competition has found expression in the need for researchers to investigate the maximum possible number of potentially useful drug

candidates. In addition, since the resources available to researchers are limited, there is also the related need to identify dead-ends faster. To achieve a measure of success in such a scenario, researchers have resorted to applying a host of new technologies like bioinformatics, cheminformatics, proteomics, genomics, and pharmacogenomics. These new fields are varying combinations of biological sciences, computer science, and information technology. Together with novel advanced algorithms, high-end computers, and automated processes and systems, these young technologies permit scientists to mass-produce huge amounts of gene sequence, protein structure, and small molecule data for the purpose of analysis and comparison.

The new approach promises to enable greater productivity in drug-design and drug-discovery. However, the deluge of genetic information has also meant the setting up of new infrastructure that is capable of handling the quantum of data that is involved. In particular, it has necessitated a) specialized computer databases to store, organize, and index the data, and b) specialized tools to access, view, and analyze the data. Issues like size and number of databases, complex database queries, efficient data mining, higher throughput, data visualization, and systems integration have taken center stage with the goal of bringing together diverse sets of researchers who might be distributed through diverse geographic locations. In the initial stages of the genomics revolution, the focus was on the creation and maintenance of a database to store biological information such as nucleotide and amino acid sequences. This involved both complicated design issues and also the development of a suitable interface permitting easy submission and revision of sequence data. With the sequencing of particularly the human genome nearing completion, however, the spotlight is shifting to analysis of the substantial already-

available data. Researchers can use sequence data to, for example, first obtain a 'normal' state of cellular affairs and then, by comparison with an 'abnormal' (or diseased) state, attempt to hypothesize reasons for the difference between the two states. Thus, increasing emphasis is now being laid on analyzing and interpreting nucleotide sequences, protein domains, and protein structures. Computational biology is the field that has come to be identified with this process of data analysis and interpretation. According to the National Center for Biotechnology Information (NCBI), the related fields of bioinformatics and computational biology include:

- The development and implementation of tools that enable efficient access to, and use and management of, various types of information; and
- The development of new algorithms (mathematical formulas) and statistics with which to assess relationships among members of large data sets, such as methods to locate a gene within a sequence, predict protein structure and/or function, and cluster protein sequences into families of related sequences.[5]

The emergence of the Worldwide Web as the optimal medium for information transmission over large distances was a development that coincided with the needs of scientists working in genomics. The geographic spread of scientists working on and of those interested in using sequence data made imperative an Internet based approach to data dissemination. Indeed, the alternative--developing similar solutions to similar problems at different locations--would not only be counterproductive to a synergistic research effort but also nearly impossible. The earliest endeavors in genomics tried a similar approach but recognized the utility of universal collaborations and public involvement when they realized the sheer magnitude of the data that they faced. The

sequencing of small genomes such as those of viruses and bacteria (typically 1-10 kbp) and smallest eukaryotes (typically 10^3 kbp) has been relatively fast given their small size. In 1996, yeast, with a genome size of 10^3, became the first free-living eukaryote was completely sequenced marking an important milestone in genomics. In comparison, the Human Genome Project, a multinational effort to map the human genome, began in 1988 and is expected to be completed some time in 2003. A large effort such as this would have been extremely difficult for a single laboratory to handle and a consortium of laboratories working together at the problem is a more efficient way of chipping away at the monumental task. The coordination of such an effort involves frequently updatable databases, consistent information from all the laboratories involved, and fast access to data. The possibility of errors in sequenced data is also a major concern with the current target being obtaining at most one error per 10,000 bases. Given the complexities involved in maintaining data accuracy when accessed by scientists worldwide, much importance has also been attached to proper data transmission over the Internet.

The ready availability of sequence data has tremendous implications for the way the genetic flow of information within a cell would be interpreted in future. However, the detection and management of obscure patterns within such growing biological data necessitates suitable computational tools and databases. The next section will illustrate some of the more important public domain databases available to researchers around the world. Such repositories of biological data have the potential to facilitate the detection of novel trends and new principles that could eventually explain some of the complexities of biological systems. In addition, significant levels of collaboration exist among the major international databases, permitting easier access to biological data.

2.2 Important Biological Databases

Two main life sciences servers that manage and disseminate biological data are the National Center for Biotechnology Information (NCBI) in the U.S. and the European Bioinformatics Institute (EBI) in England. Both provide reliable access to data and also analytical database mining tools. Besides these two servers there are numerous other locations around the world that maintain mirror sites for data and analysis software.

2.2.1 NCBI [6]

The NCBI was established in 1988 after the U.S. senate recognized the need for computerized data processing in the life sciences and passed legislation that lead to the formation of NCBI as a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH), the world's largest biomedical research facility. At the time, NLM already had significant experience in the creation and management of biomedical databases. While NLM continues to focus on maintaining biomedical databases, NCBI is involved in developing novel analytical tools to assist in the understanding of processes at the molecular and genetic levels that play a key role in disease.

Figure 2.3[7] shows the major components of NCBI. The NCBI's diverse responsibilities include:

- conducting research on basic biomedical problems at the molecular level using mathematical and computational methods;
- developing automated systems for biological data analysis and storage;
- developing and promoting standards for databases and biological nomenclature;

- fostering usage of database and analytical software among the scientific community; and
- coordinating worldwide efforts to collect biological data.[8]

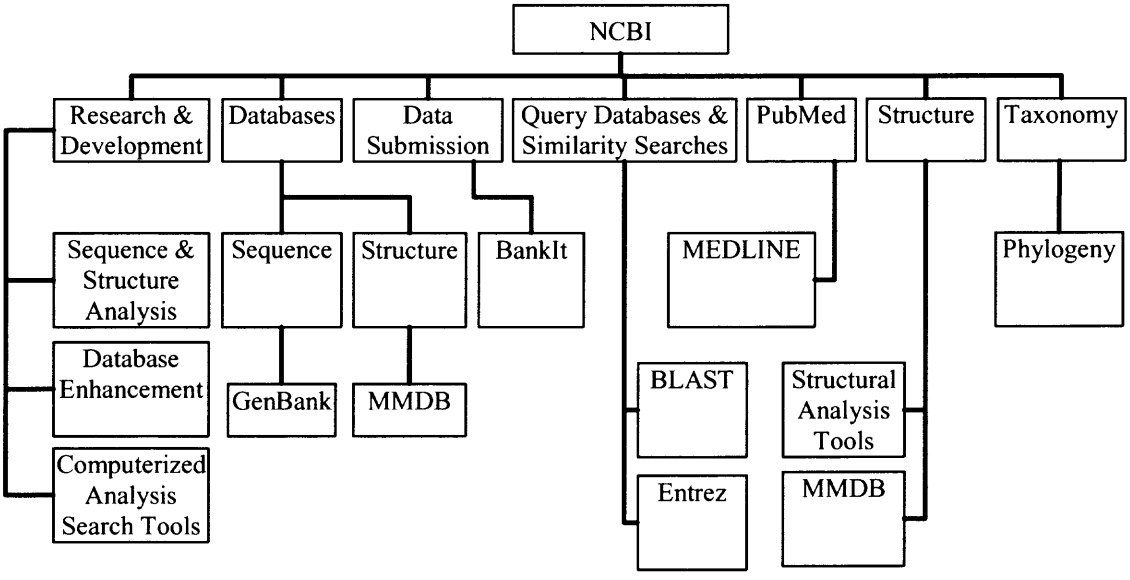

Figure 2.3 Synopsis of the NCBI

Computer scientists, molecular biologists, biochemists, structural biologists, mathematicians, and research physicians work with NCBI on basic and applied research in computational molecular biology. Some of the problems they work on include gene organization, sequence analysis, and structure prediction. Research projects at NCBI include detection and analysis of gene organization, repeating sequence patterns, protein domains and structural elements, creation of a gene map of the human genome, mathematical modeling of the kinetics of HIV infection, analysis of effects of sequencing errors for database searching, development of new algorithms for database searching and

multiple sequence alignment, construction of non-redundant sequence databases, mathematical models for estimation of statistical significance of sequence similarity, and vector models for text retrieval.

Since its inception, NCBI has developed into a mammoth storehouse for sequence and associated data. Figure 2.4 illustrates the various databases and tools that can be accessed through NCBI. The NCBI's literature databases (Figure 2.4a) form an extended searchable library of the life sciences literature. PubMed provides access to more than 12 million citations from MEDLINE and other life sciences journals and is a service of the NLM. It also has links to many other sites that provide full text articles and other related resources. PubMed Central (PMC) is NCBI's initiative to preserve and maintain open access to the electronic literature. It is a digital archive of life sciences journal literature developed and managed by NCBI. The NCBI collaborates with authors and publishers and adapts textbooks and monographs for the Web and links them to PubMed, the biomedical bibliographic database, and to other resources links. This effort has resulted in Bookshelf, which provides background information to users of PubMed. The Online Mendelian Inheritance in Man (OMIM) database catalogs human genes and genetic disorders and contains textual information and references in addition to links to MEDLINE and sequence records in the Entrez system. Proteins Reviews on the Web (PROW) features authoritative short, structured reviews on proteins and protein families. Together, these reviews provide nearly 20 standardized categories of information, such as abstract, biochemical function, ligands, references, etc., for each protein.

The Entrez system (Figure 2.4b) is a retrieval system that is designed for searching several linked databases. It also provides graphical views of sequences and

chromosome maps. A powerful feature of Entrez is its ability to retrieve related sequences, structures, and references. The protein sequence database is a collection of protein sequence entries that have been assembled from various different sources such as Swiss-Prot, PIR, PRF, PDB, as well as translations from annotated coding regions in GenBank and RefSeq. Similarly, the nucleotide sequence database provides access to nucleotide sequences compiled from various sources such as GenBank, RefSeq, and PDB. Entrez Genome database helps locate information for whole genomes of more than 1000 viruses and over 100 microbes. The Molecular Modeling Database (MMDB) is NCBI's structure database and permits access to empirical 3D biomolecular structures that have been obtained from X-ray crystallography and NMR-spectroscopy.

The taxonomy database contains the names of all organisms that are represented in the genetic databases and have at least one nucleotide or protein sequence. Population study data sets are DNA sequences that were collected for the analysis of the evolutionary relatedness of a population. ProbeSet links to NCBI's Gene Expression Omnibus (GEO) gene expression and hybridization array repository. It is intended to enhance searching on the GEO database and link the search results to internal and external resources. The 3D Domains database contains protein domains from the Entrez Structure database. UniSTS is an NCBI resource that reports information about markers, or Sequence Tagged Sites (STS), primer sequences, product size, and maps. SNP is a repository for both single nucleotide substitutions (SNPs) and short insertion and deletion polymorphisms. The Conserved Domain Database (CDD) is a collection of sequence alignments and profiles that represent protein domains that were conserved during molecular evolution. The Journals database permits searches for a journal and also

provide links to records for that journal in the database. UniGene is an experimental system for dividing GenBank sequences into a non-redundant set of gene-oriented clusters. These clusters represent unique genes and contain related information such as tissue types and map location.

The NCBI's sequence databases (Figure 2.4c) receive sequence data from worldwide sequencing projects. Sequences submitted by individual laboratories worldwide and by exchange of data with two other international databases, the European Molecular Biology Laboratory (EMBL) and the DNA Database of Japan (DDBJ), are incorporated by NCBI into the GenBank DNA sequence database. GenBank is an annotated collection of all publicly available nucleotide and amino acid sequences. dbEST is a division of GenBank containing sequence data and other information on Expressed Sequence Tags (ESTs), or short, single-pass genomic sequences, from many organisms. The genome survey sequences (GSS) database is similar to the dbEST except that most of its sequences are genomic in origin instead of being cDNA based. HomoloGene is a gene homology tool that compares nucleotide sequences between pairs of organisms so that potential orthologs can be identified. The High Throughput Genomic (HTG) Sequences database is a collection of finished and unfinished high throughput genome sequences from the sequence databases of NCBI, EMBL, and DDBJ. The SNPs database is a central repository for single-base nucleotide substitutions and short deletion and insertion polymorphisms. RefSeq is an important database of non-redundant reference sequences standards, which include genomic DNA contigs, mRNAs, and proteins for known genes. Data collection is enhanced by multiple collaborations within NCBI and with outside laboratories. dbSTS is an NCBI database containing

sequence and mapping data about short genomic landmark sequences (STS) that are operationally unique in the genome.

The NCBI provides several data retrieval and submission tools through its website. These data mining tools are: Text Term Searching, Sequence Similarity Searching, Taxonomy, and Sequence Submission (Figure 2.4d). Text Term Searching is done through Entrez, LinkOut, Cubby, and Citation Matcher. Entrez provides integrated access to nucleotide and protein sequence data from over 100,000 organisms, along with 3D protein structures, genomic mapping information, and PubMed MEDLINE. LinkOut is a registry service for the creation of links from specific articles, journals, or biological data in Entrez to resources on external websites including full-text publications, biological databases, consumer health information, and research tools. Cubby permits Entrez users to store and update searches using a Stored Search feature. It also allows customization of LinkOut display to include or exclude links to providers. Given the article's bibliographic information, Citation Matcher provides the ability to find the PubMed ID or the MEDLINE UID of any article in the PubMed database.

Sequence similarity searching can be performed by using one of the many variants of the popular program BLAST (Basic Local Alignment Search Tool). BLink displays the results of BLAST searches that have been done for every protein sequence in the Entrez Protein database. The output could include the positions of up to 200 BLAST hits on the query sequence, scores, and alignments. Blastcl3 is a BLAST network client that can access the NCBI BLAST search engine. It can search all the sequences in a FASTA file and produce one-to-many alignments in text or HTML format. It can also

carry out searches against multiple databases. For faster searching, a stand-alone version of BLAST can be downloaded to a local computer for local use.

The Taxonomy Browser allows searching of the NCBI's Taxonomy database. Taxonomy BLAST can use the classification of source organisms by NCBI's Taxonomy database to group BLAST hits. A summary of BLAST taxonomy data can be obtained with the help of the TaxTable utility, which can also display the relationship of the organism to others through a color-coded graph. ProtTable provides a summary of protein coding regions in a genome. TaxPlot provides a three-way view of genome similarities.

Sequence submission capability is provided by two data submission tools. Sequin is a data submission tool that includes ORF (open reading frame) Finder, an alignment viewer and editor, and a link to PowerBLAST. BankIt is an Internet submission tool for one or simple sequence submissions.

The NCBI provides access to various sequence analysis tools (Figure 2.4e). Clusters of Orthologous Groups (COGs) are a system of gene families derived from complete genomes. They were delineated by comparing protein sequences encoded in 43 complete genomes, representing 30 major phylogenetic lineages. COGnitor is a program used to compare a user sequence to the COGs database to identify the cluster of orthologous groups to which it might belong. Gene Expression Omnibus (GEO) is a gene expression data repository and online resource for that retrieval of gene expression data from any organism or artificial source. HomoloGene compares nucleotide sequences between pairs of organisms to identify putative orthologs. The Conserved Domain Database (CDD) is a collection of sequence alignments and profiles representing

protein domains conserved in molecular evolution. LocusLink provides a single-query interface to curated sequences and descriptive information about genetic loci by searching a browsable list that includes items such as gene names, descriptive terms, and LocusID numbers. The Mammalian Gene Collection (MGC) is a recent effort by the National Institutes of Health (NIH) to generate full-length complementary DNA (cDNA) resources. Clone Registry is a database used mainly by participating human and mouse genome sequencing centers to record those clones selected for sequencing, clones currently in sequencing pipeline, and clones that are finished and represented by sequence entries in GenBank. The raw sequence data underlying sequences generated by various genome projects can be stored using Trace Archive. The ORF Finder is a graphical analysis tool that finds all open reading frames of a selected minimum size in a user's sequence or in a sequence already in the database. VecScreen is a tool for identifying segments of a nucleic acid sequence that may be of vector, linker, or adapter origin before using Tools for Sequence Analysis or submission. Electronic-PCR (e-PCR) can be used to compare a query sequence to mapped STSs to find a possible map location or the query sequence.

Various genetic and physical maps can be accessed through NCBI's several map options (Figure 2.4f). Map Viewer provides comprehensive views of chromosome maps for 17 organisms. It displays one or more maps that have been aligned to each other based on shared marker and gene names and, for the sequence maps, based on a common sequence coordinate system. GeneMap'99 is a physical map of over 30,000 human gene-based markers. It was constructed by the International Radiation Hybrid Mapping Consortium and provides a glimpse of some of the most important parts of the genome.

Model Maker permits a user to build an mRNA sequence from genomic data, select exons identified by alignments of mRNAs and ESTs, edit the model, test open reading frames, and save results. The OMIM Gene Map contains the cytogenetic locations of genes that have been reported in the literature and determined by different mapping methods. The OMIM Morbid Map presents an alphabetical listing of diseases and their corresponding cytogenetic map locations, with links to OMIM entries. The Human-Mouse Homology maps contain a table comparing genes in homologous segments of DNA from human and mouse, sorted by position in each genome. Additionally, several maps can be accessed for genomes of organisms such as thale cress, fruit fly, human, malaria, mosquito, mouse, nematode, rat, and zebrafish.

The NCBI presents several tools for 3D structure display and similarity searching (Figure 2.4g). The Conserved Domain Search Service (CD-Search) can be used to identify the conserved domains present in a protein sequence. The Cn3D utility is a web browser plug-in and a 3D structure and sequence alignment viewer for NCBI structure databases. It runs on Windows, Macintosh, and UNIX operating systems and also has powerful annotation and alignment editing features. The Conserved Domain Architecture Retrieval Tool (CDART) displays the functional domains that make up a protein and lists proteins with similar domain architectures. The VAST Search is a structure-structure similarity search service that compares 3D coordinates of a newly determined protein structure to those already present in the MMDB or PDB databases. The NCBI threading is ongoing work on algorithms for protein fold recognition.

The collaboration of NCBI and the National Cancer Institute (NCI) has resulted in a number of projects (Figure 2.4h). The Spectral Karyotyping (SKY) and Comparative

Genomic Hybridization (CGH) database is a repository of publicly submitted SKY and CGH data. SKY and CGH are complementary fluorescent molecular cytogenetic techniques. The Cancer Chromosome Aberration Project (CCAP) hopes to speed up the definition and detailed characterization of the chromosomal alterations that are associated with malignant transformation. The Cancer Genome Anatomy Project (CGAP) is an interdisciplinary program to identify human genes expressed in different cancerous states. The Mitelman Database of Chromosome Aberrations in Cancer is a genome-wide map of chromosomal breakpoints in human cancer. Serial Analysis of Gene Expression (SAGE) is an experimental technique for quantitatively measuring gene expression. The SAGEmap website provides a differential analysis of CGAP SAGE libraries. It also includes a comprehensive analysis of SAGE tags in human GenBank records, in which a UniGene identifier is assigned to each human sequence that contains a SAGE tag.

The NCBI also lists several regularly updated statistics on its web pages (Figure 2.4i). GenBank growth statistics and human genome sequencing progress keep track of the number of sequences in GenBank and the sequencing of the human genome respectively. LocusLink statistics summarize gene records for fruit fly, human, mouse, rat, and zebrafish. PubMed, RefSeq, and Taxonomy statistics also show regular updates.

The NCBI provides access to the whole genomes of over 1,000 organisms. These genomes represent both completely sequenced organisms and those for which sequencing is in progress.

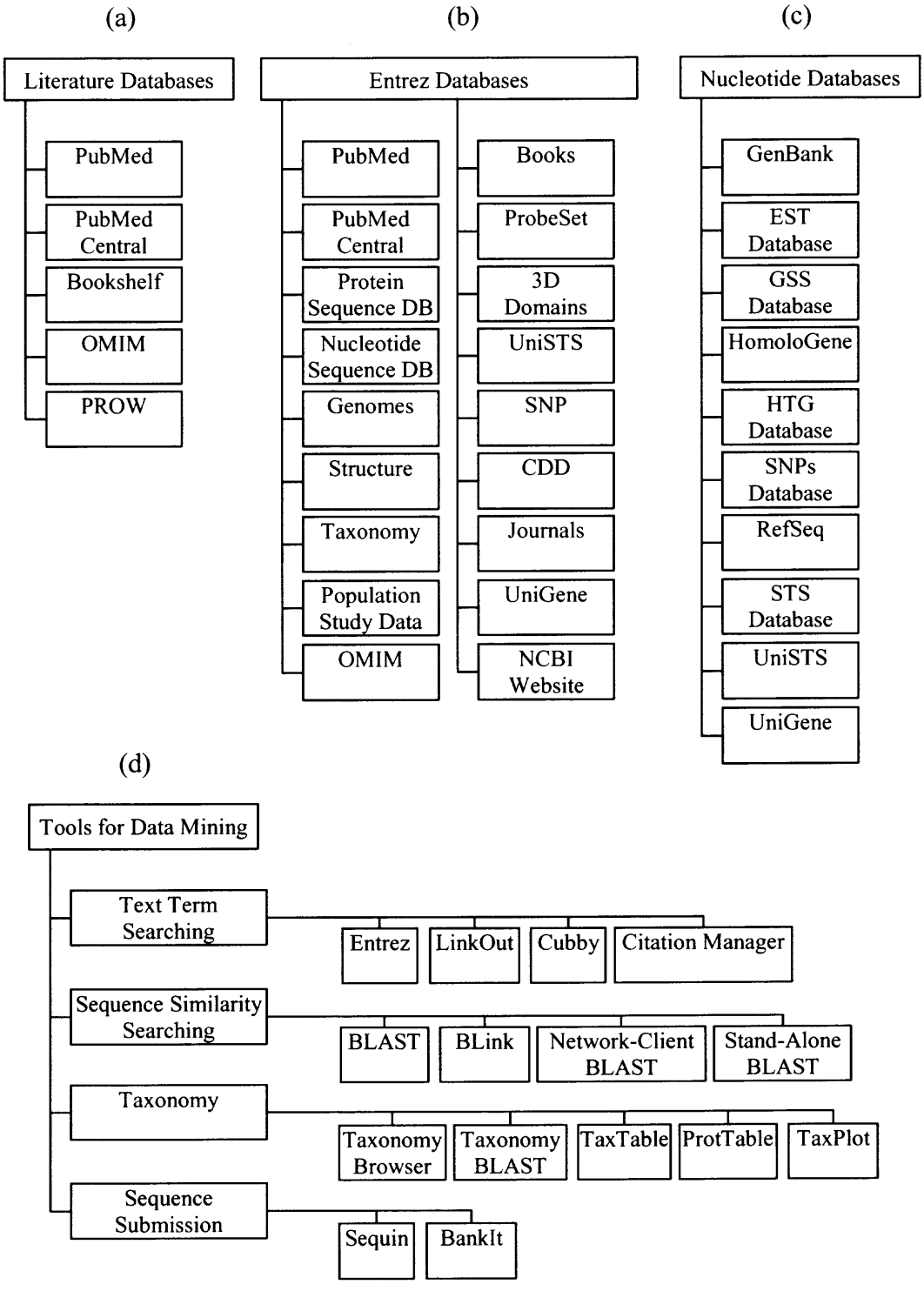

Figure 2.4 NCBI databases and tools

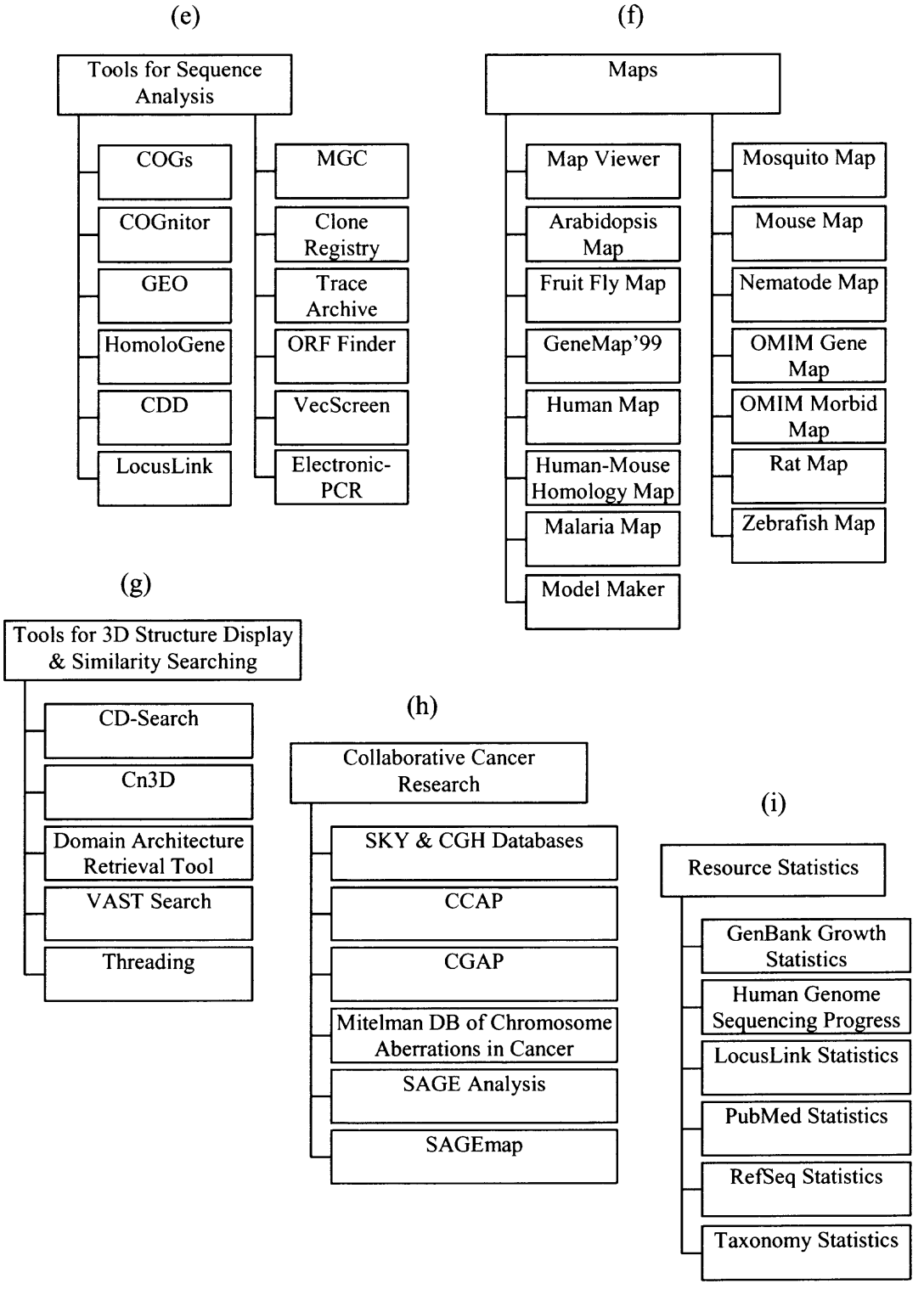

Figure 2.4 NCBI databases and tools (Continued)

2.2.2 EBI [9]

As a non-profit academic organization, European Bioinformatics Institute (EBI) forms part of the European Molecular Biology Laboratory (EMBL), which is an international collaboration of research institutes funded by fifteen countries for research in molecular biology. The EBI can trace its roots to the EMBL Nucleotide Sequence Data Library that was established in 1980 in Heidelberg, Germany and was the world's first nucleotide sequence database. The EBI itself was established in 1992 after the EMBL recognized that the enormity of the task of sequencing required special research facilities and financial security. It was decided to locate the institute at the Wellcome Trust Genome Campus in the United Kingdom, near the major sequencing efforts at the Sanger Institute and the Human Genome Mapping Project (HGMP) Resource Center.

The EBI conducts research and provides information about bioinformatics to the scientific community. Figure 2.5[7] provides a synopsis of the activities of the EBI. These activities are similar to those of the NCBI and include:

- developing bioinformatics tracking technology;
- conducting research and development of bioinformatics software;
- training and supporting its subscribers; and
- providing relevant bioinformatics services.

Research at the EBI includes the study of molecular evolution, genome comparison, gene prediction, protein motifs, metabolic pathways, sequence-structure relationships, the application of parallel computing in molecular biology, the analysis of biomolecular sequences and 3D structures, new biological databases, and navigation tools for linking databases.

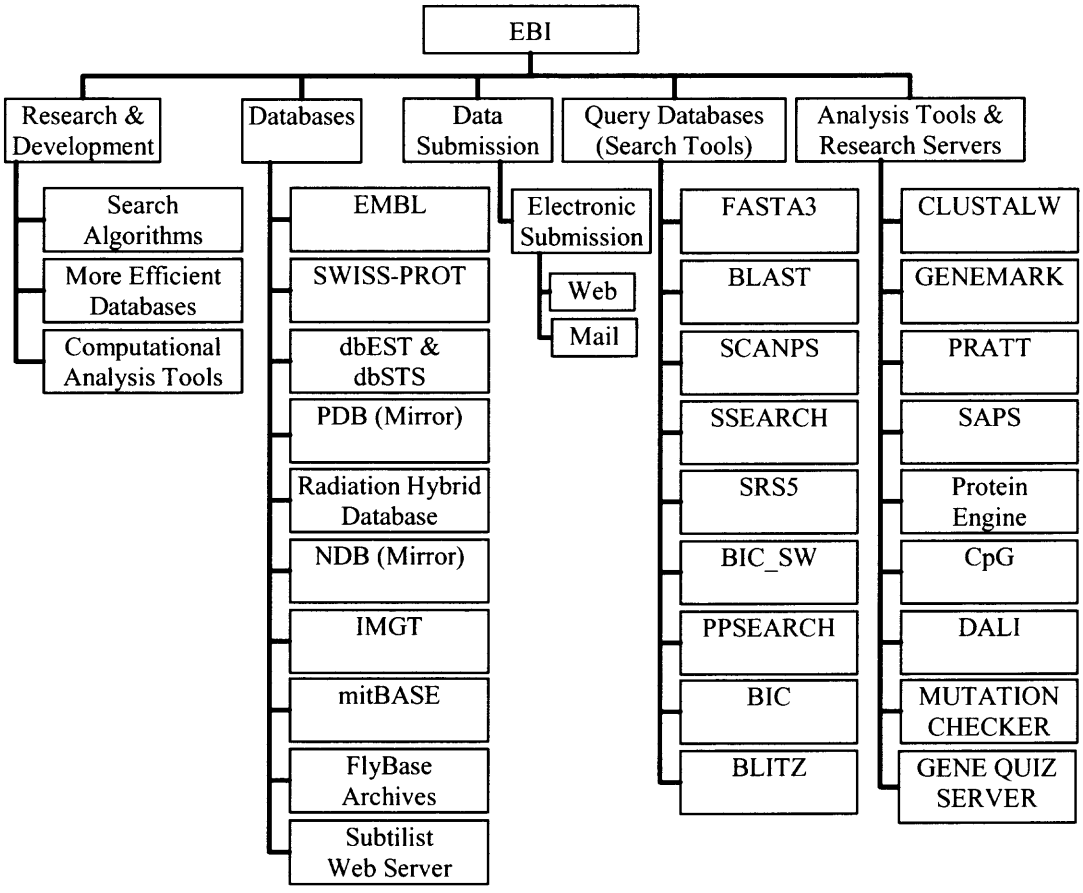

Figure 2.5 Synopsis of the EBI

The EBI maintains versions of most major public domain sequence database searching and analysis tools. The EBI services are constantly reviewed and upgraded through research and development activities. Figure 2.6 illustrates the various databases and tools that can be accessed through the EBI. The EBI provides access to many tools for browsing and retrieving biological sequence and literature data (Figure 2.6a). The Sequence Retrieval System (SRS) is useful for browsing the various EBI sequence and literature databases. It provides access to large volumes of assorted biological data stored in over 400 internal and public domain databases. The EMBL Sequence Version Archive

(EMBL-SVA) server is a repository of all entries that have ever appeared in the EMBL Nucleotide Sequence Database. The Generic Database Entry Retrieval, dbfetch or emblfetch, permits the user to retrieve up to 50 entries at a time from various databases. One entry at a time can be retrieved from the MEDLINE literature reference database through medlinefetch by entering a MEDLINE ID or PMID number into the search dialog.

The EBI has developed and maintains a number of protein sequence databases (Figure 2.6b). The Swiss-Prot Protein Knowledgebase is a curated and annotated protein sequence database that was established in 1986. It provides high levels of annotation and integration with other databases but a low level of redundancy. It is maintained collaboratively by the EBI and the Swiss Institute for Bioinformatics (SIB). The Translated EMBL (TrEMBL) database is a computer-annotated protein sequence database that supplements Swiss-Prot. It is a temporary storehouse of all the translated coding sequences (CDS) that are present in the EMBL Nucleotide Sequence Database but that have not yet been included in Swiss-Prot for information quality reasons. InterPro contains information about protein families, domains, and functional sites and can be used for whole-genome analysis and proteome analysis. CluStr (Clusters of Swiss-Prot and TrEMBL proteins) provides an automatic clustering of Swiss-Prot and TrEMBL proteins into groups of related proteins, bases on an analysis of all pair-wise comparisons between protein sequences. The International Protein Index (IPI) is a non-redundant human proteome set constructed from Swiss-Prot, TrEMBL, Ensembl, and RefSeq. GOA (Gene Ontology Annotation) provides assignments of Gene Ontology (GO) terms to gene products for all organisms with completely sequenced genomes by a combination

of electronic assignment and manual annotation. The Proteome Analysis database provides comprehensive statistical and comparative analyses of the predicted proteomes of fully sequenced organisms. The Human Proteomics Initiative (HPI) is an ongoing project that aims to annotate, describe, and distribute a large amount of highly curated information about human protein sequences. For each known protein, information including description of function, domain structure, subcellular location, post-translational modifications, variants, and similarities to other proteins, is sought to be provided. The Integrated relational Enzyme database (IntEnz) will contain enzyme data for the purpose of creating a single relational enzyme database. The TrEMBLnew database is a weekly update to TrEMBL that is produced from all new nucleotide sequences deposited in the EMBL Nucleotide Sequence Database. The SP_ML database is the Swiss-Prot and TrEMBL protein sequence databases in XML format. NEWT is a taxonomy database that integrates taxonomy data compiled at NCBI and data specific to the Swiss-Prot database. Protein and Associated Nucleotide Domains with Inferred Trees (PANDIT) is a collection of multiple sequence alignments and phylogenetic trees covering many common protein domains.

Figure 2.6c shows the different databases that fall under EBI's nucleotide sequence databases. The Alternative Splicing Database (ASD) contains annotated data on alternatively spliced exons and aims to help investigations of the mechanism of alternative splicing on a whole-genome scale. The EMBL Nucleotide Sequence Database is the premier nucleotide sequence resource of the EBI. The primary sources for the DNA and RNA sequences in this database are direct submissions from individual researchers, genome sequencing projects and patent applications. The database is

produced through the collaboration between the EBI, the NCBI, and the DDBJ (DNA Database of Japan), with each group collecting a portion of the total sequence data reported worldwide and all new and updated database entries being exchanged between the groups on a daily basis. Ensembl is a joint project between the EMBL-EBI and the Wellcome Trust Sanger Institute that seeks to develop a system that maintains automatic annotation of large eukaryotic genomes. The Genomes Server provides access to completed genomes at the EBI. The Genome Monitoring Table (MOT) monitors the worldwide progress of large genome sequencing projects and presents in a table the status of several such projects. The EMBL-Align is a multiple sequence alignment database. The Parasites database provides access to various parasite genomes. The Mutations database will eventually have data generated by the Sequence Variation Database Project. The Immunogenetics database (IMGT) consists of the IMGT/LIGMDB database of immunoglobulins and T-cell receptors, the IMGT/HLA database of the human MHC complex, and IMGT/MHC database of the MHC complex of non-human species.

The EBI has developed and maintains several protein structure related databases (Figure 2.6d). The Macromolecular Structure Database (MSD) group is responsible for the deposition and validation of new protein structures. The MSD collects, manages, and distributes macromolecular structure data derived in part from the Protein Data Bank. The MSD collaborates with members of the Research Collaboratory for Structural Bioinformatics (RSCB) in the U.S. and with the PDBj in Japan to support the PDB. The 3D sequence alignment server (3Dseq) provides annotated alignments between the Nucleotide Sequence Database and the PDB. The Fold classification based on Structure-Structure alignment of Proteins (FSSP) database is based on exhaustive all-against-all 3D

structure comparisons of protein structures currently in the PDB. The Distance Matrix Alignment (DALI) server is an automatic service for the comparison of 3D protein structures. The 3Dee database contains protein domain definitions.

The EBI has a microarray informatics group that is engaged in managing, storing and analyzing microarray data. ArrayExpress (Figure 2.6e) is a public repository for microarray data and aims to storing well-annotated data following the recommendations of the Microarray Gene Expression Data (MGED) Society. Also available through the EBI are various tools for querying and submitting microarray data, such as MIAMExpress and Expression Profiler.

Figure 2.6f shows the various bioinformatics associated literature databases accessible through the EBI. The MEDLINE database is the NLM's premier bibliographic database stretching over the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the pre-clinical sciences. It contains bibliographic citations and author abstracts from over 4,000 biomedical journals published in the U.S. and 70 other countries. The database contains over 11 million citations dating back to the 1960's and is updated weekly. The Biocatalog is a software directory of general interest in molecular biology and genetics. The FlyBase database is a repository of genetic and molecular data for Drosophila (fruit fly). It includes data on all species from the family Drosophilidae though the main represented species in Drosophila melanogaster.

Figure 2.6g lists the various tools available at the EBI for conducting protein functional analysis. CluSTr Search permits searching of Swiss-Prot and TrEMBL by accession numbers and cluster identification numbers. InterProScan allows searching for

protein sequences in the InterPro member databases. FingerPRINTScan identifies a queried protein sequence as a member of a known family, inferring a wealth of known information about that family and its members. It also classifies sequences using familial definitions from the PRINTS database. ppsearch allows search for protein motifs given a protein sequence as input. It can also perform comparisons against all patterns stored in the PROSITE pattern database and help in determining the function of an uncharacterized protein. The GeneQuiz system provides highly automated analysis of biological sequences. It is an integrated system for large-scale biological sequence analysis (from protein sequence to biochemical function), using a variety of search and analysis methods and up-to-date protein and DNA databases. Pratt allows the user to search for patterns (protein motifs) conserved in sets of unaligned protein sequences. The Rapid Detection and Alignment of Repeats (RADAR) tool uses an algorithm for partitioning a query sequence into repeats. It identifies short composition biased as well as gapped approximate repeats and complex repeat architectures involving many different types of repeats in the query sequence.

The various sequence analysis tools available at the EBI are given in Figure 2.6h. The ClustalW tool is a popular general purpose multiple sequence alignment program for DNA or proteins. It computes the best match for selected sequences and lines them up so that regions of identity, similarity, and difference can be observed. Pair-wise global (Needleman-Wunsch) or local (Smith-Waterman) alignment can be performed using EMBOSS-Align. GeneMark is a gene prediction service that uses an algorithm based on Makov models of coding and non-coding regions within a sliding window. Such an approach is sensitive to local variations of coding potential and can illustrate details of

the coding potential distribution as well as gene identification. GeneWise is a program that compares a protein sequence or a protein profile Hidden Markov Model (HMM) with a DNA sequence. The DNA Block Aligner compares two DNA sequences assuming collinear blocks and can be used for finding protein motifs such as promoter regions. PromoterWise also compares two DNA sequences but permits inversions and translocations and is useful for finding promoters. Mutation Checker is a sequence validation utility. Genetic Code Viewer reviews genetic code differences. The CpG Plot/CpG Report is a CpG Island finder and plotting tool. Transeq is a DNA sequence translation tool. Reverse Translator is a reverse complement checker. Pepstats/Pepwindow/Pepinfo are EMBOSS programs for basic protein sequence analysis. Statistical Analysis of Protein Sequences (SAPS) statistically evaluates a wide variety of protein sequence properties such as compositional biases, clusters and runs of charged and other amino acid types, different kinds and extents of repetitive structures, locally periodic motifs, and anomalous spacings between identical residue types.

The EBI provides several sequence homology and similarity tools (Figure 2.6i). The FASTA algorithm is a sequence similarity and homology searching algorithm for nucleotide and protein databases. WU-Blast2 is the BLAST 2.0 (with gaps) version of Washington University. NCBI-Blast2 is different from WU-Blast2 and is actually the blastall program. Blast2-EVEC checks sequences for vector contamination. Genome and Proteome FASTA server provides access to completed genomes and proteomes for FASTA searches. MPsrch is a fast implementation of the Smith and Waterman algorithm. It performs a comprehensive search of protein sequence databases in a short time. In contrast with BLAST and FASTA that use a heuristic algorithm, MPsrch

employs an exhaustive algorithm. Scanps2.3 is a new version of Scanps Fast implementation of the Smith and Waterman algorithm for protein database searches. Parasites Blast is the Parasites Genomes blast server. EGI blast is the blast server of EST clusters and alignments based on the EuroGeneIndexes. SNP-Fasta server provides FASTA searches of the European SNP database (HGBASE).

The EBI also provides several tools and services for the determination of a protein's 2D or 3D structure (Figure 2.6j). As noted above, DALI is a network service for comparing 3D protein structures using distance matrix alignment, FSSP is used for fold classification and structural alignment or proteins, and 3Dseq is the 3D sequence alignment server that produces annotated alignments. MaxSprout is an algorithm for generating protein backbone and side chain 3D coordinates from a C(alpha) trace. The backbone is assembled using known structure fragments and side chain conformations are optimized using an approximate potential energy function. The Protein Quaternary Structure (PQS) query allows searching of likely quaternary structures generated at the EBI. CHEMPDB is a ligand library at EBI and contains complete chemical description of all the distinct chemical components found in the PDB. The Secondary Structure Matching (SSM) tool allows comparison of protein chains or structures and searching for similar ones in the whole PDB archive or among SCOP (Structural Classification of Proteins) domains. The OCA server permits the user to rapidly search through the contents of the entire PDB Archive. PDBLite provides easy access to the PDB. The Biotech Validation Suite for Protein Structures provides a comprehensive check report of a protein.

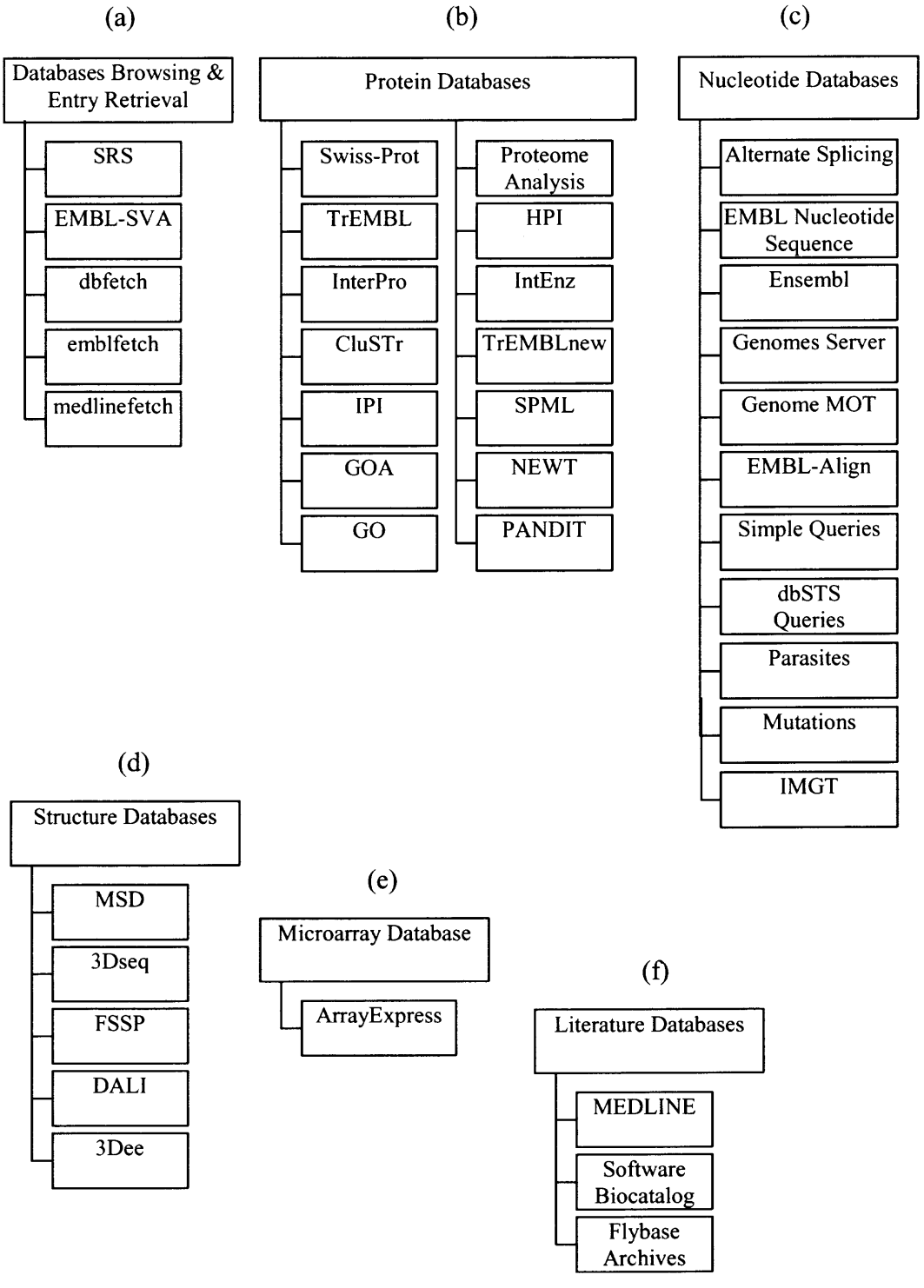

Figure 2.6 EBI databases and tools

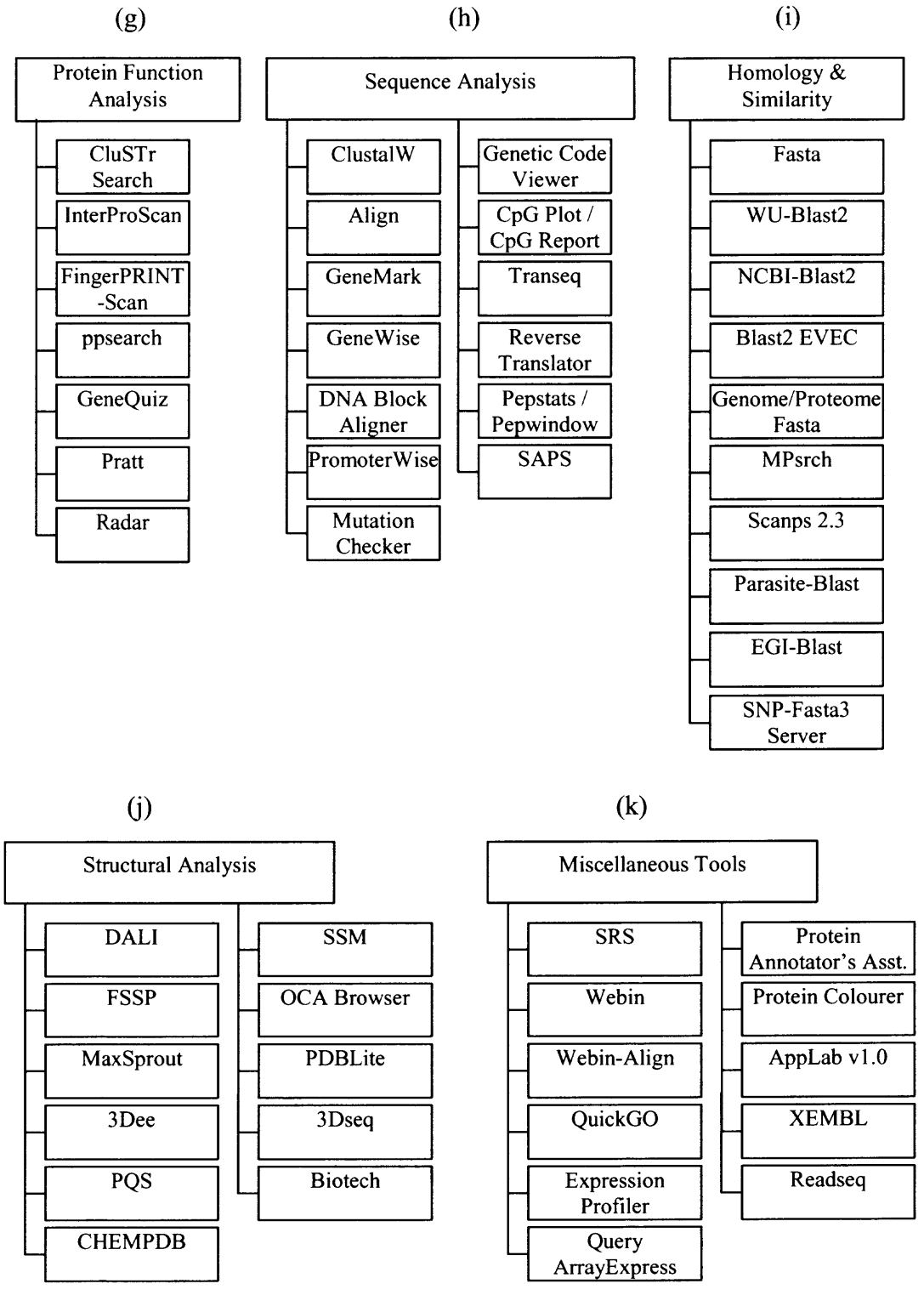

Figure 2.6 EBI databases and tools (Continued)

Figure 2.6k lists miscellaneous bioinformatics related tools available at the EBI. Webin allows submission and annotation of nucleotide sequences to the EMBL database. Webin-Align allows submission and annotation for nucleotide and amino acid multiple sequence alignment to the EMBL database. QuickGo is a web base viewer for the Gene Ontology data. Expression Profiler is a set of tools for clustering, analysis, and visualization of gene expression and other genomic data. Query ArrayExpress permits searching of the ArrayExpress microarray database. Protein Annotation Assistant is a tool that helps with protein annotation. Protein Colourer is a tool for coloring amino acid sequences. AppLab v1.0 is a CORBA-Java based application wrapper. XEMBL provides EMBL records in BSML or AGAVE XML format. Readseq is a sequence format conversion utility.

2.3 Sequencing Techniques

This section discusses some of the prevalent techniques for sequencing genomic regions and their assembly into a genome. DNA sequencing is the process of obtaining the string of bases that a DNA molecule contains. Typically, sequencing involves cutting, copying, and reading DNA. The result of sequencing is a readable data unit, the DNA sequence, which is then entered into a database as a digital record.

Cutting: DNA is typically fragmented at specific sites using restrictions enzymes (or endonucleases), which are proteins that catalyze the hydrolysis of DNA at specific restriction sites that are determined by the local base sequence at such sites. Special sequence segments, called palindromes, exist where the sequence in one direction is the same as its reverse complement. A particular restriction enzyme will cleave the DNA at

the same point in the sequence on both strands resulting in the generation of unique "sticky" ends. For example, the sequence 5'...GACGTC...3' and its reverse complement are palindromes. When a restriction enzyme that cleaves a DNA strand between G and A acts upon this double stranded palindrome, it produces an uneven cut in the double strand with one strand breakage producing 5'...G—AGCTC...3' and the other strand breakage producing 3'...CTGCA—G...5'. Thus, for both strands, the 5' end at the cut point will be four bases shorter than the 3' end. These "sticky" ends are then free to pair up with other DNA fragments similarly cleaved by the same restriction enzyme. This technique is frequently employed in genetic engineering for the production of recombinant DNA. Some bacteria use restriction enzymes as a defense mechanism against viral attacks by cutting the viral DNA before damage. The DNA of such bacteria is protected against its own restriction enzymes by the methylation of some of its bases.

In the above example, the target sequence for the restriction enzyme contained 6 base pairs. It is more common to come across 4-, 6-, and 8-cutter restriction enzymes than it is to find an odd-cutter restriction enzyme. This is because sequences containing an odd number of base pairs cannot be palindromes.

Another commonly employed method to break apart DNA molecules is the so-called "shotgun method". This method involves using an abrasive method, such as high vibration levels, upon a solution that contains purified DNA. This process results in a random breakage of each DNA molecule at several points in its sequence resulting in DNA fragments of varying lengths. A sample of these fragments is then extracted by filtration and subjected to further processing leading to their amplified copying (or cloning). The sequence determination of these fragments and then assembling the

resulting sequences to produce a cloning library is a standard way of finding the sequence of DNA.

Copying: This process is also known as DNA amplification and involves generation of numerous copies of a given fragment of DNA. Important rationales for making a large number of copies of a DNA fragment include the availability of a sufficient quantity of starting DNA material for any chemical reactions involving the DNA fragment, and the availability of some DNA material for storage purposes for use at a later date. Two common copying methods are recombinant DNA cloning and the polymerase chain reaction.

Recombinant DNA is produced by inserting a DNA fragment into the genome of a host organism (or vector). This fragment will be multiplied as the host reproduces and multiplies its own DNA. The host is then destroyed and only the clones of the originally inserted fragment are extracted. Common vectors include plasmids, cosmids, phages, and yeast artificial chromosomes (YACs). Plasmids have a limitation on the size of the inserted fragment of about 15 kilobasepairs (kbp). The insert size limit for phages is about 25 kbp and for cosmids about 50 kbp. YACs can be used for inserts on the order of 1000 kbp (one million base pairs).

The polymerase chain reaction (PCR) is a way to produce large numbers of a DNA fragment without cloning. It involves using the enzyme DNA polymerase that catalyzes the elongation of a single strand of DNA given the template DNA to which this single strand is attached. Thus, the method uses base-complementarity to elongate a small single strand until both strands achieve the same length. PCR consists of two basic alternating steps: the denaturing by heat step in which a double-stranded DNA is

separated into two single strands, and the elongation step in which each single strand is then converted into a double strand by the addition of a small amount of "primer" (short double-stranded DNA) followed by polymerase action. Since each repetition of these two steps produces two new molecules it results in an exponential growth of the number of copies of the original DNA fragment.

Reading: A DNA sequence is read by using gel electrophoresis that is a technique based upon the size separation of molecules and uses the fact that nucleic acid molecules are charged in solution and will move in a specific direction upon the application of an electric field. The smaller the molecule the greater the distance it will travel within the gel. As an example, given the DNA fragment CTGAATCTAGTCCTTTGA, and a restriction enzyme that cleaves fragments immediately after every A, the collection of sub-fragments will contain CTGA, CTGAA, CTGAATCTA, and the original sequence. Gel electrophoresis of this collection of sequences will separate them by size. If this process is repeated using restriction enzymes for each of the other nucleotides as well, giving different collections of sequences that end in each of the four bases, then it is possible to determine the precise base composition of the original DNA sequence. Up to 700 bp fragments can be read in this way. The DNA fragments could also be labeled using radioactive isotopes or fluorescent dyes, which would permit the production of a graphic record of the sequence positions.

CHAPTER 3

MICROARRAYS

3.1 Concept

Microarrays[10,11] are small, solid supports upon which sequences from thousands of different genes can be tethered at precise locations. The supports are usually glass microscope slides but can also be silicon chips or nylon membranes, with the DNA being printed, spotted or synthesized directly on the support. Each spot on the array is a small deposit of DNA, cDNA, oligonucleotide, or even protein, and can be used to identify a particular gene sequence. A DNA microarray, for example, works by exploiting the ability of a given mRNA molecule to bind specifically (or hybridize) to its parent DNA template. Thus, by using one microarray that contains many DNA samples, a single experiment can determine the expression levels of thousands of genes within a cell by measuring the amount of mRNA bound to each site on the microarray.

3.2 History and Applications

Microarrays can trace their roots back to 1975 when E. M. Southern introduced the Southern blots as a method to separate restriction enzyme digested genomic DNA. Southern's key insight was the use of labeled nucleic acid for the purpose of probing nucleic acid molecules tethered to a solid support.[12] In Southern blotting, the denatured DNA is transferred on to a supporting membrane on which DNA fragments are then identified by their hybridization with gene-specific probes. Thus, the Southern blot became the first array. The introduction of a one-to-one correspondence between clones

and hybridization signals was achieved by the subsequent use of filter-based screening of clone libraries. Next came gridded libraries that were stored in microtitre plates and stamped on to filters in fixed positions, and were used to identify each clone uniquely. Such advances made possible applications like expression analysis by the hybridization of mRNA to cDNA libraries that were gridded on nylon fibers.

The main thrust in the level of interest in array technologies came in the form of two major innovations: miniaturization and high-density synthesis of oligonucleotides. Miniaturization was made possible by the using non-porous solid supports like glass. Some protocols permit robotic spotting of about 10,000 cDNAs on to a microscope slide and their hybridization with a double labeled probe. Solid supports also allow fluorescence-based detection of hybridization. High-density spatial synthesis of oligonucleotides has been developed by several groups using photolithographic masking techniques imported from semiconductor manufacturing,[13] or by developing *in situ* synthesis with reagents that are delivered by ink-jet printer devices. These innovations have made possible the development of *micro*arrays.

While the earliest microarrays were typically used for gene screening and target identification,[14,15] microarray applications have since been expanded in range. Microarrays are now used in disease characterization,[16] developmental biology,[17,18], pathway mapping, mechanism of action studies, and toxicology. Microarrays have fast gained popularity in cancer research where they have made possible the molecular characterization of tumors on a genomic scale and promise more reliable diagnosis and better treatment of cancer.[19-27] In immunological studies,[28,29] microarrays permit studies of host genomic responses to bacterial infections and can offer useful insights into the

process of reversing immunity. Typical transcript-level microarray studies involve comparison of mRNA levels in different types of cells by varying tissue (e.g., liver vs. brain), treatment (e.g., drugs X, Y, and Z), cell-state (e.g., tumor vs. non-tumor, or development stage), organism (e.g., different strains of yeast), or timepoint. Still other potential applications include use of microarrays in personalized medicine through profiling studies, in molecular diagnosis of disease, and in predicting differences in drug efficacy and toxicity from person to person. Large scale protein-DNA interactions, genotyping, and biochemical pathways investigations also offer important areas of research using microarrays.

Three important microarray application categories are described below based on the type of DNA microarray used: Comparative genomic hybridization (CGH),[30] which is widely used in tumor classification, risk assessment, and prognosis prediction; expression analysis,[31] which is used in drug development, drug response, and therapy development; and mutation analysis,[32] which is used in drug development, therapy development, and disease progression tracking. These microarrays differ from each other in the kind of immobilized DNA that is used to make the array and, therefore, in the kind of information that the array can provide. Two of these applications (CGH and mutation analysis) use genomic DNA sequences as samples while the third (expression analysis) uses transcriptomic (i.e., derived from mRNA) samples.

3.2.1 Comparative Genomic Hybridization (CGH)

Mutations are usually tackled by DNA repair genes, which, therefore, play a big role against cancer. Mutations within DNA repair genes themselves often result in lost or broken chromosomes and it is thought that certain chromosomal gains and losses are

related to cancer progression. Comparative genomic hybridization is a technique for determining genomic gains and losses by looking for a change in the number of copies of a particular disease-state gene. Large pieces of genomic DNA are immobilized on the microarray and become the target for fluorescently labeled genomic DNA extracted from both normal (i.e., control) and diseased (i.e., sample) tissue. If the number of copies of a particular target gene has gone up in the diseased state relative to the normal state, then a large amount of sample DNA (labeled red, for example) will hybridize to those spots on the microarray that represent the gene involved in that disease. Only small amounts of control DNA (labeled green, for example) will hybridize to those same spots. Thus, there will be a preponderance of red fluorescence over green fluorescence indicating that the number of copies of the disease gene has gone up.

3.2.2 Expression Analysis

Expression analysis is the process that determines the level at which a gene is expressed. Microarrays used for such analyses are called expression arrays. The immobilized DNA is the complimentary DNA (cDNA) that is derived from the mRNA of known genes. The hybridization mixture contains cDNA taken from both normal (green) and diseased (red) tissue. Over-expression of a gene during a particular disease-state would result in the hybridization of more sample cDNA and less control cDNA to the spots representing that expressed gene. Thus, that spot will fluoresce red with greater intensity than it will fluoresce green. Such analyses for various genes involved in different diseases allows cDNA derived from any individual to be tested to determine whether the expression pattern of the gene from that individual matches the expression pattern of a known

disease. Treatment can be initiated if such a test comes out positive. Expression arrays are widely used in cancer fundamental research and can also be used in new-drug design.

3.2.3 Mutation Analysis

Microarrays can be used to detect mutations in a gene sequence. In this case, the immobilized DNA is usually that of a single gene, though the sequence immobilized in one spot on the array will be different from a sequence immobilized on a different spot. The difference in two such sequences could only be of one or a few specific nucleotides (representing mutations). Single Nucleotide Polymorphism (SNP) is one type of sequence that is commonly used in this type of analysis. SNP microarrays can be used to test an individual for a particular disease expression pattern for which an associated SNP pattern has been previously established. Given a particular microarray that contains various SNPs, genomic DNA from an individual will hybridize with greater frequency only to the specific SNPs associated with that individual. Such a test might provide clues about the susceptibility of the individual to developing the disease(s) linked to those particular SNPs.

3.3 Types of Microarrays

There are two basic types of microarrays: spotted (or cDNA or Stanford) microarrays and oligonucleotide microarrays. Spotted microarrays are made by immersing pens into a concentrated DNA solution and physically, by means of a robot, depositing the DNA on to regularly spaced spots on a glass microscope slide. Spotted microarrays have longer deposited sequences that can often be complete cDNA or EST sequences. Oligonucleotide microarrays usually are higher-density and have 25-nucleotide (25-mer)

or shorter sequences that are packed on precise locations on the microarray. The microarrays analyzed in this study are one of each kind. Thus, a spotted microarray from Compugen® is compared to a high-density oligonucleotide array from Affymetrix®. In the following discussion, the same probe-target terminology for both spotted and oligonucleotide microarrays has been preserved. Thus, for both types of microarrays, the immobilized sequences on the support are called probes and the hybridizing sample sequences are called targets.

3.3.1 Spotted Microarrays (Compugen®)

Figure 3.1 illustrates the overall process of cDNA microarray preparation. The overall process contains four sub-processes: Preparation of the slide that would support the probe oligonucleotide sequences, preparation from mRNA of cDNA targets for hybridization, hybridization reaction of the probes, and the targets, and image data capturing and analysis.

The cDNA spotted microarrays involve glass microscope slides that can contain on their surfaces hundreds to thousands of immobilized cDNA sequences. Glass supports have certain unique advantages over other types of supports such as nylon. A treated glass surface permits covalent binding of DNA resulting in effective immobilization of the probes and, subsequently, a strong hybridization with the targets. Glass is a durable substance that can handle high temperatures and high pH changes, besides being non-porous, which means that hybridization volumes can be minimized resulting in better kinetics of the hybridization process. Glass has low fluorescence and, therefore, contributes minimally to background noise. These properties of glass make it a good choice as supports.

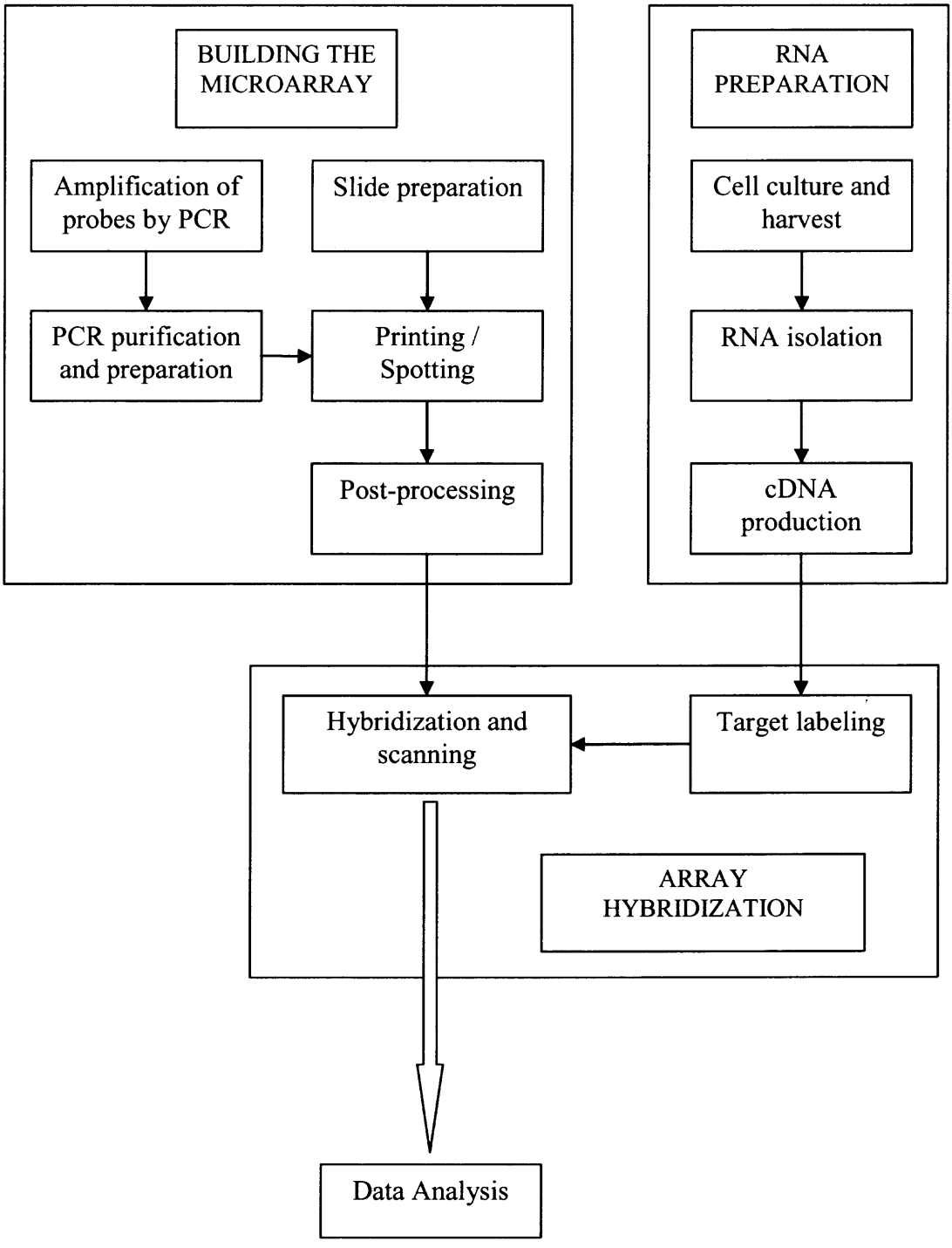

Figure 3.1 cDNA microarray synopsis

Robotic arrayers are needed to place the probes on to precise locations on the glass slides. Typically, the robots automatically collect samples from either 96- or 384-well microtitre plates with the help of several pens, pins, tweezers, capillaries, or ink-jets. The pen tips are usually fragile and great care is needed to manufacture, maintain, and operate them. Each pen collects a small volume of sequence solution and deposits it in spots ranging between 50 and 200 microns in diameter. The arraying system is usually controlled by computer programs running on common operating system environments like Microsoft Windows NT. Several slides can be arrayed simultaneously and at least 100 arrays with 10,000 features (spots) can be produced per day using some arrayers. The slides are first aligned in a uniform pattern on a table before the plates. The robot then proceeds with the spotting process with the pens collecting solution from the microtitre plates and depositing them sequentially on each of the slides. This is followed by a wash and dry process and then the whole procedure is repeated with a new set of samples until all samples have been arrayed. Following spotting, the slide is dried and the probes are immobilized using ultraviolet (UV) irradiation to form covalent bonds between certain residues in the DNA and charged groups on the slide.

Preparation of cDNA targets involves the preparation of control (e.g., from normal tissue) and sample (e.g., from diseased tissue) cDNA sequences using reverse transcription or polymerase chain reaction. As shown in Figure 3.2, mRNA is extracted from normal tissue and the cDNA manufactured from it is labeled with green fluorescent dye. Similarly, mRNA is extracted from sample tissue and the cDNA made from it is labeled with red fluorescent dye. The labeled cDNA sequences are then combined in equal amounts to produce the hybridizing solution.

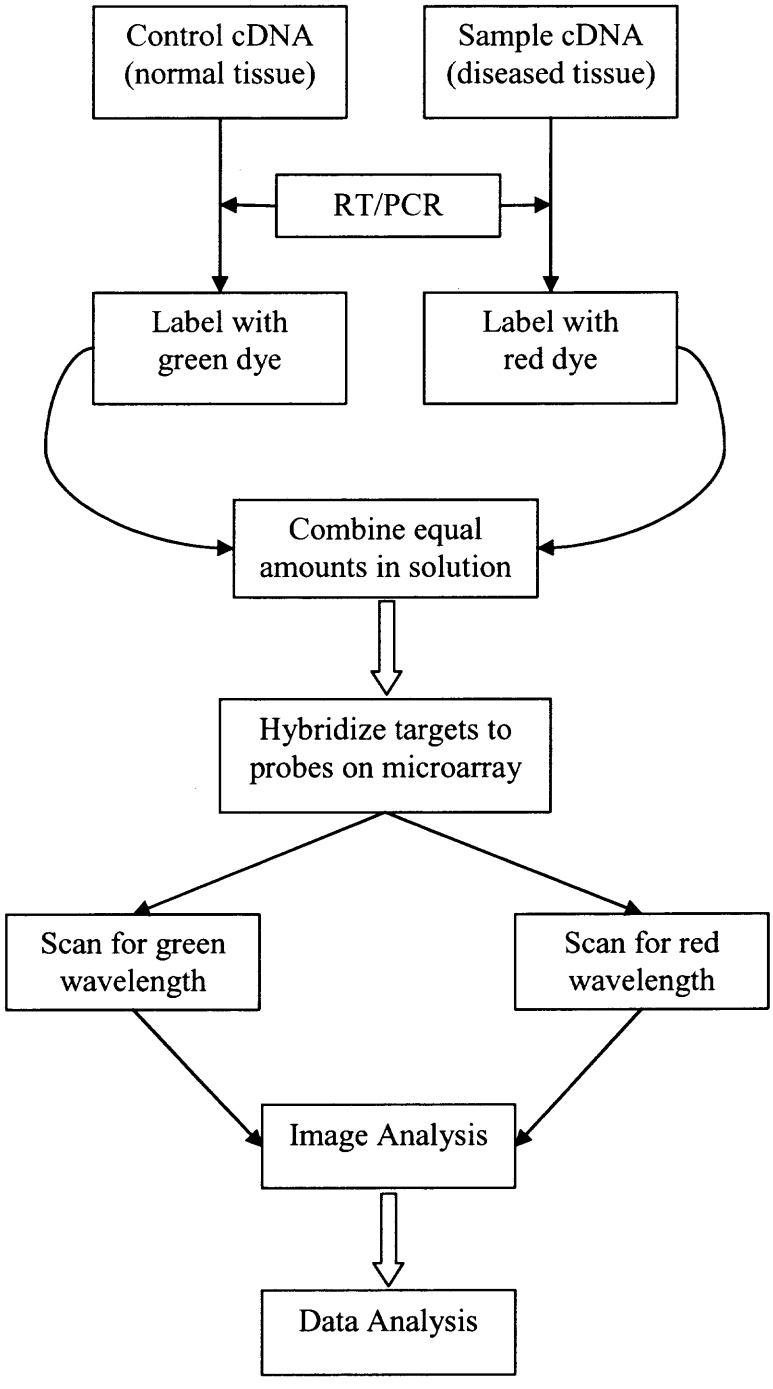

Figure 3.2 cDNA microarray data generation

The probes on the microarray are soaked with the hybridizing solution containing the targets. Hybridization is aided by shaking or vibrating the microarray to allow sufficient mixing of the probes and the target containing solution. The binding of the probe to the target is typically detected by scanning the microarray using a either a scanning confocal laser or a charge coupled device (CCD) camera-based reader. The scanner scans for the relative abundance of a probe sequence in the control and sample solutions by monitoring the differential hybridization of the targets to the probes. The ratio of green and red fluorescence intensities for each spot (probe) on the microarray indicates the relative abundance of the corresponding DNA probe in the control and sample targets. This ratio, say M, is given by $M = \log_2(\text{Red/Green}) = \log_2(\text{Red}) - \log_2(\text{Green})$. If M is negative, which is the case when there is a preponderance of green intensity over red intensity, then that indicates that the gene represented by this probe sequence is over-expressed in the green-labeled sample compared to the red-labeled sample. If M is zero, which is the case when both red and green intensities are comparable, then that indicates that the gene represented by this probe is equally expressed in both samples. If M is positive, which is the case when red intensity predominates over green intensity, then that indicates that the gene is over-expressed in the red-labeled sample compared to the green-labeled sample.

Some of the advantages of spotted glass slide microarrays are low cost per microarray, freedom to select custom genes, freedom to select any species for analysis, competitive hybridization, and an open architecture. Some of the disadvantages of these microarrays are quality control issues, clone management, and cost of cloning.

3.3.2 Oligonucleotide Microarrays (Affymetrix®)

Affymetrix® was established in 1991 in Santa Clara, California and completed its initial public offering in June 1996. It provides its oligonucleotide microarrays through its GeneChip® integrated system to pharmaceutical, biotechnology, agrichemical, and consumer products companies, and to researchers in academia, government, and other non-profit research institutes.

The manufacturing process of the oligonucleotide microarrays from Affymetrix® borrows technologies from the semiconductor industry. It involves the use of photolithography and solid-phase combinatorial chemistry to generate microarrays that contain hundreds of thousands of oligonucleotide probe sequences packed at very high densities. According to Affymetrix®, these probes are designed in a way that maximizes sensitivity, specificity, and reproducibility, which permits consistent discrimination between specific and background signals, and between target sequences that might be closely related to each other. Appendix A contains descriptive images about Affymetrix microarrays. In particular, Figure A.5 illustrates the target labeling process for an array.

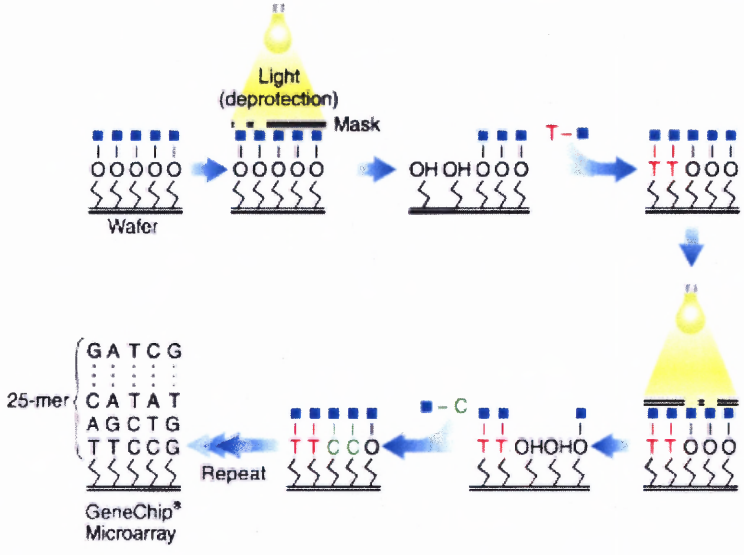

Figure 3.3 GeneChip® manufacture

Figure 3.3[33] shows Affymetrix's process of making their oligonucleotide microarray, also called GeneChip®. The process begins with the washing of a 5-inch quartz wafer to allow uniform hydroxylation across its surface. Being naturally hydroxylated, quartz forms a good substrate for the tethering of chemicals, such as linker molecules, that are used subsequently to place the probes on the microarray. The wafer is immersed in a silane solution and the silane chemically reacts with the hydroxyl groups on the quartz resulting in the formation of a matrix of covalently bonded molecules. The probe density is determined by the distance between these silane molecules, with some microarrays being able to support over 500,000 probe locations (or features) inside 1.28 cm^2. Millions of identical oligonucleotides are present in each of these features (see also Appendix A, Figure A.6).

When linker molecules attach to the silane matrix, they form a light sensitive surface, which is a property subsequently utilized in the array manufacturing process. The building of the oligonucleotides on the microarray is an *in situ* process, with the different nucleotides A, C, T, or G being added to several probes simultaneously. In each step, photolithographic masks that have 18 to 20 square micron windows (corresponding to the size of individual features) are place over the coated wafer with the windows exposing only those features that would receive a nucleotide (see Appendix A, Figure A.4). The illumination of these masks with ultraviolet light results in the de-protection of the exposed linkers which become available for nucleotide coupling in the next step. The activated features are washed with a solution that contains a single type of deoxynucleotide with a detachable protection group. The synthesis is initiated with the nucleotide linking with the activated linkers. In the next synthesis step, a different mask

might be placed on the wafer leading to de-protection and coupling all over again, using the next nucleotide in the oligonucleotide sequence. This process is repeated until the oligonucleotide reaches up to 25-mer length. Upon completion of the synthesis of all oligonucleotides on all probe features, the wafer is de-protected and cut into smaller squares that form individual microarrays. A single wafer can produce between 49 and 400 microarrays, depending on specific requirements of the number of features per array.

A gene on a GeneChip® is represented by 16-20 oligonucleotides that are each up to 25-mer long. Million of copies of a single such 25-mer probe populate each probe cell (feature) on the microarray. Probe cells are square-shaped having a side of 18-50 microns (see also Appendix A, Figures A.6 – A.8). The selection of probes that suitably represent genes is an important design issue that affects the reliability of GeneChips®. Other important issues in the manufacture of Affymetrix microarrays include optimal pH, salt, and temperature conditions. Affymetrix® also incorporates computer models based on empirical data to predict the intensity and concentration dependence of probe hybridization, in order to minimize probe number while maximizing data quality. The process of synthesizing oligonucleotides on to the microarray is also called probe tiling.

A unique feature of Affymetrix microarrays is the Perfect Match/Mismatch probe strategy. Each probe oligonucleotide that is designed to be exactly complimentary to a target sequence is called the Perfect Match probe (PM). Each PM is spatially partnered with a probe oligonucleotide that is identical except for a single base mismatch in its center (at the 13^{th} base position) and that is called the Mismatch probe (MM). The single base mismatch is homomeric, substituting A for T and vice versa or C for G and vice versa. Each pair of PM and MM probes is called a probe pair. A probe set contains 16-

20 probe pairs. This is illustrated in Figure 3.4. These PM/MM probe pairs permit the measurement of background noise and of hybridization signals that might be caused by non-specific cross-hybridization. According to Affymetrix®, the difference in hybridization signals between the PM and MM probes and the intensity ratios of these probe pairs can indicate the abundance of a specific target.

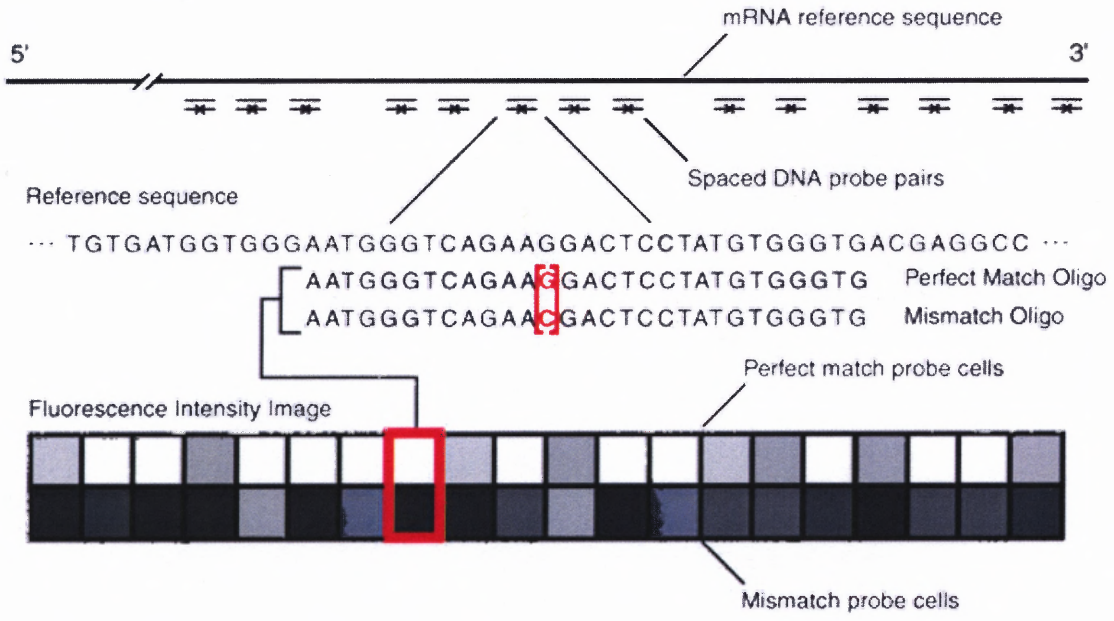

Figure 3.4 Perfect Match/Mismatch strategy

As shown in Figure 3.4, the oligonucleotide probes are ultimately derived from a reference mRNA sequence corresponding to a particular gene. The selection of probes and the design of a GeneChip® depend upon the chip's intended use. For gene expression arrays, for example, sequence and annotation data are culled from various databases. Global genetic activities can be monitored using Affymetrix GeneChips® for a variety of organisms, including yeast, Arabidopsis, Drosophila, mice, rats, and humans. Affymetrix® also caters to requests for custom expression microarrays for other

organisms. For human microarrays, the databases from which expressed human genome sequences are extracted include GenBank, RefSeq, and dbEST. These sequences are clustered by similarity and using the UniGene database as reference, the clusters are further divided into sub-clusters that represent distinct transcripts (mRNAs). Such a process involves the alignment of sequences to the human genome in order to get information about splicing and polyadenylation variants. High quality consensus sequences are generated using the alignment and the annotation provided by the databases. Representative sequences, or exemplars, could also be selected by quality ranking for the purpose of probe design. According to Affymetrix®, 11-16 probes are selected among all possible 25-mers that can represent each transcript (mRNA).

In some cases, probes can be selected from genomic regions that could be shared by multiple splice or polyadenylation variants so that a comprehensive idea about a gene's activity can be obtained. In other cases, the preference is for unique probes that can differentiate between variants. Another metric considered in the probe selection process is the distance between two probe sequences. In general, probe sequences are well spaced out so as to sample different regions of every transcript (mRNA).

Affymetrix's genotyping microarrays employ a different set of probe design strategies. These strategies include using multiple probes to interrogate individual nucleotides in a sequence, or using only one or two probes that represent specific alleles to check for the presence of a consensus sequence.

Data acquisition from a gene expression experiment is usually initiated with Affymetrix® Microarray Suite (MAS) software, which controls the microarray scanners provided by Affymetrix® (see Appendix A, Figures A.2, A.3, and A.9). The software

allows image acquisition and analysis for all GeneChip® microarrays as well as an interface for the storage and management of image data. MAS 5.1 software includes statistical algorithms for expression data analysis. According to Affymetrix®, these algorithms provide p-values for statistical significance for target detection levels, confidence limits for expression change values, user-modified parameters for stringency of data analysis, and other standard statistical techniques. The software allows a user to view the scanned image of a microarray interactively. For example, the user can click with a computer mouse on an individual probe cell's pixels and the software will provide the probe hybridization intensity and other probe cell information. The software can also display distributed probe set microarray that contain probes from a single probe set placed in different areas of the microarray.

CHAPTER 4

AFFYMETRIX VS. COMPUGEN HUMAN GENOME MICROARRAYS

4.1 Chip Design

Numerous short (about 200 to 500 base pairs), unique sequences called sequence tagged sites (STSs) occur only once in the human genome and their location and base composition are known. Expressed sequence tags (ESTs) are STSs that have been derived from cDNAs. STSs and ESTs are useful for localizing and orienting the mapping and sequence data that gets reported from various research laboratories worldwide. STSs and ESTs serve as landmarks on the physical map of the human genome. ESTs can act as identifiers of specific genes in the human genome. As a consequence, microarrays are frequently designed based on a selection of ESTs drawn from public or proprietary databases, such as the databases mentioned in Chapter 2.

Compugen's chip design strategy for its expression analysis microarrays is to use EST data as the information source. A gene is "constructed" by clustering ESTs that represent that gene using clustering algorithms that filter out repeats and detects chimeric sequences. Subsequently, mRNA transcripts are derived from each constructed gene since the microarray will be used for analyzing the expression of mRNA levels. Compugen® places great emphasis on the possibility of **alternative splicing**, which is the process by which a single gene may be transcribed into multiple mRNA sequences. Figure 4.1 shows three splice variants of a single gene. It has been found that alternative splicing occurs in at least 30% of all genes.[34] Each such gene has an average of 5 different splice variants of which 60% have been validated by sequencing PCR products

from only a few different tissues. Compugen® asserts that given such a high frequency of alternative splicing, a process that is germane to mRNA creation, having an exact sequence of each mRNA splice variant is crucial for probe selection. The probe selection process of Compugen® selects sequences that exclusively represent each mRNA.

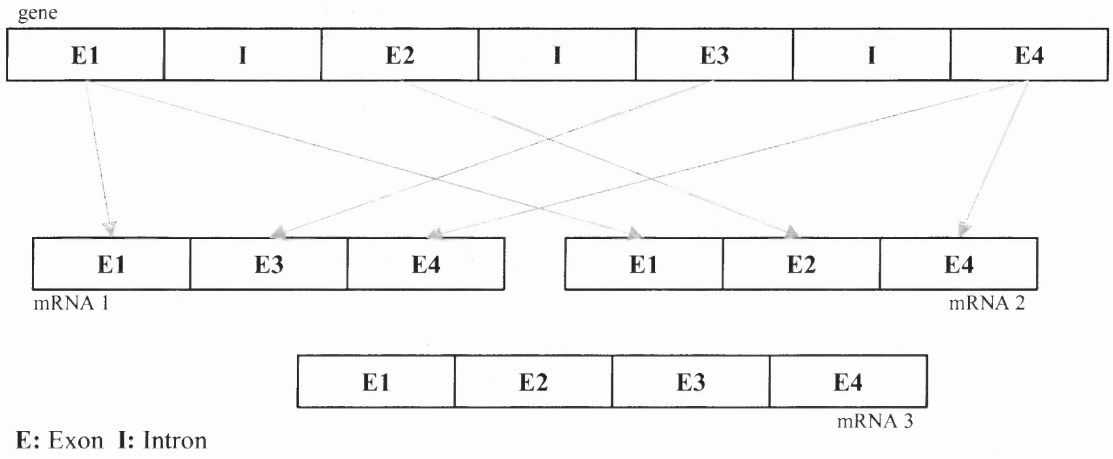

Figure 4.1 Alternative splicing

Compugen® spots 1 nmole of 60 bp long oligonucleotides on its human genome microarray. The total number of oligonucleotides on this microarray is 18,861, and these represent 46,581 mRNAs and 979,931 ESTs. The oligonucleotides are arranged on the microarrays according to standard Gene Ontology (GO) assignments, which permits easily locatable of oligonucleotides derived from genes with similar function. The number of unique genes represented on the human genome chip of Compugen® is 18,664.

Affymetrix® incorporated a number of new features in the probe selection strategy for its new Human Genome U133 Set (HG-U133). Appendix B provides some details about the probe selection process of Affymetrix®. Figure B.1 lists the probe sequence selection steps for the U133 Set. Sequences culled from various databases were aligned

to the draft human genome assembly and were clustered using UniGene clusters as seed clusters. Following a number of sub-clustering steps, candidate probe selection regions were identified. The final choice of the probes was dictated by the quality of the sequences and annotations. Figure B.2 illustrates the criteria used by Affymetrix® to determine multiple probes for a given sequence region. Table B.1 lists a number of differences between various human genome microarrays from Affymetrix®. For its previous HG-U95 Set, Affymetrix® relied on consensus sequences for probe selection. A consensus sequence is derived from the sequence of clusters (e.g., UniGene clusters) and is a base-by-base sequence call from the most 5'- to the most 3'-position in the cluster. The UniGene build 95 data were cleaned by first removing ESTs from very large clusters and then aligning and sub-clustering the remaining sequences using a Cluster and Alignment Tool (CAT). To be included in the consensus sequence, a given base at a certain position in the sequence had to agree for at least 75% of the aligned sequences. If this was not the case then that base position was identified with an "N" and was not included in probe selection. One or two consensus sequences per cluster were then tiled to create the microarray.

For its HG-U133 Set, Affymetrix® used 3,873,773 sequences out of a total of 7,665,267 human sequences available from UniGene, dbEST, WUSTL, GenBank, and RefSeq. In all, over six million sequences were considered for inclusion in the U133 Set. According to Affymetrix®, this set has improved consensus sequence quality by identifying and removing low-quality ESTs. The U133 Set contains the U133A and U133B microarrays. The U133A microarray contains exemplar sequences from the RefSeq database as well as sequences related to those on the older U95Av2 microarray.

The U133A microarray contains 22,283 oligonucleotides including 8,645 consensus, 13,570 exemplar, and 68 control oligonucleotides. The U133B microarray contains mainly consensus sequences from EST clusters. It contains 22,645 oligonucleotides including 20,991 consensus, 1,586 exemplar, and 68 control oligonucleotides. According to Affymetrix®, the two microarrays together contain representatives of more than 39,000 transcripts derived from about 33,000 well-defined human genes.

Table 4.1 shows the classifications and counts of sequences that have been mapped by Affymetrix® on the U133 Set. The probe selection process for this set does not use the heuristic rules that were used for previous microarray sets. Instead, the probe selection for the U133 set uses a multiple linear regression (MLR) model that was derived from a thermodynamic model of nucleic acid duplex formation and that predicts probe binding affinity and linearity of signal changes in response to different target concentrations. This model-based probe selection forms the basis for Affymetrix's assertion that it offers a physical and mathematical foundation for systematic and large-scale probe selection.

Table 4.1 Classifications and Counts of Sequences on the HG-U133 Set [35]

Classification	HG-U133A	HG-U133B	Total
UniGene Clusters	14,593	19,318	31,728
Additional Potential Full Lengths	513	198	707
Subclusters	18,462	21,070	38,903
Full Length Including UTR	13,049	1,556	14,529
Extended Full Length	171	58	229
Strongest evidence for polyadenylation	3,228	6,929	10,140
Complete CDS Consensus End	570	74	643
Non-EST Consensus End	2,526	2,755	5,278
Evidence for polyadenylation	993	595	1,586
EST-only clusters			
Oriented, Mapped, and 3'	279	9,153	9,432
Oriented and 3'	33	590	623
Mapped and 3'	14	76	90
3'	22	0	22
Opposite Consensus End	683	619	1,301
Distant Consensus End	176	150	326

4.2 Effective Genome Coverage

The data used for this research were provided by two sources. The Center for Applied Genomics (CAG), Newark, New Jersey provided data for the human genome microarray from Compugen®. The data for the HG-U133 Set was downloaded from the website of Affymetrix®. The CAG made available a list (as a Microsoft Excel® worksheet) of all oligonucleotide sequences that are mapped on the Compugen microarray. The list corresponds to Compugen's 96-well human oligonucleotide library data and, for each oligonucleotide, includes fields for plate ID, location ID, GenBank accession number, UniGene ID, brief description of the gene represented, Compugen gene ID, and the actual 60-mer oligonucleotide sequence.

The HG-U133 Set, consisting of the U133A and U133B datasets, was downloaded as several files from the website of Affymetrix®. For each of U133A and U133B sets, separate zipped files containing the consensus, exemplar, and control sequences were downloaded and unzipped. This resulted in the generation of several large text files, with the largest file being about 35 megabytes. Such large text files were opened using the robust and useful text editor TextPad® from Helios Software Solutions. The unzipped text files contained multiple sequences, each in FASTA format with a unique header that contained information including a unique ID, GenBank accession number, and other descriptive information about the sequence. Using TextPad, each header line was "bookmarked", all bookmarked lines copied, and all copied lines pasted into a new text file. Using the block-select mode and regular expressions in TextPad, the GenBank accession number for each sequence represented on the U133A and U133B sets was extracted and pasted into a Microsoft Excel® worksheet.

The GenBank accession numbers for the oligonucleotides represented on the Compugen and Affymetrix human genome microarrays were thus obtained in Microsoft Excel® file format. Based on a comparison of these GenBank accession numbers, it was verified that the Compugen microarray contains a much smaller set of GenBank identification numbers (18,861) than the two U133 Set Affymetrix microarrays (41,253 unique GenBank identification numbers and 3,539 repeats). The two different sets of microarrays share only 239 GenBank sequences. Table 4.2 shows that Compugen® uses 60-mer sequences to represent 18,664 genes and Affymetrix® uses 25-mer sequences to represent about 33,000 genes. This translates to a genome-wide coverage of 1,119,840 bases with the single Compugen microarray and of 825,000 bases with the two microarrays in the U133 Set of Affymetrix®. Thus, while the Compugen microarray interrogates almost half the number of genes than do the Affymetrix microarrays, they probe a larger total region within the human genome. Put another way, Affymetrix human genome microarrays interrogate many more genes but end up probing a much smaller total genomic region than the Compugen microarray. *It is pertinent, therefore, to ask whether the HG-U133 GeneChip® Set from Affymetrix® is more efficient than the Compugen spotted array.*

Table 4.2 Genome Coverage of Affymetrix and Compugen Microarrays

	Compugen	Affymetrix
Number of oligonucleotides	18,861	41,253
Number of genes represented	18,664	33,000
Length of oligonucleotides	60	25
Genome coverage (no. of bases)	1,119,840	825,000

Designing an experiment to compare the effectiveness of two different types of microarrays would require the consideration of a number of important issues. A direct comparison between a Compugen-type spotted cDNA array and an Affymetrix-type high-density oligonucleotide array cannot be made because of obvious array-to-array and within-array normalization and scaling considerations: The physical dimensions, biochemical treatment, and response identification of the two types of arrays are different and the direct response intensity comparisons would not be meaningful.

For the purpose of a strict comparison between the two sets of microarrays, a comprehensive selection of post mortem tissues from a single source could be made. Though all cells in the body contain the same complement of genes, for example, the cells in brain tissue contain the same DNA as those in liver tissue, cell differences arise because of the different genes that are expressed in the cells comprising different tissues. Thus, a suitable tissue set can be selected from which genomic samples can be extracted and probed under identical laboratory conditions by the two different microarray sets (Figure 4.2). The hybridization mixture thus generated would be used separately with Compugen and Affymetrix microarrays. Subsets of genes identified by the two microarray sets could be evaluated using quantitative real-time PCR and regions of overlap could be obtained. In such an experiment, it can be conceived that the gene subsets identified by the two microarray sets will differ substantially since Compugen's human genome microarray can probe only 18,664 genes as against the U133 Set's much larger 33,000 genes targeting capacity. However, such an experiment could only be used to verify the difference in the number of gene targets and not to quantify the effectiveness of the two types of microarrays.

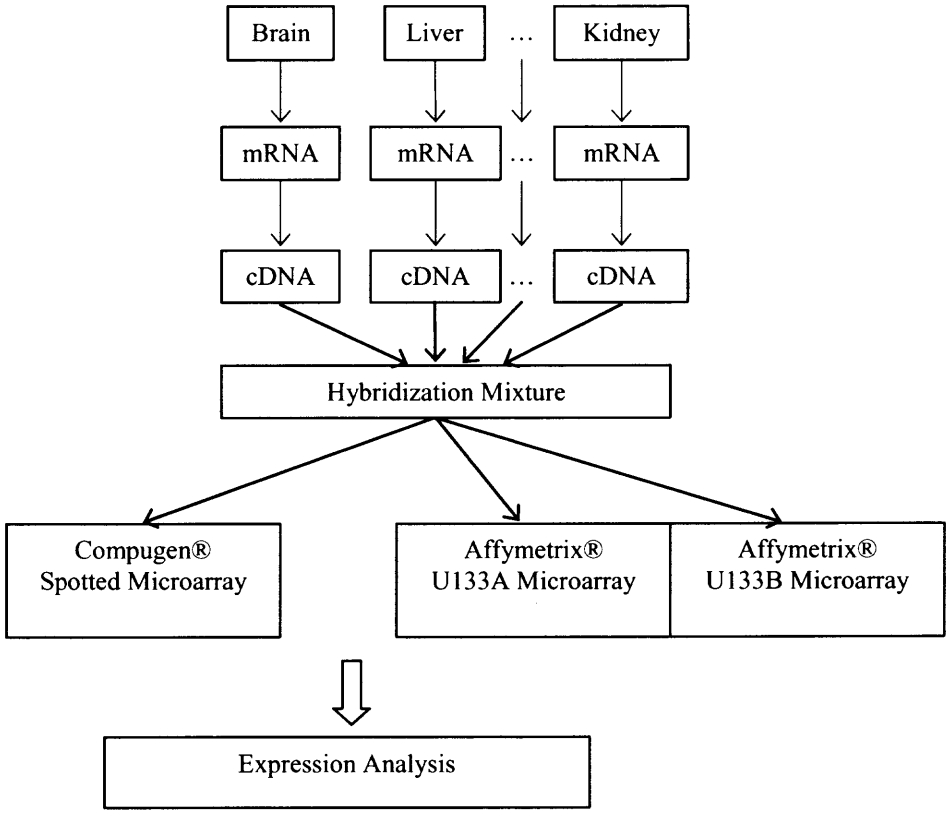

Figure 4.2 Experiment design for microarray comparison

An effectiveness comparison experiment could be based upon the fact that gene expression differences also exist between healthy and diseased tissues and between healthy and drug-adapted cells. In such an experiment, the hybridization mixture in Figure 4.2 would contain genetic information from normal, diseased, and/or drug-adapted tissues. Separate analyses of the microarrays would provide insights about which microarray gives better results.

CHAPTER 5

CONCLUSION

In this report, a comparison between the human genome microarrays from Compugen® and Affymetrix® was performed in the context of the emerging field of computational biology. The two main database servers for sequence data collection and dissemination, the NCBI and the EBI, were discussed in some detail. A description of the databases and tools available from the NCBI and the EBI was provided in an attempt to convey the complex data management issues related to the huge amount of sequence data that has been generated world-wide over the last several years. In addition, popular sequencing techniques that are used to populate sequence databases were also briefly discussed.

The concept of microarrays was examined from a brief historical perspective. Several modern applications of microarrays including expression analysis, comparative genomic hybridization, and mutation analysis were discussed. The two main types of microarray, spotted cDNA microarrays represented by Compugen microarrays and high-density oligonucleotide microarrays represented by Affymetrix microarrays, were described. The differences in array supports, array design, including sequence selection, sequence collection and analysis, and probe selection process were explored. Common arraying strategies of both Compugen® and Affymetrix® were discussed.

The respective chip design of the two types of microarrays was analyzed. It was found that Compugen's human genome microarray contains probes that interrogate 1,119,840 bases in the human genome while Affymetrix's HG-U133 Set probes a far fewer 825,000 bases in the human genome. Since the Compugen microarray represents

18,664 genes as against the Affymetrix U133 Set's 33,000 genes, the efficiency of Affymetrix's 25-mer probes as against Compugen's 60-mer probes was questioned. The difficulty of designing a cross-array comparison experiment was noted.

APPENDIX A

IMAGES OF THE AFFYMETRIX MICROARRAY

All images courtesy Affymetrix®. These images were obtained from the website of Affymetrix® at http://www.affymetrix.com/corporate/media/image_library/index.affx.

Figure A.1 Affymetrix GeneChip® probe array

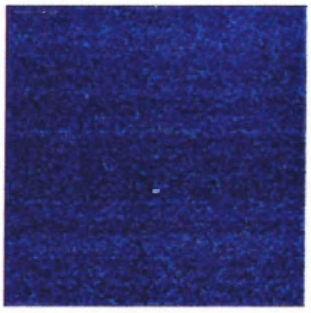

Figure A.2 Affymetrix GeneChip® Rat 230A array image

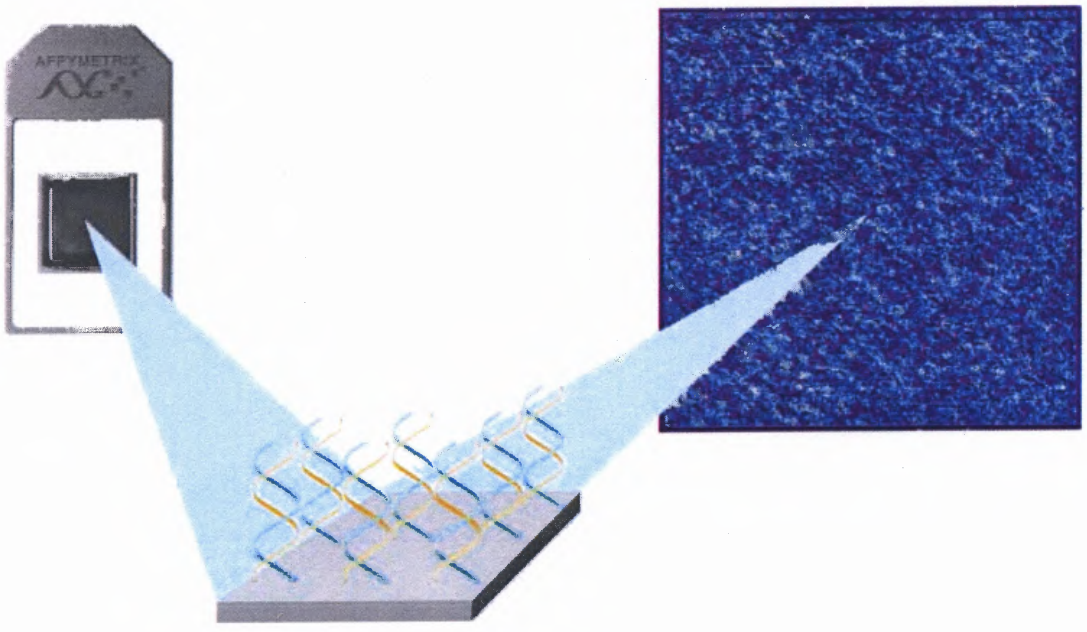

Figure A.3 Gene expression on a single GeneChip®

Figure A.3 depicts data from an experiment showing the expression of thousands of genes on a single GeneChip® probe array. Figure A.4 illustrates photolithography; GeneChip® probe arrays are manufactured through a unique and robust process, a combination of photolithography and combinatorial chemistry.

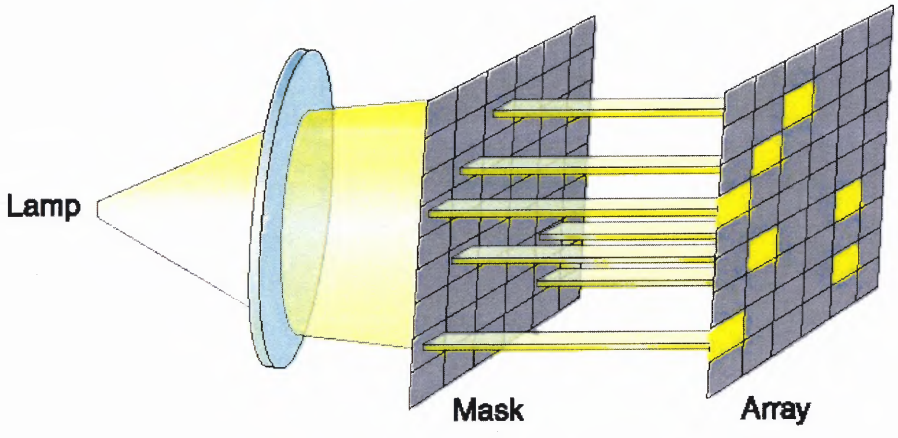

Figure A.4 Photolithography

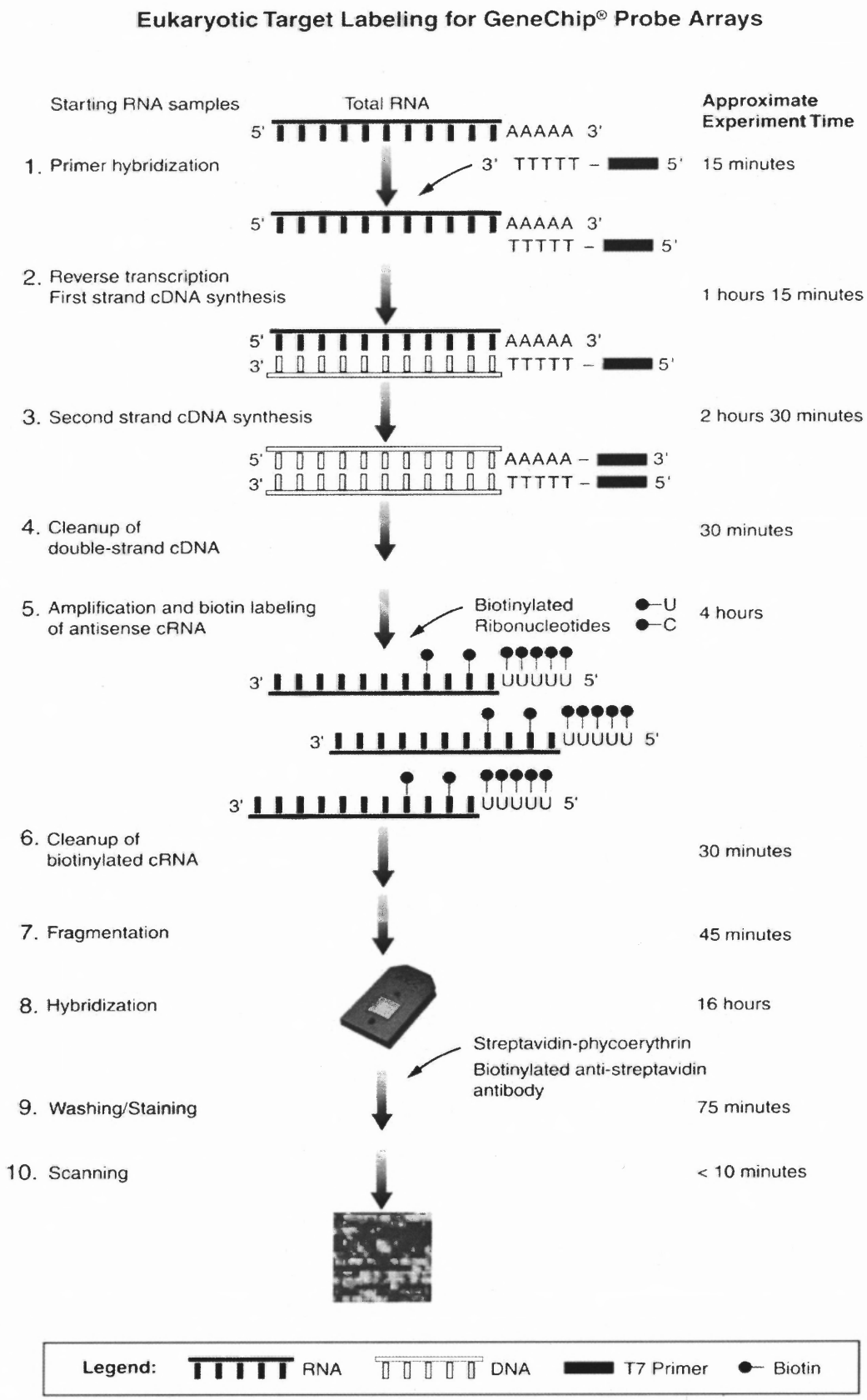

Figure A.5 Overview of eukaryotic target labeling for GeneChip® expression arrays

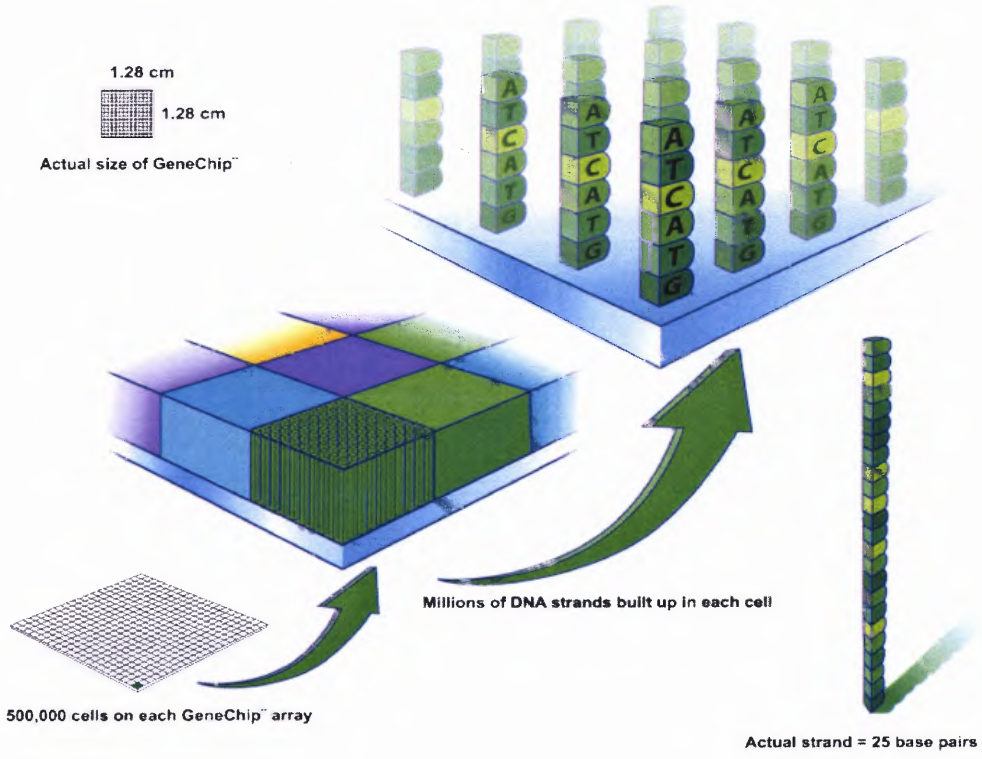

Figure A.6 A single feature on an Affymetrix GeneChip®

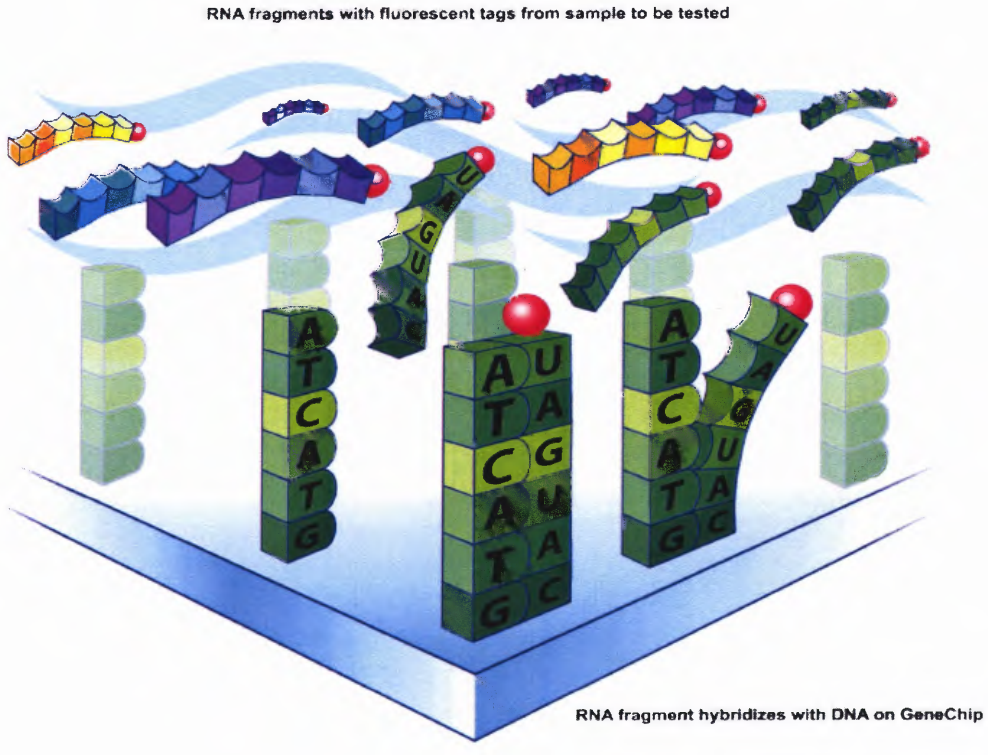

Figure A.7 Hybridization of tagged probes to Affymetrix GeneChip®

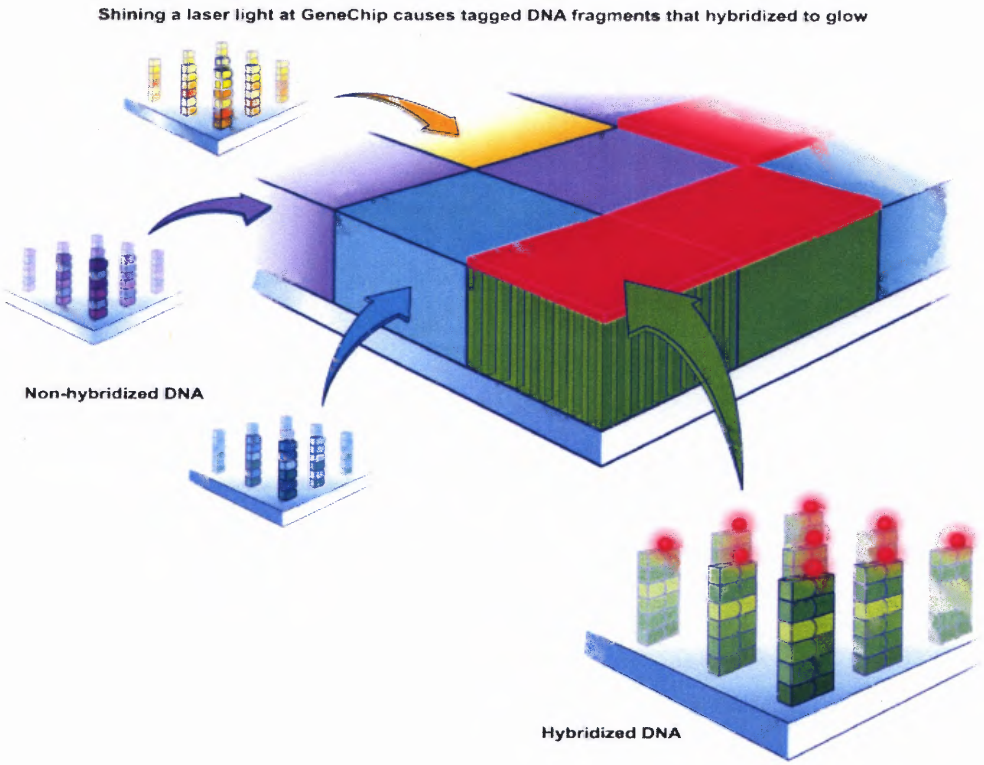

Figure A.8 Scanning of tagged and un-tagged probes on an Affymetrix GeneChip®

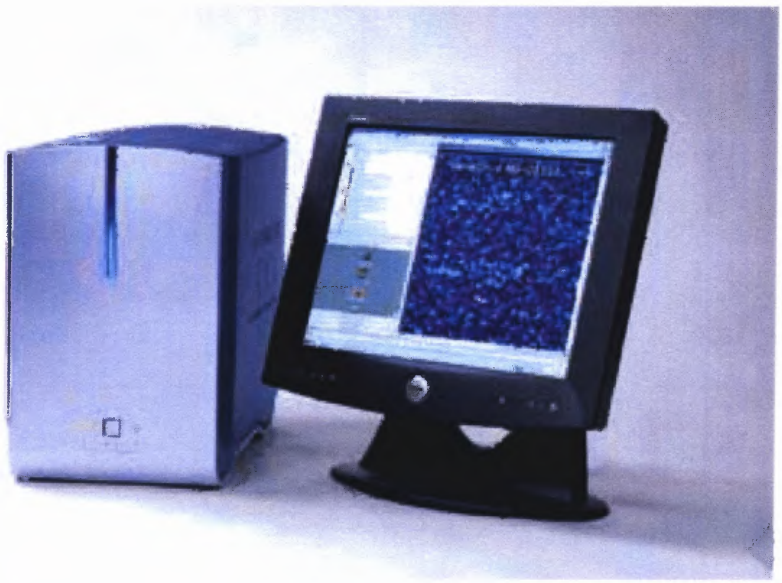

Figure A.9 Affymetrix GeneChip® Scanner 3000 with workstation

APPENDIX B

AFFYMETRIX PROBE SELECTION DETAILS

This appendix provides probe selection and chip design information from Affymetrix®. It is sourced from Affymetrix's Gene Expression Monitoring Technical Note.[35]

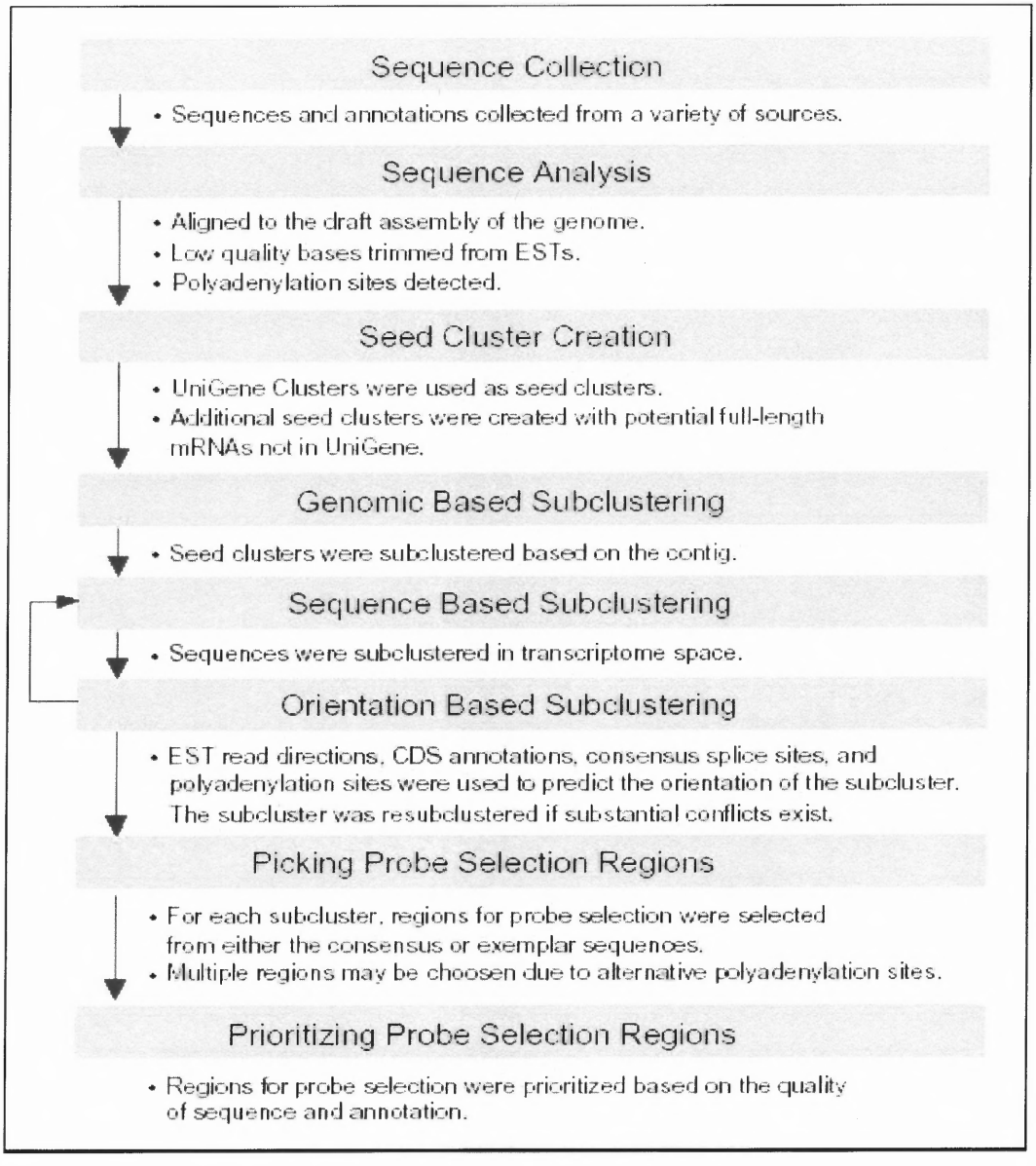

Figure B.1 Sequence selection for Affymetrix microarrays

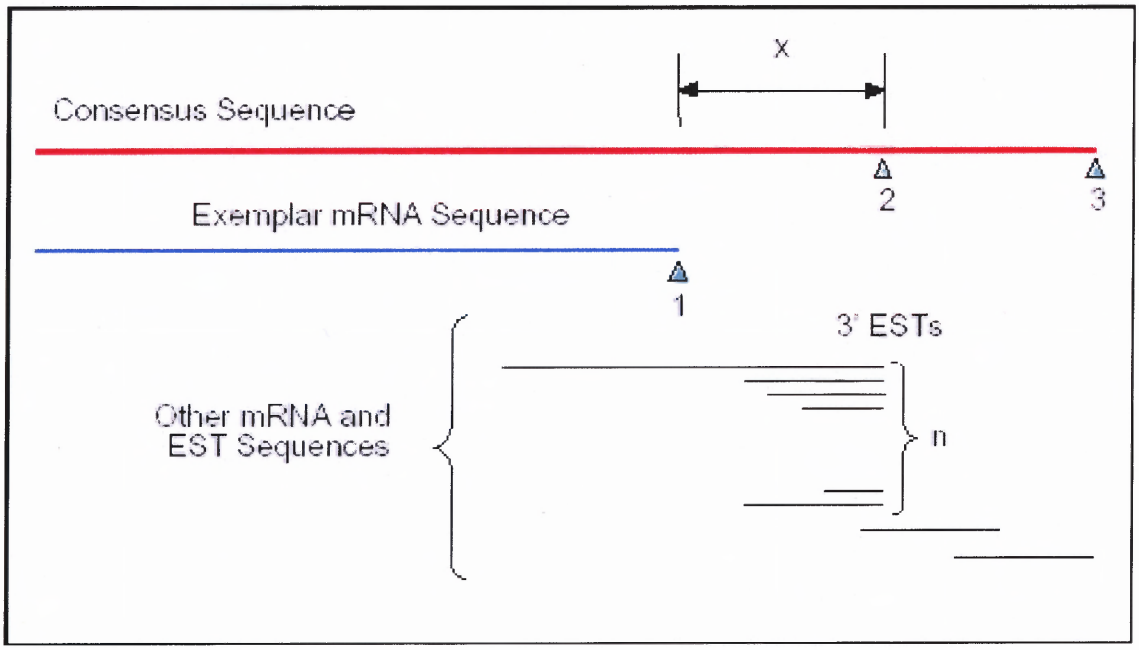

Figure B.2 Multiple probe selection regions

If the full-length exemplar has 3' UTR, then probes are picked from the 600 bp region ending at 1. For the stack of 3' ESTs (annotated 3', or possessing polyadenylation site/signal), if $n > 8$, and $x > 400$, then probes are also picked from region 2. Otherwise, if $n < 8$ or if $x < 400$, then region 2 is skipped. If the exemplar has no 3' UTR sequence then region 1 is not picked. Consequently, if $n > 8$, region 2 is picked, otherwise region 3 is picked.

Table B.1 Differences Between U95 and U133 Sets

	U95 and HuGene FL	U133	Justification
Sequence Sources	UniGene, GenBank	Unigene, RefSeq, Genbank, dbEST, WUSTL, Golden Path Draft Assembly	Improved annotation, classification, and sequence quality
Sequence Curation	Filtered for repeats, vector	Repeats and vector screening, EST quality trimming	Avoid low quality EST sequence regions, thereby improving consensus sequence quality
Sequence Subclustering	Subcluster by similarity and orientation	Similarity, orientation, and genomic position	Reduces false clusters of homologs
Sequence Orientation	According to CDS annotation and EST read direction	Genomic sequence, poly-A prediction, CDS, and EST read direction	Improves orientation calls by using sequence-based methods in addition to annotations.
Sequence Selection Region	600 base region from end of consensus	600 base region from end of exemplar full length mRNA or consensus. Multiple poly-A sites selected.	Full-length exemplars may be of higher sequence quality than consensus. Multiple poly-A sites improve sensitivity for alternative transcripts.
Probe Quality	Heuristic rules (e.g. not more than 10 A's in a probe). Probe quality is assessed as a binary (yes/no) function.	Thermodynamic multiple linear regression model predicts intensity of probes. Probe quality assessed on a continuous scale.	Improve selection of probes that hybridize well to the correct target and reduce non-specific cross hybridization
Probe Uniqueness	Probes which have 21 or more bases out of 25 matching targets expected to be in RNA samples are too similar, and will be avoided.	Probes which have two 8-mer matches, including at least one 12-mer match will be avoided.	Minimize specific cross hybridization to similar targets from unintended sequences.
Probe Spacing	Approximately equally spaced.	Spacing weighted to favor high quality and independent probes.	Ensure multiple probes give independent measurements of the target.
Number of Probes	16-20	11	Combined with algorithm and probe quality improvements, allows greater information density without reduction in information quality.
Probe Set Annotation	_s, _g, _f, _n, _r, _i	_s, _x Discontinued: _r, _i, _n Transformed: _g→_s _f→_x	Non-unique probe set types were simplified and adjusted to account for improvements in probe selection rules.
Feature Size	20 microns	18 microns	Allow greater information density without reduction in information quality

REFERENCES

1. C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379-423 and 623-656, July and October, 1948.

2. International Society of Computational Biology, http://www.iscb.org, March, 2003

3. GenBank Statistics, NCBI, http://www.ncbi.nlm.nih.gov/Genbank/genbankstats .html, February, 2003.

4. PDB Current Holdings, PDB, http://www.rcsb.org/pdb/holdings.html, April, 2003.

5. A Science Primer, NCBI, http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics .html, March, 2003.

6. NCBI, http://www.ncbi.nlm.gov, March, 2003.

7. H. H. Rashidi and L. K. Buehler, *Bioinformatics Basics: Applications in Biological Science and Medicine*, CRC Press, 2000.

8. Our Mission, NCBI at a Glance, NCBI, http://www.ncbi.nlm.nih.gov/About/glance/ ourmission.html, March, 2003.

9. EBI, http://www.ebi.ac.uk, March, 2003.

10. M. Shena et al., "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, vol. 270, issue 5235, pp. 467-470, October 20, 1995.

11. D. Shalon et al., "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization," *Genome Res.*, vol. 6, issue 7, pp. 639-645, July, 1996.

12. E. M. Southern, "Detection of specific sequences among DNA fragments separated by gel electrophoresis," *J. Mol. Biol.*, vol. 98, pp. 503-517, 1975.

13. S. P. A. Fodor et al., "Light-directed, spatially addressable parallel chemical synthesis," *Science*, vol. 251, pp. 767-773, 1991.

14. K. Lindblad-Toh et al., "Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays," *Nature Biotechnol.*, vol. 18, pp. 1001–1005, 2000.

15. K. Lindblad-Toh et al., "Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse," *Nature Genet.*, vol. 24, pp. 381–386, 2000.

16. C. Lock et al., "Gene-microarray analysis of multiple sclerosis lesions yields new targets validated in autoimmune encephalomyelitis," *Nature Med.*, vol. 8, pp. 500–508, 2002.
17. R. Miki et al., "Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays," *Proc. Natl Acad. Sci.*, vol. 98, pp. 2199–2204, 2001.
18. V. R. Iyer et al., "The transcriptional program in the response of human fibroblasts to serum," *Science*, vol. 283, pp. 83–87, 1999.
19. A. Alizadeh et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503–511, 2000.
20. M. Bittner et al., "Molecular classification of cutaneous malignant melanoma by gene expression profiling," *Nature*, vol. 406, pp. 536–540, 2000.
21. S. M. Dhanasekaran et al., "Delineation of prognostic biomarkers in prostate cancer," *Nature*, vol. 412, pp. 822–826, 2001.
22. T. R. Golub et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, 1999.
23. Hedenfalk et al., "Gene-expression profiles in hereditary breast cancer," *N. Engl. J. Med.*, vol. 344, pp. 539–548, 2001.
24. M. A. Shipp et al., "Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning," *Nature Med.*, vol. 8, pp. 68–74, 2002.
25. T. Sorlie et al., "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proc. Natl Acad. Sci.*, vol. 98, pp. 10869–10874, 2001.
26. L. J. van't Veer et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530–536, 2002.
27. M. Volm et al., "Expression profile of genes in non-small cell lung carcinomas from long-term surviving patients," *Clin. Cancer Res.*, vol. 8, pp. 1843–1848, 2002.
28. W. Lo et al., "A novel chromatin immunoprecipitation and array (CIA) analysis identifies a 460-kb CENP-A-binding neocentromere DNA," *Genome Res.*, vol. 11, pp. 448–457, 2001.
29. M. F. Shannon et al., "Transcription: Of chips and ChIPs," *Science*, vol. 296, pp. 666–669, 2002.

30. J. R. Pollack et al., "Genome-wide analysis of DNA copy-number changes using cDNA microarrays," *Nature Genet.*, vol. 23, pp. 41–46, 1999.

31. M. K. Kerr and G. A. Churchill, "Statistical design and the analysis of gene expression microarray data," *Genet Res.*, vol. 77, pp. 123–128, 2001.

32. S. A. Ahrendt et al., "Rapid p53 sequence analysis in primary lung cancer using an oligonucleotide probe array," *Proc. Natl Acad. Sci.*, vol. 96, pp. 7382–7387, 1999.

33. Affymetrix, http://www.affymetrix.com/technology/manufacturing/index.affx, April, 2003.

34. M. S. Gelfand et al., *Nucleic Acids Research*, vol. 27, pp. 301-302, 1999.

35. Gene Expression Monitoring Technical Note, Array Design for the GeneChip® Human Genome U133 Set, Affymetrix, http://www.affymetrix.com, April, 2003.