2006

Cross Chip Probe Matching Tool: A Tool For Linking Probes From Microarrays Within And Across Species

Ruchi Ghanekar
University of Alabama at Birmingham

CROSS CHIP PROBE MATCHING TOOL: A TOOL FOR LINKING PROBES FROM MICROARRAYS WITHIN AND ACROSS SPECIES

by

RUCHI GHANEKAR

GARY GRIMES, COMMITTEE CHAIR
GRIER PAGE
MURAT TANIK

A THESIS

Submitted to the graduate faculty of The University of Alabama at Birmingham,
in partial fulfillment of the requirements for the degree of
Master of Science

BIRMINGHAM, ALABAMA

2006

CROSS CHIP PROBE MATCHING TOOL: A TOOL FOR LINKING PROBES FROM MICROARRAYS WITHIN AND ACROSS SPECIES

RUCHI VIJAY GHANEKAR

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING

ABSTRACT

Traditionally, but not exclusively, a microarray experiment is conducted on a single type of microarray. However, different labs often use different chip types, and new versions of chips may be developed. Comparing experiments can be problematic in such cases, as it is not always clear which probe sets correspond to the same gene across experiments. The goal of this research was to develop a tool that allows investigators to identify probe sets that correspond to the gene within and across species.

The Cross Chip Probe Matching Tool (CCPMT) makes use of the probe sequences for all features on arrays from corporate literature, white papers, and corresponding documentation. Probes were assigned to a particular gene by using the Basic Local Alignment Search Tool (BLAST) developed at the National Center for Biotechnology Information (NCBI). The sequence alignment was achieved using database resources from The Arabidopsis Information Resource (TAIR), microarray vendor sequences, and The Institute for Genomic Research (TIGR) gene indices. For CCPMT the identification of the orthologous genes across species was performed by evaluating Eukaryotic Gene Orthologs (EGO) defined by TIGR. CCPMT allows investigators to combine data across multiple chip types and make better multi propositional inferences, as well as aiding in the meta-analysis of experiments.

ACKNOWLEDGMENTS

I want to express my gratitude towards my parents and my sister for encouraging me to pursue my higher education abroad and for making my dream come true.

I would like to thank Dr. Grier Page for providing me with the research assistantship opportunity at the Section on Statistical Genetics (SSG) at the University of Alabama at Birmingham (UAB). I express my appreciation for all of the help and advice Dr. Page has provided.

I want to take this opportunity to thank my advisor Dr. Gary Grimes for his guidance during my graduate studies. I also wish to acknowledge Dr. Murat Tanik for being on my committee.

I wish to thank everyone on the Software Developers team at SSG for providing me with software assistance and helping me learn a lot.

Special thanks go to my husband Amol Vaidya and my in-laws for always being there for me and encouraging me throughout my graduate studies.

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION

A microarray experiment is usually conducted on a single type of microarray. However, different labs often use different chip types, and new versions of chips may be developed. Comparing experiments can be problematic in such cases, as it is not always clear which probe sets correspond to the same gene across experiments. There are also cases where an investigator wants to compare microarray probes across two species and the common gene annotation identifiers are not provided by the microarray vendors. These limitations can hinder experiments that require annotation of microarray data.

The Cross Chip Probe Matching Tool (CCPMT) is developed to solve the annotation problems in microarrays. It is a tool that allows investigators to identify probe sets that correspond to the gene within and across species.

The current version of CCPMT includes two plant species, Arabidopsis *thaliana* and Poplar. CCPMT helps in probe and gene level annotations in both the plant species by using such gene identifiers as Arabidopsis Genome Initiative (AGI) and Tentative Consensus (TC). CCPMT also allows investigators to compare probe level data across the two plant species. For annotating across species, CCPMT queries on the Eukaryotic Gene Ortholog (EGO) identifiers. CCPMT makes use of the gene expression data on the microarray chips from multiple vendors to map expression data from one microarray chip to another.

CHAPTER 2

LITERATURE REVIEW

CCPMT is a tool to allow investigators to identify probe sets that correspond to a paralogous gene within the same plant species and an orthologous gene across plant species. Below is a brief summary of the existing annotation tools similar to CCPMT.

- KARMA (Keck ARray Manager and Annotator).
- RESOURCERER.
- GeneSeer.

A. KARMA

KARMA [6] is a web server application for comparing and annotating heterogeneous microarray platforms and was developed at Yale University. Data for the KARMA database was collected from these sources:

- LocusLink.
- SwissProt.
- Gene Ontology (GO).

1) Key Features: The key features of KARMA are as follows:

- Allows for cross species comparison of gene sequences.

- Allows comparison of common array platforms.
- Allows comparison of custom array platforms.
- Utilizes the UniGene Cluster identifiers to compare within species.
- Uses the HOMOLOGENE data set from NCBI to compare across species.
- Uses GenBank Accession numbers as reference IDs to annotate the genes.
- Has its own set of resources including DRAGON, SOURCE, and Unchip that allow the user to submit a set of DNA elements laid down on chips/arrays and then return the latest annotation describing those elements.

2) KARMA Limitations: The following list contains some of the limitations of KARMA and also highlights the strengths of CCPMT as compared with KARMA:

- The output of KARMA cannot be viewed as dynamic web pages. The web link for the results is sent through email. CCPMT not only sends the results as an email attachment but also displays the results in the web browser.
- KARMA does not allow querying for annotation information within the same array. CCPMT provides annotations mapping within the same array. KARMA cannot be used in cases where a probe set on a chip would map to another probe set on the same chip.

B. RESOURCERER

RESOURCERER [7] is a database for annotating and linking microarray resources within and across species and was developed at TIGR.

1) Key Features: The key features of RESOURCERER are as below.

- RESOURCERER, a microarray-resource annotation and cross-reference database, was built using the analysis of ESTs and gene sequences provided by the TGI and TIGR Orthologous Gene Alignment (TOGA) databases (now called EGO).
- RESOURCERER provides comparison between resources from the same species using TGI, UniGene, LocusLink, or RefSeq and across species using the EGO database.
- ESTs are downloaded daily from the dbEST database and are cleaned to remove the unwanted non-coding part.
- Cleaned EST and gene sequences are compared pair-wise to identify overlaps using Basic Local Alignment Search Tool (BLAST) [8].
    - BLAST was developed by Altschul et al. (1990). It is a search algorithm that is used to search protein or DNA databases. BLAST is best used for sequence similarity searching.
- The mapped sequences within each cluster are assembled to produce TC sequences that are loaded into species specific databases.
- TGI can be searched by TC number, the GenBank accession number of any EST contained within the dataset, or any gene used to build the index.
- To identify orthologs and paralogs, RESOUCERER uses high-stringency pair-wise sequence searches and a reflexive, transitive closure process to associate sequence-specific best hits with gene sequences and assembled ESTs.

2) RESOURCERER Limitations: The following list contains some of the limitations of the tool RESOURCERER and also highlights the strengths of CCPMT as compared with RESOURCERER.

- In RESOURCERER there are no options for the user to enter individual probe set IDs/gene IDs. CCPMT offers the flexibility to compare entire arrays as well as to enter individual probe set IDs/gene IDs.
- RESOURCERER mappings are based on TIGR annotations such as TCs, EGO, and TGI. CCPMT mappings are based on the gene ID level as well as the probe set ID level. In CCPMT for Arabidopsis the mappings are based on AGI as well as TC and EGO identifiers.
- RESOURCERER currently does not support any plant species. CCPMT is one of the first annotation tools exclusively for plant species.

C. GeneSeer

GeneSeer [9] is a software tool for gene names and genomic resources and was developed by the Cold Spring Harbor Laboratory.

1) Key Features: The key features of the tool RESOURCERER are as follows:

- In order to locate genomic resources for a given gene using a familiar name, GeneSeer uses the GenBank accession ID.

- GeneSeer retrieves and stores synonyms for the model organisms including human, mouse, and Arabidopsis, from the following sources.
    - GenBank.
    - FlyBase.
    - ExPASy.
    - HUGO.
    - ENSEMBL.
    - GO.
- GeneSeer has built a set of mRNA accessions that includes the known genes and expert-curated cDNAs called Set of FASTA Representatives (SOFAR).
- Gene lists are created using the above listed resources. Using BLAST, they calculate a matrix of similarity scores.
- Each gene name gets mapped to a SOFAR representative. If an accession or EST name is submitted to the system and it is not recognized, then GeneSeer downloads the sequence and uses BLAST against SOFAR to identify its name.
- To identify homologs across species and present a tree view, matrixes of similarity scores based on BLAST are pre calculated for all pairs of SOFAR proteins in the database. These matrixes are used to generate clusters of related proteins and the result of the matrix can be viewed in the phylogenetic tree structure.

2) GeneSeer Limitations: The following list contains some of the limitations of the tool GeneSeer and also highlights the strengths of CCPMT as compared with GeneSeer.

- GeneSeer does not provide information based on microarrays but on individual annotation ID or sequences, e.g., HUGO ID and protein sequences, respectively. Thus, GeneSeer cannot be used to compare arrays from various microarray vendors. CCPMT not only supports individual annotation ID based querying but also supports queries based on microarray probe set data.

D. Key Features of CCPMT

- Different labs often use different chip types, and new versions of chips may be developed, thus comparing experiments can be problematic. In such cases it is not always clear which probe sets correspond to the same gene across experiments. CCPMT is an annotation tool for microarrays that can overcome this problem of mapping probe sets to gene across experiments.
- CCPMT is one of the first tools exclusively meant for plant species annotations.
- A web based tool, CCPMT can be accessed through the internet.
- Investigators can query at probe level with probe set IDs or even for gene level data with the following gene identifiers:
    - AGI.
    - EGO.
    - TC.
- In CCPMT, an investigator can either enter individual or multiple probe set/gene identifiers (separated by commas) in the textbox to query the CCPMT database.

- Checkboxes for microarray vendors provide the option of selecting multiple arrays while querying the CCPMT database.
- It is possible to carry out a one-to-one comparison of arrays in CCPMT.
- Results are displayed immediately in the web browser.
- Results are sent through email in a machine readable file format.
- CCPMT has a flexible database design and in future applications, the database could be modeled for mammalian data.

CHAPTER 3

INTRODUCTION TO GENETICS

A. Microarrays and Genetics

With the advancements in the field of genetics, more and more genetic data is available in the public domain. In recent years many of the plant genomes have been decoded. Arabidopsis *thaliana* and Poplar are some of the plant species that have had their genomes decoded.

Microarray technology is a tool one can use to study large number of genes of interest and their interactions with each other [1]. This technology uses a robot to precisely apply tiny droplets containing functional deoxyribonucleic acid (DNA) to glass slides. Researchers then attach fluorescent labels to DNA from the cell they are studying. The labeled probes are allowed to bind to complementary DNA (cDNA) strands on the slides. The slides are put into a scanning microscope to measure the amount of specific DNA material in the array.

Microarrays build bridges on which biologists and clinicians can meet to understand, diagnose, and treat diseases. Microarray based technologies allow for rapid access to molecular pathways, more precise diagnosis and prognosis of disease, better understanding of drug action, and the ability to better define therapeutic strategies.

DNA has the shape of a double helix and contains the necessary information for the biological growth of organisms. DNA is passed on from parent to offspring, thus

passing genetic traits through the progeny. DNA is made up of four building blocks, i.e. adenine (A), cytosine (C), guanine (G), and thymine (T). Each strand of DNA contains the chain of these chemicals, and such a chain is called as nucleotide chain. In the double helix form the bases A and T and the bases G and C form bonds. The process of converting a DNA to messenger ribonucleic acid (mRNA) is called transcription. RNA contains the bases adenine, uracil (U), cytosine, and guanine.

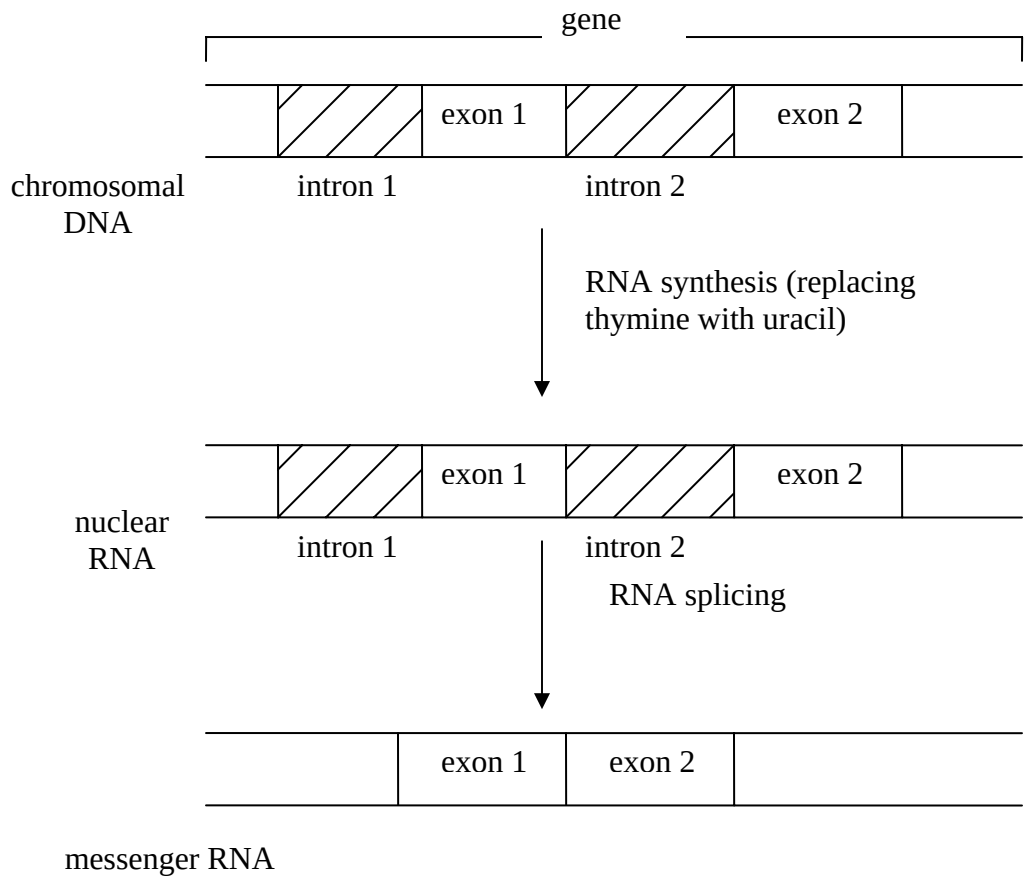

Fig.1. Steps during DNA transcription.

A gene (a chain of nucleotides) contains coding and non coding regions as shown in Fig.1. The coding regions are called exons and the non coding regions are called

introns. Genes also contain regions called as un-translated regions (UTRs). UTRs are sequences on the ends of mRNA and are not translated into protein during the translation process.

During the first step of RNA synthesis, the DNA helix uncoils, the base thymine is replaced with uracil, and a nuclear RNA is formed. An important step during transcription is RNA splicing. In many genes the DNA coding regions, i.e., exons are interrupted by long stretches of non coding DNA called introns. During the second step of transcription the introns are removed by a process called RNA splicing. The edited sequence is called mRNA. The mRNA carries the gene's instructions and dictates the protein production by ribosomes. cDNA are synthesized by complementing the bases in a given strand of mRNA. cDNA represents the parts of gene that are expressed in a cell to produce a protein.

In CCPMT, the EGO database, along with other TIGR identifiers, has been used to compare mappings across plant species. The EGO is a database for orthologous genes in eukaryotes. EGO is generated by pair-wise comparison between the tentative consensus (TC) sequences that comprise the TIGR gene indices (TGI) from individual organisms. The reciprocal pairs of the best match were clustered into individual groups, and multiple sequence alignments were displayed for each group [4].

TC sequences are created by assembling expressed sequence tags (ESTs) into virtual transcripts. In some cases, TCs contain full or partial cDNA sequences obtained by classical methods. TCs contain information on the source library and abundance of ESTs and in many cases represent full-length transcripts. Alternative splice forms are built into separate TCs. TCs are actual assemblies, with a consensus sequence, and not

simply clusters of overlapping sequences. ESTs are partial, single-pass sequences from either end of a cDNA clone. The EST strategy was developed to allow rapid identification of expressed genes by sequence analysis [5].

B. Paralogs and Orthologs

CCPMT helps investigators find microarray probe set level information within the same species, as well as across plant species.

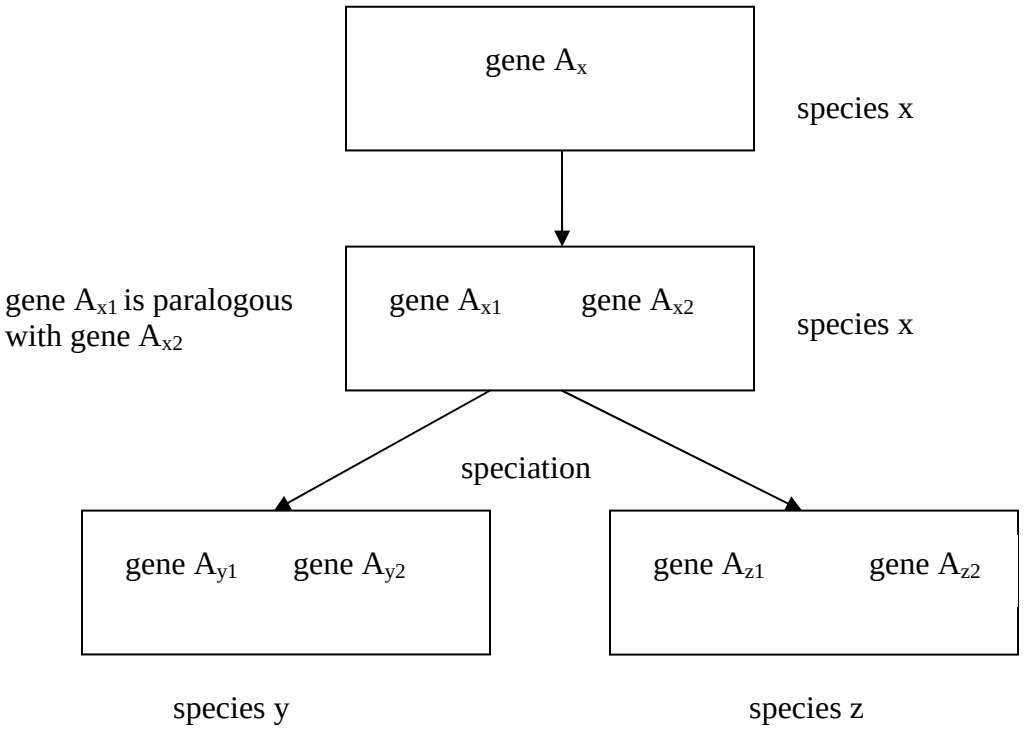

Fig.2. Paralogous and orthologous genes.

Fig. 2 illustrates that two genes are called paralogs if they have evolved as a result of gene duplication within the same species. In Fig. 2 the genes A_{x1} and A_{x2} are said to be

paralogs, as they have descended from a common gene A_x and are within the same species. If two genes during evolution have descended from a common gene and now, after speciation, exist in two different species, then such genes are orthologs of one another. In Fig. 2 the gene A_{y1} is orthologous to A_{z1}, as both of them have evolved from a common gene A_x , but they exist in different species, namely species y and species z.

For Arabidopsis, CCPMT makes use of AGI IDs to find the paralog genes, and for Poplar the TIGR TCs are used to find paralogous genes. CCPMT makes use of TIGR EGO data to map probes across species.

C. Plant Species in CCPMT

CCPMT makes use of the gene expression data on the microarray chips from multiple vendors to map expression data from one microarray chip to another. Arabidopsis *thaliana* was selected as the first plant species to be included in the CCPMT tool database.

1) Arabidopsis thaliana: Arabidopsis thaliana is the most researched plant in the field of plant genomics [2]. Arabidopsis, as it is commonly known, belongs to the mustard family (brassicaceae). Arabidopsis is neither economically important nor an agronomically significant plant, but it offers great advantages in the field of genetics and molecular biology. The reasons for selection of Arabidopsis are as follows:

- Arabidopsis genome is comparatively smaller than other plant species and was sequenced in the year 2000.

- Extensive genetic and physical maps of all five Arabidopsis chromosomes are available.
- The Arabidopsis life cycle is just six weeks germination to mature seed, and the plants can be cultivated in restricted spaces.
- The Research organizations TAIR, and TIGR and other laboratories have an extensive research community.
- Many microarray vendors have Arabidopsis as one of the plant species on their chip and probe sequences, and annotation information is publicly available.

The following microarray vendors were identified for Arabidopsis:

- Affymetrix Arabidopsis Genome (8k).
- Affymetrix Arabidopsis Genome ATH1 (25K).
- AFGC Arabidopsis Array.
- Operon Arabidopsis Genome Oligo.
- Agilent Arabidopsis 2 Oligo.
- CATMA - Complete Arabidopsis Transcriptome MicroArray.

2) Poplar: Black cottonwood *(Populus trichocarpa)* is a hardwood tree and has the distinction of being the largest of the American Poplars and also the largest hardwood tree in western North America. The genome of the Poplar is about 500 million letters of genetic code. It is the first hardwood tree whose DNA sequence was decoded, because Poplar's genome was relatively compact as compared to other hardwood trees, like pine. Thus the Poplar tree makes an ideal model system for trees.

The Poplar genome is divided into 19 chromosomes, and the Poplar genome is four times larger than the Arabidopsis genome. For CCPMT, Poplar was selected as the second plant, because about 27% of the Poplar's sequences had significant homology to Arabidopsis protein-coding sequences [3].

In Arabidopsis, the standard gene nomenclature is known as the Arabidopsis Genome Initiative (AGI) ID. An effort is currently being made to have a standard gene nomenclature for the Poplar gene on the genome at the Joint Genome Institute (JGI). Since Poplar does not have a universal gene nomenclature, the microarray vendors and research organizations follow their own set of annotations. In CCPMT the TIGR EGO identifiers are used to map Poplar probe sets with the corresponding Arabidopsis probe sets. In CCPMT, the Affymetrix Poplar Genome Array data was selected as the microarray data source for the plant species Poplar.

CHAPTER 4

IMPLEMENTATION DETAILS

This chapter explains the steps taken for building the CCPMT tool. The first task was to select the plant microarray data for CCPMT. It was decided to start with Arabidopsis plant species, as it is the most researched plant species and has a well studied genome. After Arabidopsis, Poplar was chosen as the next plant, as Poplar's genome had around 27% overlapping regions with the Arabidopsis genome.

A. Data Import

1) Arabidopsis: The data in CCPMT is from microarray vendors that provide the probe set level (transcriptome) data freely in the public domain. The microarray vendors for Arabidopsis are as follows:

- Affymetrix.
    - Arabidopsis Genome (8K).
        - Commonly referred as AG array.
    - Arabidopsis Genome ATH1 (25K).
        - Commonly referred as ATH1.

  - Affymetrix provides two files for each of their arrays. The first file provides all of the target sequences (nucleotide sequences) in that array, and the second file provides probe set level annotation information.

- Agilent
  - Agilent is a commercial microarray vendor; they do not provide the nucleotide sequences for their probe sets, but they do provide a file with probe set annotations (mapping of the probe sets with the AGI gene IDs). The mapping information provided directly by Agilent was used.
- Arabidopsis Functional Genomics Consortium (AFGC) array
  - The data for the AFGC array was downloaded from the TAIR website. TAIR also provided the AFGC probe sequences and an annotation file with probe set IDs and corresponding AGI mappings.
- Complete Arabidopsis Transcriptome MicroArray (CATMA) array
  - The data for CATMA arrays was downloaded from the TAIR website. TAIR also provided the CATMA probe sequences and an annotation file with probe set IDs and corresponding AGI mappings.
- Operon
  - The Operon data was downloaded from their official website. Operon provided a mega file that had the probe set IDs along with mappings of the corresponding AGI and their nucleotide sequences.

Table 1 contains the reference websites from which the microarray data for Arabidopsis was downloaded.

TABLE 1
REFERENCE WEBSITES FOR ARABIDOPSIS MICROARRAY DATA

	Affymetrix-AG	Affymetrix-ATH1	Operon
Target Sequence File location	http://www.affymetrix.com/support/technical/byproduct.affx?product=atgenome1	http://www.affymetrix.com/support/technical/byproduct.affx?product=arab	http://omad.operon.com/download/index.php
Vendor provided Probe-AGI mapping location	http://www.affymetrix.com/support/technical/byproduct.affx?product=atgenome1	http://www.affymetrix.com/support/technical/byproduct.affx?product=arab	http://omad.operon.com/download/index.php
	CATMA	AFGC	Agilent
Target Sequence File location	ftp://ftp.arabidopsis.org/home/tair/Microarrays/CATMA/	ftp://ftp.arabidopsis.org/home/tair/Microarrays/AFGC/	NA (do not provide sequence files)
Vendor provided Probe-AGI mapping location	ftp://ftp.arabidopsis.org/home/tair/Microarrays/CATMA/	ftp://ftp.arabidopsis.org/home/tair/Microarrays/AFGC/	http://www.chem.agilent.com/Scripts/PDS.asp?lPage=37068

2) Poplar: The Poplar species genome is not as widely researched as the Arabidopsis genome. Poplar does not have a universal gene ID annotation that all microarray vendors follow uniformly. Currently CCPMT has only Affymetrix Poplar Genome Array data.

The Poplar target sequence file and the annotation file were downloaded from the Affymetrix website, which is referred to in Table 2.

TABLE 2
REFERENCE WEBSITE FOR POPLAR MICROARRAY DATA

	Affymetrix Poplar Genome Array
Target Sequence File location	http://www.affymetrix.com/support/technical/byproduct.affx?product=Poplar
Vendor provided Annotation file	http://www.affymetrix.com/support/technical/byproduct.affx?product=Poplar

B. Data Preprocessing and Quality Control

This section deals with the various steps that were taken during the data preprocessing phase of CCPMT. Tools like NCBI BLAST were used for sequence alignment. Statistical software SAS and Microsoft Excel were used for quality control of the data. After passing through the quality control filters, the data was uploaded to a MySQL database.

Once all of the microarray data was downloaded, the next step was to decide either to use the vendor's data annotation (mapping between probe set and AGI IDs) or to get a new set of probe set and AGI mapping using the NCBI BLAST sequence alignment tool.

1) Arabidopsis: The microarray vendor's annotation files provide probe set level information. In the case of Arabidopsis all vendors provided the mappings between their probe sets and the corresponding AGI gene identifiers. For CCPMT it was decided to get

a new set of mappings between the probe sets and the corresponding AGI IDs. This mapping was accomplished using the NCBI BLAST blastn program. Blastn compares a nucleotide query sequence against a nucleotide sequence database. In BLAST the S score signifies the measure of similarity of the query to the sequence [8]. The term E value can be defined as the probability, that there is another alignment with a similarity greater than the given S score due to chance. Percent identity is defined as the extent of similarity between two sequences can in terms of percentage.

The aim of this research was to use BLAST for aligning the vendor sequence files against a sequence database and to compare these new set of mappings with the ones provided by the microarray vendors. The TAIR dataset was used as the database during sequence alignment.

The Affymetrix and Operon array sequences do not contain introns and UTRs, hence the AGI CDS [10] database at TAIR was used as the sequence database. The AGI CDS contains all of the Arabidopsis coding sequences but lack the introns and the UTRs. The AGI CDS dataset belongs to the TAIR6.0 release version and was released in November 2005. For the vendors AFGC and CATMA, the AGI Transcripts dataset was chosen as sequence database during sequence alignment using NCBI BLAST. The AGI Transcripts dataset includes all of the coding sequences from Arabidopsis and also contains the UTRs. This dataset does not contain the non coding regions, i.e., the introns. The AGI Transcripts dataset belongs to the TAIR6.0 release version and was released in November 2005.

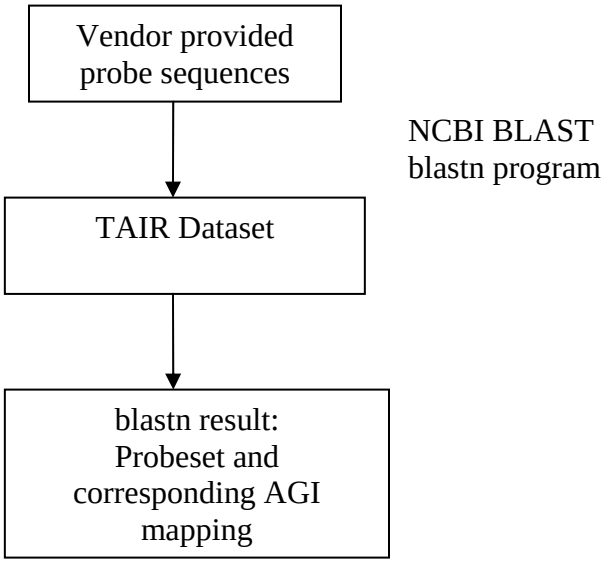

Fig.3. Arabidopsis BLAST workflow.

In the case of Arabidopsis, the probe set sequences from all other vendors (except Agilent) were aligned with the AGI sequences. The sequence alignment was performed using NCBI BLAST blastn program (Fig. 3).

For the Affymetrix AG and ATH1 arrays, the expected value cut-off was set at E^{-4} and the percent identity cut-off was set at 98%. Microarray vendors Operon, CATMA, and AFGC each had the blastn expected value cut-off set at E^{-4} .Each of the above vendors were tested for two sets of percent identity cuts-off, i.e., one at 75% and the next at 98%, Table 3 summarizes the mapping results.

The microarray vendors CATMA, Operon and AFGC had their percentage identity cut-offs lowered to 75% as all these three arrays are susceptible to polymorphisms (Table 3). With the percent identity lowered to 75%, an increase in the number of "*OneVendorBlastMany*" mappings was observed. Thus, for these three

vendors, reducing the cut-offs yielded a higher mapping results. In CCPMT, for vendors Operon and CATMA, it was decided to use the mappings percent identity 98%. For AFGC array the percent identity was set at 75%.

TABLE 3
COMPARISON BETWEEN THE MICROARRAY VENDOR AND CCPMT MAPPINGS

	AG	ATH1	Operon(75%)	Operon(98%)	CATMA(75%)	CATMA(98%)	AFGC(75%)	AFGC(98%)
Vendor Mapping Numbers	8297	22810	29954	29954	24576	24576	19108	19108
Nil Entries from vendor(No Mapping for these probes)	141	250	936	936	2969	2969	2823	2823
Exact Match	6932	20193	24352	26138	14788	19551	9979	*6413*[*]
Present-Vendor Absent-Blast	850	930	2062	2335	2008	2990	1497	10952
Absent-Vendor Present-Blast	0	0	0	0	0	0	8	1
One-Vendor Many-Blast	338	896	2353	480	3613	408	6521	368
Many-Vendor One-Blast	124	584	0	0	28	30	49	117

Keywords explaining terms from Table 3

Vendor Mapping Numbers: The total number of mappings provided by the microarray vendor.

Nil Entries from vendor *(No Mapping for these probes):* The number of nil mappings provided by the vendor

E.g.

Vendor probe ID	Vendor AGI mapping
At5784_at	---

Exact Match: Number of mappings that are same in both microarray vendor as well as blastn result

E.g.

Vendor probe ID	Vendor AGI mapping	BLAST probe	BLAST AGI mapping
123_at	AT4G14240	123_at	AT4G14240

Present-VendorAbsent-Blast: The number of mappings that are present in microarray vendor mapping list, but absent in blastn parsed result.

E.g.

Vendor probe ID	Vendor AGI mapping
123_at	AT4G14240, this means the probe 123_at is not present in the blastn result.

Absent-VendorPresent-Blast: The number of mappings that are absent in the microarray vendor mapping file, but present in blastn result. Ideally this number should be zero.

One-VendorMany-Blast: The blastn result shows the multiple number of AGI mapped to a probe set as compared with mapping done by the microarray vendors.

E.g.

Vendor probe ID	Vendor AGI mapping	BLAST probe	BLAST AGI mapping
123_at	agi_1; agi_2	123_at	agi_1; agi_3; agi_4; agi_2

Many-VendorOne-Blast: The blastn result maps a probe set to fewer number of AGI as compared with the microarray vendors mapping.

E.g.

Vendor probe ID	Vendor AGI mapping	BLAST probe	BLAST AGI mapping
123_at	agi_1; agi_2	123_at	agi_1

*Note**

- In case of CATMA and AFGC, these files were obtained from TAIR at ftp://ftp.arabidopsis.org/home/tair/Microarrays/
- The probe set to AGI mapping files for CATMA and AFGC have a timestamp of January 2006.
- The files that provide the probes and the sequences in FASTA format have a time stamp of April 2004.

- Thus there seems to be a difference in numbers when comparing the blastn results with the mapping file provided by TAIR.

2) Poplar:

The Poplar species does not have any "gene ID" as a universal gene annotation, so for CCPMT it was decided to use the EGO mappings. Poplar probe sets can, thus be mapped within the Poplar species using TC IDs and across plant species using the EGO database.

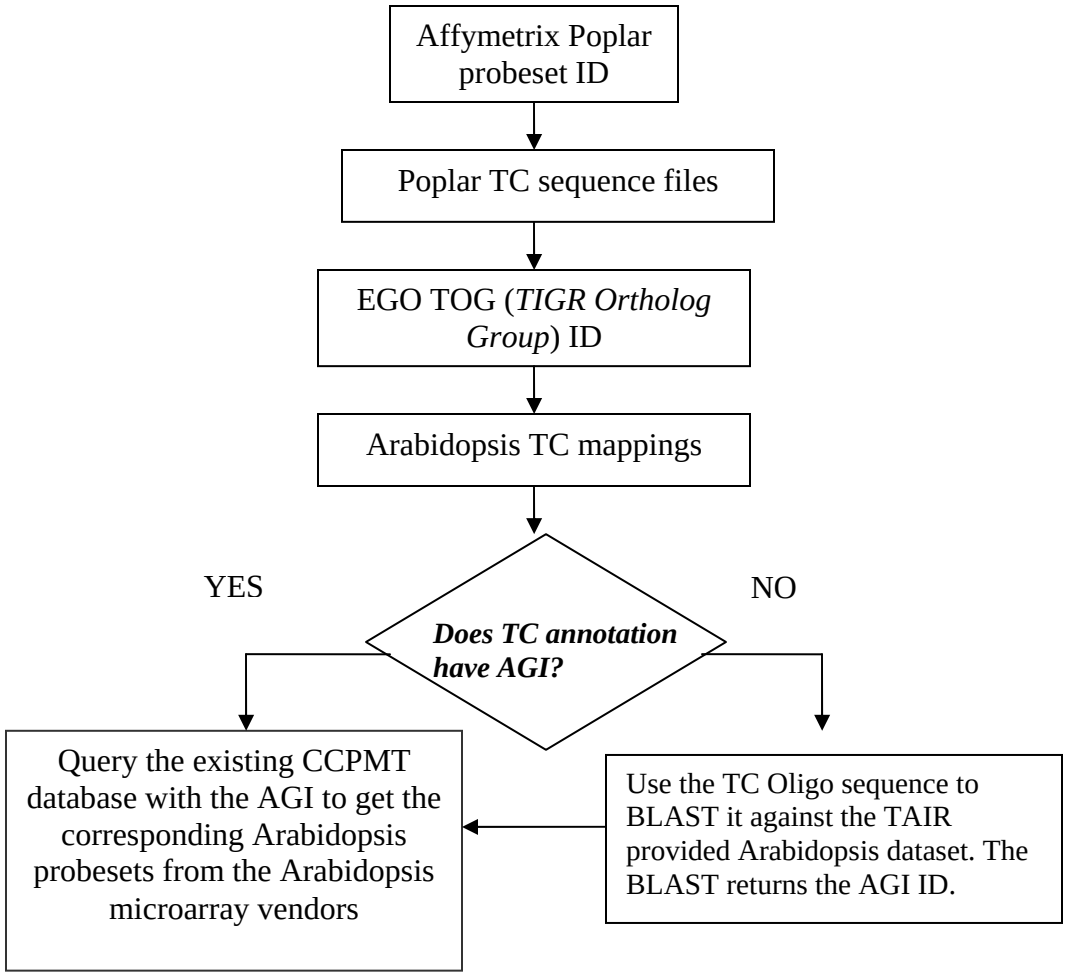

Fig.4. Poplar - Arabidopsis mapping.

The Poplar target sequences were sequence aligned with the TIGR Poplar TC dataset using the blastn program (Fig.4). The blastn expected value and percent identity cut-off were E^{-4} and 95%, respectively. TIGR also provides a file with a mapping of the EGO ID and the corresponding TCs for all species. From this file the mappings between EGO IDs and the corresponding Arabidopsis and Poplar TCs were parsed. In the future any plant species having an EGO ID can be easily incorporated into CCPMT.

The next step was to get a mapping between the Arabidopsis TCs and their corresponding AGI IDs. This was achieved by using the Arabidopsis TC sequences (TIGR provides this file) and sequence aligning it with the TAIR “AGI Transcripts” dataset. This sequence alignment produces a file that maps the Arabidopsis TCs with the corresponding Arabidopsis AGI IDs.

Table 4 explains the mapping of the Affymetrix Poplar Genome Array with the Affymetrix AG and Affymetrix ATH1 arrays.

TABLE 4
AFFYMETRIX POPLAR ARRAY MAPPED WITH ARABIDOPSIS AFFYMETRIX ATH1 AND AG ARRAYS

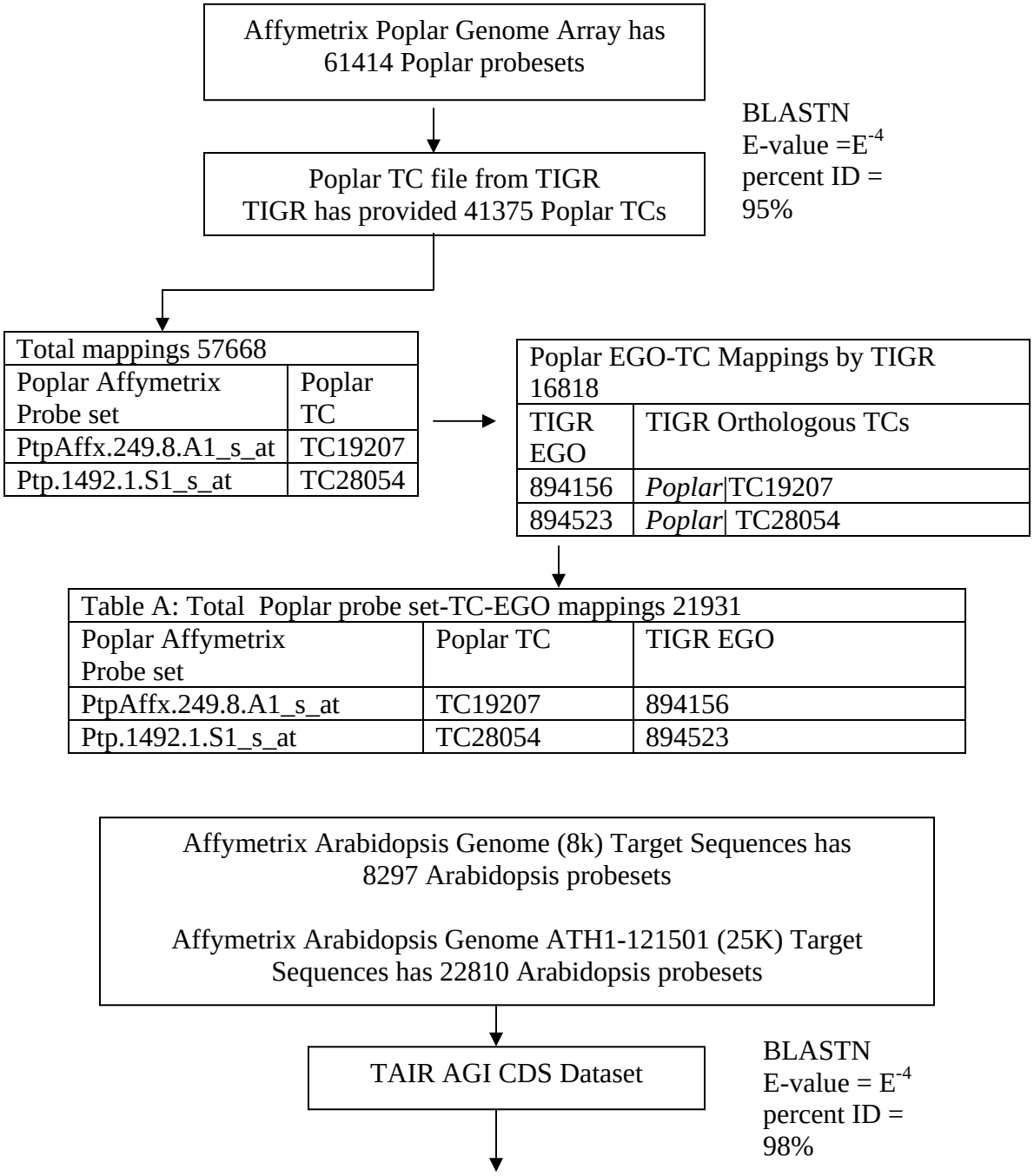

Table B: Total mappings of AG – AGI 7998 Total mappings of ATH1 – AGI 23664		
Affymetrix Arabidopsis Genome (AG)	Arabidopsis Affymetrix ATH1 probe set	AGI
12936_s_at	264474_s_at	AT5G38410
12752_s_at	254386_at	AT4G21960

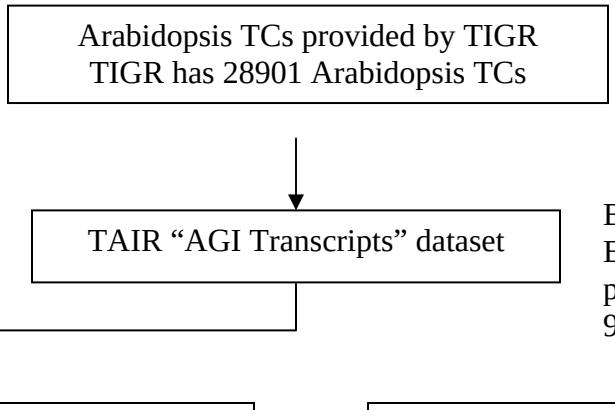

Table C: Total TC-AGI mappings 41651	
Arabidopsis TC	AGI
TC261045	AT5G38410
TC251315	AT4G21960

Table D: Arabidopsis EGO-TC Mappings by TIGR 18551	
TIGR EGO	TIGR Orthologous TCs
894156	*Arabidopsis*|TC261045
894523	*Arabidopsis*|TC251315

Union of Table B, Table C and Table D

Table E: Total AG-AGI-TC-EGO mappings 7823 Total ATH1-AGI-TC-EGO mappings 20051				
AG Probe set	ATH1 Probe set	AGI	Arabidopsis TC	EGO
12936_s_at	264474_s_at	AT5G38410	TC261045	894156
12752_s_at	254386_at	AT4G21960	TC251315	894523

Union of Table A and Table E

7744 mappings between Affymetrix Poplar Genome Array probe sets and Affymetrix AG probe sets 17297 mappings between Affymetrix Poplar Genome Array probe sets and Affymetrix ATH1 probe sets					
Poplar Affymetrix Probe set	Poplar TC	Arabidopsis TC	Arabidopsis AGI	Arabidopsis Affymetrix AG probe set	Arabidopsis Affymetrix ATH1 probe set
PtpAffx.249.8.A1_s_at	TC19207	TC261045	AT5G38410	12936_s_at	264474_s_at
Ptp.1492.1.S1_s_at	TC28054	TC251315	AT4G21960	12752_s_at	254386_at

C. Programming Details

1) CCPMT Mapping: CCPMT uses its own set of mappings to map probe sets within and across species. In case of Arabidopsis, many microarray vendors provide mappings between the probe sets and the corresponding AGI IDs, but since the goal was to use a new set of mappings, the NCBI BLAST blastn program was used for sequence alignment. The output of the blastn is in a special format, hence a Java parser was written to extract the significant data.

2) Data Comparison: Once the mappings were parsed from the BLAST output, this data was compared with the microarray vendor mappings. To carry out the comparative analysis between CCPMT and the microarray vendor mappings, SAS and Microsoft Excel were used.

3) CCPMT application: The core code for CCPMT was written in Java programming language because Java is platform independent. In the future the CCPMT tool will work with other existing software tools. If the tools are written in the same software language, tool integration is comparatively simpler. CCPMT was developed in three stages:

- Web Pages.
- Core methods.
- Database (Backend).

The front end of CCPMT was designed as a web application. The web pages were written in JSP. Once the user hits the submit button all of the data entered is sent to the servlets. Servlets act as the core methods that process the information received from the JSP pages and query the database. MySQL was used as the backend database.

CHAPTER 5

RESULTS

The preliminary data from the microarray vendors was mapped using the sequence alignment NCBI BLAST tool. Table 5 gives the summary of the microarrays used in CCPMT.

TABLE 5
MICROARRAY VENDORS IN CCPMT

Microarray Vendor	Total probes in array	Total mappings in CCPMT (one to many)	Plant Species
Affymetrix Arabidopsis Genome (8K)	8297	7998 (probe set – AGI mapping)	Arabidopsis
Affymetrix Arabidopsis Genome ATH1-121501 (25K)	22810	23666 (probe set – AGI mapping)	Arabidopsis
Agilent - Arabidopsis 2 Oligo Microarray (V2) G4136B	21500	21500 (probe set – AGI mapping)	Arabidopsis
Arabidopsis Functional Genomics Consortium (AFGC) array	19108	32861 (probe set – AGI mapping)	Arabidopsis
Complete Arabidopsis Transcriptome MicroArray (CATMA) array	24576	22621 (probe set – AGI mapping)	Arabidopsis
Operon - Arabidopsis Genome Oligo Set Version 3.0	29954	27691 (probe set – AGI mapping)	Arabidopsis
Affymetrix Poplar Genome Array	61414	57699 (probe set – TC Mapping)	Poplar

The one-to-one microarray comparison in CCPMT has been summarized in Table 6. The arrays Affymetrix Arabidopsis Genome (8K) and Affymetrix Arabidopsis Genome ATH1-121501 (25K) have been abbreviated to AG and ATH1, respectively.

TABLE 6
SUMMARY TABLE COMPARING MAPPINGS OF MICROARRAYS IN CCPMT

	AG	ATH1	AFGC	Agilent	CATMA	Operon	Affymetrix Poplar Genome Array
AG	--	7828	12170	7018	7193	8361	7744
ATH1	7827	--	30066	19188	20521	24636	17279
AFGC	12171	30066	--	29622	26070	30509	17793
Agilent	7018	19188	29622	--	18563	21371	16913
CATMA	7192	20521	26070	18561	--	23082	16378
Operon	8362	24636	30509	21371	23081	--	17505
Affymetrix Poplar Genome Array	7744	17279	17793	16912	16378	17504	--

.

A. CCPMT Web Application

CCPMT can be queried either at the probe set level or with identifiers such as the AGI ID, TIGR EGO, or TC IDs. One can also compare entire arrays by selecting the input array and the output array from the drop-down menu. As CCPMT is a web application, users can enter their queries in a textbox and, upon submission the queries, the results are displayed in a browser friendly format.

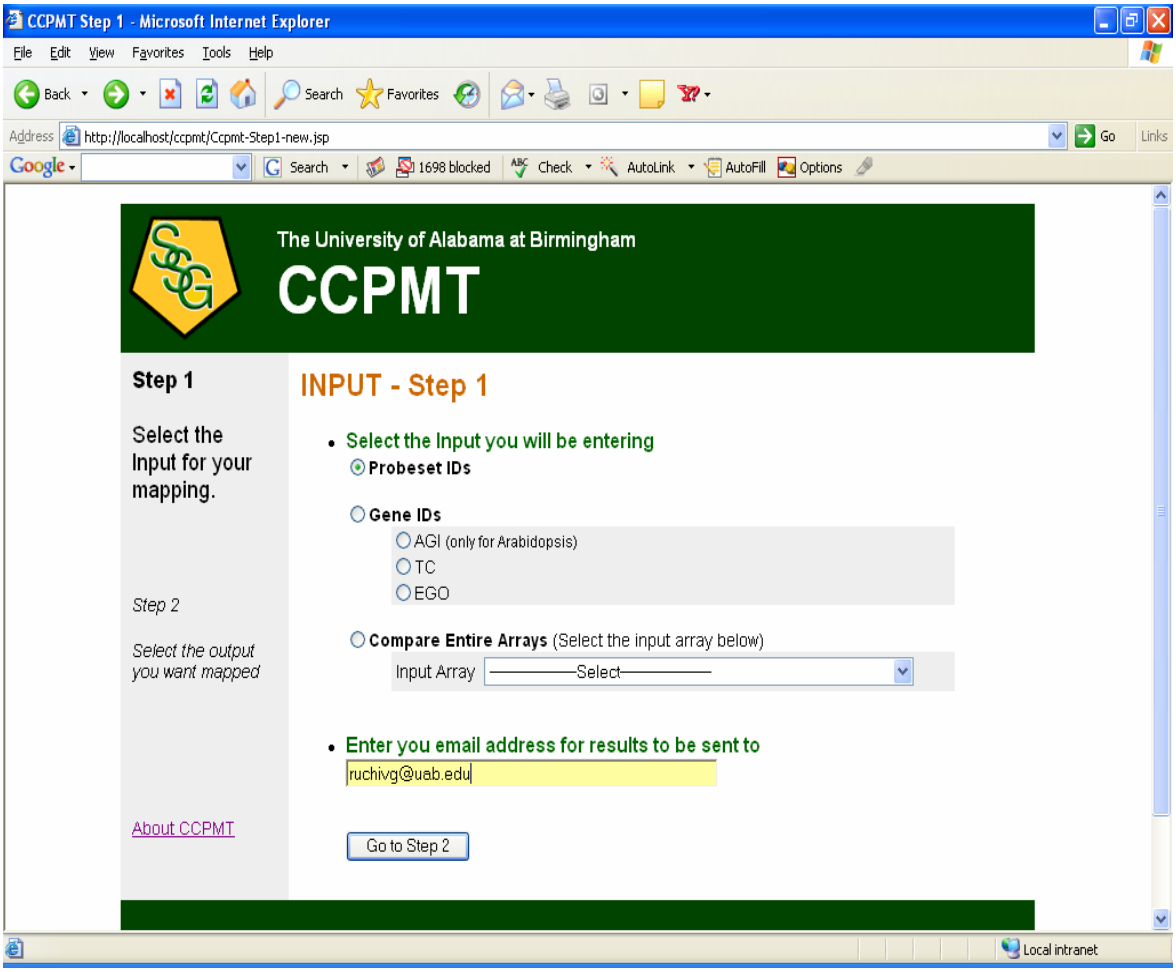

Fig.5. CCPMT screenshot for input parameters.

Fig. 5 illustrates the index page of the CCPMT web application. The user can chose to map either at the probe set level, or at the AGI, TC, or EGO levels. The example in Fig. 5 illustrates that the user wants to map the input data at the probe set ID level. This step also requires the user to enter his or her email address. The results displayed will be sent to the user as an attachment via email and will be in the comma separated file format. The user can also compare arrays by selecting the input array from the drop down menu on the index page of the CCPMT tool.

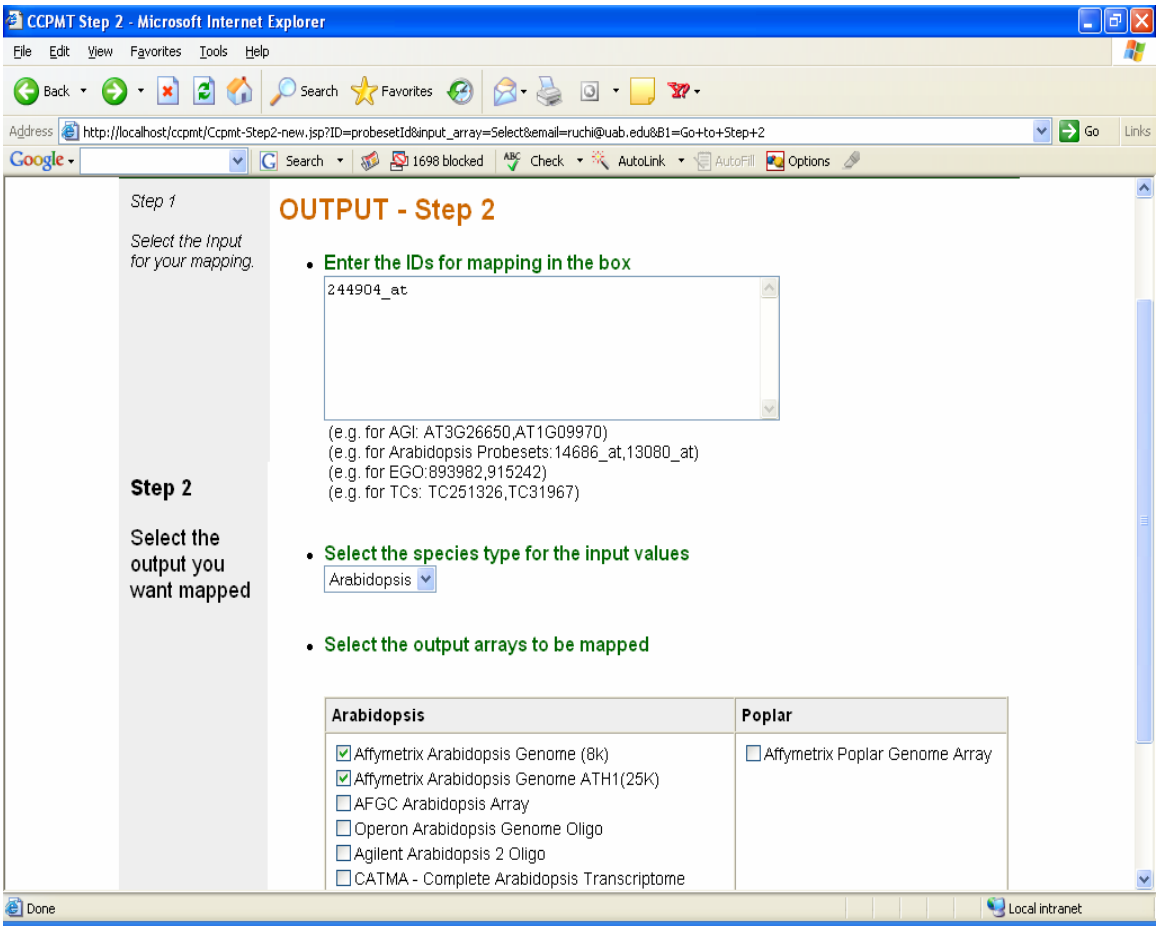

Fig.6. CCPMT screenshot for output parameters.

After entering the input, the next step is to enter the output parameters that the input should be mapped with. Fig. 6 illustrates the example where the probe set ID 244904_at needs to be mapped to other probe set IDs in the Affymetrix AG and the Affymetrix ATH1 arrays. The user also has to enter the plant species of the input probe set ID. In Fig. 6 the user has selected Arabidopsis as the plant species of the input probe 244904_at. This information is useful in deciding if the mapping should be done within the same plant species or across plant species.

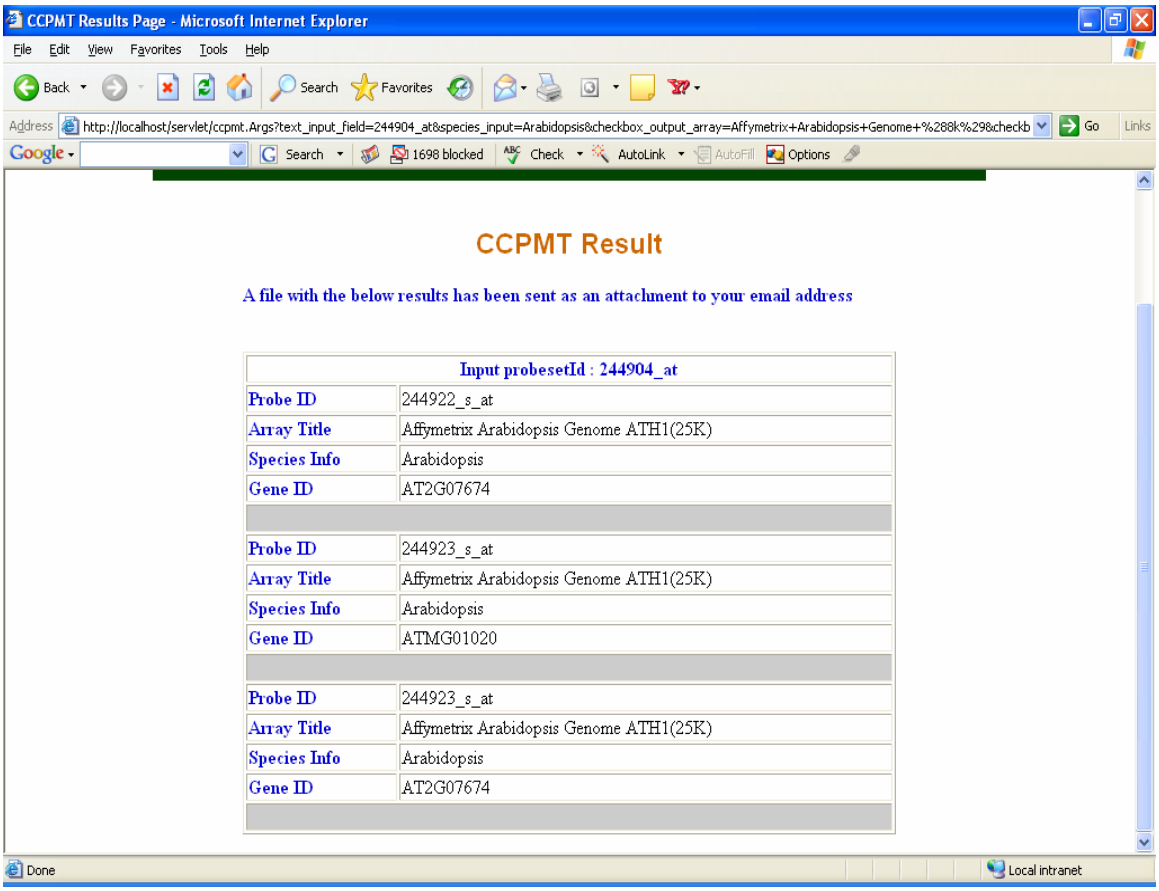

Fig.7. CCPMT result screenshot.

Once the user clicks the submit button the result is displayed. This page also contains the message that the result was sent as an attachment to the user through the

email address. Fig. 7 shows that the probe set 244904_at was mapped to 244922_s_at, 244923_s_at, 244923_s_at through the respective AGI IDs.

CHAPTER 6

CONCLUSION

The goal for CCPMT was to allow investigators to identify probe sets that correspond to the genes within and across species. CCPMT was developed as a web application to allow easy accessibility. An investigator can query the CCPMT database online and also get the results as an email attachment. The current version of CCPMT supports two plant species and their probe set level data from seven vendors. The CCPMT mappings were also compared with the vendor mappings. CCPMT can query not only at the probe set level but even at the gene level, i.e., using AGI, TC, and EGO IDs.

CCPMT currently has six Arabidopsis microarray vendors and one Poplar microarray vendor. The tool was designed in such a way that one can easily incorporate a new microarray vendor for the current plant species as well as for new plant species. Table 7 contains plant species that have been short listed for future inclusions in the CCPMT tool.

CCPMT can also be modeled to include and compare mammalian data, e.g., human and mouse microarrays data. Coexpression Tool, a database of 500+ *Arabidopsis* ATH1 microarrays from the Nottingham Arabidopsis Stock, has been developed at SSG. CCPMT will facilitate better integration of the data into the Coexpression Tool. HDBStat! is a tool designed at SSG for the statistical analysis of microarray data using

methods that take into account non-normal data and small sample sizes, and make use of mixture models. CCPMT will be integrated into HDBstat! for easier annotation of the microarrays.

TABLE 7
PLANT MICROARRAY RESOURCES FOR CCPMT

	Affymetrix	Agilent	Nimblegen	Operon	cDNA
Arabidopsis	P	P	A	P	A
Barley	A	--	--	--	--
Maize	A	--	--	A	A
Soybean	A	A	--		
Rice	A	A	--	A	A
Sugarcane	A	--	--		
Tomato	A	--	--	A	A
Wheat	A	--	--	--	--
Onion	--	--	--	--	--
Grape	--	--	A	--	--
Medicago	A	--	--	A	A
Poplar	P	--	--	--	A
Spruce	--	--	--	--	A

P – Already present in CCPMT

A – Available for future addition in CCPMT

-- – Not available with the microarray vendor

REFERENCES

[1] D. V. Nguyen, A. B. Arpat, N. Wang, and R. J. Carroll, "DNA Microarray Experiments: Biological and Technological Aspects," *Biometrics*, vol. 58, issue 4, pp. 701-17, Dec. 2002.

[2] "TAIR – Home Page" [Online] Available: http://www.arabidopsis.org. [Accessed Jan 14, 2006].

[3] B. Stirling, Z. K. Yang, L. E. Gunter, G. A. Tuskan, and H. D. Bradshaw, Jr, "Comparative sequence analysis between orthologous regions of the Arabidopsis and Populus genomes reveals substantial synteny and microcollinearity," [Online] Available: http://cjfr.nrc.ca. [Accessed Feb 15,2006].

[4] "Eukaryotic Gene Orthologs - The Institute for Genomic Research" [Online] Available: http://www.tigr.org/tdb/tgi/ego [Accessed Feb 22, 2006].

[5] "TIGR Gene Indices Information Page" [Online] Available: http://www.tigr.org/tdb/tgi/definitions.html [Accessed Feb 22, 2006].

[6] K. H. Cheung, J. Hager, D. Pan, R. Srivastava, S. Mane, Y. Li1, P. Miller and K. R. Williams, "KARMA: a web server application for comparing and annotating heterogeneous microarray platforms," *Nucleic Acids Research*, vol. 32, pp. 441-444, July. 2004.

[7] J. Tsai, R. Sultana, Y. Lee, G. Pertea, S. Karamycheva, V. Antonescu, J. Cho, B. Parvizi, F. Cheung and J. Quackenbush, "RESOURCERER: a database for annotating and linking microarray resources within and across species," *Genome Biology*, vol. 2, Oct. 2001.

[8] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D.J. Lipman, "Basic local alignment search tool," *J. Mol. Biol*, vol. 215, pp. 403-410, Feb.1990.

[9] A. J. Olson, T. Tully and R. Sachidanandam, "GeneSeer: A sage for gene names and genomic resources," *BMC Genomics*, vol. 6, pp. 134. Sept. 2005.

[10] “TAIR – FTP site” [Online] Available:
ftp://ftp.arabidopsis.org/home/tair/home/tair/Sequences/blast_datasets/README
[Accessed Jan 14, 2006].