

AffyProbeMiner: a web resource for computing or retrieving accurately redefined Affymetrix probe sets

Hongfang Liu^{1,2,\$}, Barry R Zeeberg^{1,\$}, Qu Gang³, A Gunes Koru⁴, Alessandro Ferrucci^{1,5}, Ari Kahn^{1,6}, Michael C Ryan⁶, Antej Nuhanovic^{7,8}, Peter J Munson⁷, William C Reinhold¹, David W Kane⁸, John N Weinstein^{*,1}

¹Genomics and Bioinformatics Group, Laboratory of Molecular Pharmacology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA

²Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University Medical Center, 4000 Reservoir Road, NW, Washington, DC 20007, USA

³Department of Electrical and Computer Engineering, University of Maryland, College Park, College Park, MD 20742, USA

⁴Department of Information Systems, University of Maryland, Baltimore County, 1000 Hilltop Circle, MD 21050, USA

⁵Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, 1000 Hilltop Circle, MD 21050, USA

⁶Department of Bioinformatics, George Mason University, Fairfax, Virginia, 20110, USA

⁷Mathematical and Statistical Computing Laboratory, Division of Computational Bioscience, Center for Information Technology, National Institutes of Health, Bethesda, MD 20892, USA

⁸SRA International, 4300 Fair Lakes Court, Fairfax, VA 22033, USA

^{\$}Contributed equally

ABSTRACT

Motivation: Affymetrix microarrays are widely used to measure global expression of mRNA transcripts. That technology is based on the concept of a probe set. Individual probes within a probe set were originally designated by Affymetrix to hybridize with the same unique mRNA transcript. Because of increasing accuracy in knowledge of genomic sequences, however, a substantial number of the manufacturer's original probe groupings and mappings are now known to be inaccurate and must be corrected. Otherwise, analysis and interpretation of an Affymetrix microarray experiment will be in error.

Results: AffyProbeMiner is a computationally-efficient platform-independent tool that uses all RefSeqs and validated complete coding sequences in GenBank to (1) regroup the individual probes into consistent probe sets and (2) remap the probe sets to the correct sets of mRNA transcripts. The individual probes are grouped into probe sets that are 'transcript-consistent' in that they hybridize to the same mRNA transcript (or transcripts) and, therefore, measure the same entity (or entities). About 65.6% of the probe sets on the HG-U133A chip were affected by the remapping. Pre-computed regrouped and remapped probe sets for many Affymetrix microarrays are made freely available at the AffyProbeMiner web site. Alternatively, we provide a web service that enables the user to perform the remapping for any type of short-oligo commercial or custom array that has an Affymetrix-format Chip Definition File (CDF). Important features that differentiate AffyProbeMiner from other approaches are flexibility in the handling of splice variants, computational efficiency, extensibility, customizability, and user-friendliness of the interface.

Availability: The web interface and software (GPL open source license), are publicly-accessible at <http://discover.nci.nih.gov/affyprobeminer>.

Contact: hl224@georgetown.edu or barry@discover.nci.nih.gov.

1 INTRODUCTION

Microarrays are widely used for large-scale, quantitative molecular profiling of biological specimens (Stoughton 2005). However, because of poor reproducibility across different platforms or different generations of the same platform (Carter, Eklund et al. 2005; Kong, Hwang et al. 2005; Biotechnology 2006), the validity of microarray results remains a subject of concern to the scientific community. Because hybridization to a microarray is sequence-dependent, it is affected by mis-assignment of probes, ambiguous assignment of probes, and alternative splicing of target transcripts. With respect to the latter, the percentage of genes exhibiting alternative splicing is high - variously estimated as 30% and 99% (Boue, Letunic et al. 2003; Lee and Roy 2004). Accurate quantitation requires knowledge of both the identity of the genes and the splice forms that are expressed. Ideally, probes in a probe set should all target the same set of transcripts. With improved genomic knowledge, we have become increasingly aware of the deviation from that ideal. Table 1 shows examples of the two major classes of mapping problems that necessitate redefinition of probe sets on the Affymetrix HG-U133A chip:

1. **A probe set contains some probes that match multiple transcripts** - Probes within a probe set do not all target the same set of transcripts. The expression levels measured by

those probes will introduce inconsistencies in quantitation algorithms. Table 1 illustrates the type of complication that can arise:

- Affymetrix had originally represented the human genes CLEC2D by one probe set, 220132_s_at, and NPM1 by two probe sets, 221691_x_at and 200063_s_at.
 - Currently, three RefSeqs represent CLEC2D, and three RefSeqs represent NPM1.
 - The entries in Table 1 for each probe set (column) identify the probes that match the RefSeqs (rows). For example, all 11 probes in probe set 220132_s_at match NM_013269.
 - The level of hybridization to probe set 200063_s_at provides a consistent estimate of the composite expression for RefSeqs NM_002520 and NM_199185 of NPM1. None of the probes in the set hybridize with RefSeq NM_001037738. However, expression of RefSeq NM_001037738 is reflected in the hybridization of probe set 221923_s_at.
 - In contrast, if we use probe set 221691_x_at to measure the expression of transcripts of NPM1, the level of hybridization to the probe set could reflect cross-hybridization with RefSeqs of CLEC2D.
2. **Some probes in a probe set do not match the target transcript(s)** - Several probes within a probe set may not match any of the transcripts for the gene that Affymetrix had originally designated for the probe set. The expression levels measured by those probes do not reflect the composite expression of the transcripts of the intended gene and therefore introduce an inconsistency in quantitation. For example, again in relation to Table 1:
- Probes 7 and 8 of 221691_x_at do not target NM_199185, which represents NPM1, but they do target all three transcripts of CLEC2D.
 - Therefore, the expression levels measured by 221691_x_at do not consistently reflect the composite expression of the RefSeqs of the intended gene.

After the probe sequence information was made public by Affymetrix, several publications addressed those mapping problems by redefining probe sets (Gautier, Moller et al. 2004; Carter, Eklund et al. 2005; Dai, Wang et al. 2005; Harbig, Sprinkle et al. 2005; Kong, Hwang et al. 2005). For example, Harbig (Harbig, Sprinkle et al. 2005) used BLAST to match probes with documented and postulated human transcripts, and redefined about 37% of the probes on the HG-U133 plus 2.0 array. They found that the original Affymetrix annotation was compromised because of the potential for cross-hybridization with splice variants or transcripts of other genes that contained matching sequences. More than 5000 probe sets were shown to hybridize with multiple transcripts. They proposed a sequence-based identification method in which probe sets were remapped to the most closely-related RefSeqs. An R package, *altcdfenvs*, was developed by Gautier (Gautier, Moller et al. 2004) to provide alternative mapping of probes. As proof of concept, the package was applied to those genes that are represented in RefSeq. Probes that matched multiple RefSeqs were eliminated from consideration. Dai (Dai, Wang et al. 2005) recently provided redefinitions under several conditions. They sug-

Table 1. Example of ambiguous and mismatched probes in the original assignments for the Affymetrix HG-U133A chip

| Gene | RefSeq | Probe set | | | |
|--------|--------------|-------------|-------------|-------------|-------------|
| | | 220132_s_at | 221691_x_at | 200063_s_at | 221923_s_at |
| CLEC2D | NM_013269 | 1-11 | 1,3,4,7,8 | - | - |
| | NM_001004419 | 1-11 | 1,3,4,7,8 | - | - |
| | NM_001004420 | 1-11 | 1,3,4,7,8 | - | - |
| NPM1 | NM_002520 | - | 1-9 | 1-11 | - |
| | NM_199185 | - | 1-6,9 | 1-11 | - |
| | NM_001037738 | - | 1-11 | - | 1-11 |

Table 2. The number of raw complete coding sequences extracted for each organism when considering RefSeq only, or RefSeq and GenBank. The coverage for some organisms (e.g., Rat and Rice) is low. That low coverage could result in (1) probes being discarded because of no target transcript; and (2) probe sets failing to be split because of the low resolution. We expect those problems to be less severe as coverage becomes more complete.

| Organism | Number of Records | |
|--------------------------|-------------------|------------------|
| | RefSeq | RefSeq & GenBank |
| Arabidopsis thaliana | 30,443 | 88,580 |
| Bos Taurus | 7,015 | 10,018 |
| Caenorhabditis elegans | 23,122 | 24,315 |
| Canis familiaris | 781 | 995 |
| Danio rerio | 10,957 | 15,114 |
| Drosophila melanogaster | 19,854 | 22,468 |
| Gallus gallus | 4,069 | 4,805 |
| Glycine max | 0 | 677 |
| Homo sapiens | 24,413 | 73,885 |
| Hordeum vulgare | 0 | 144 |
| Macaca mulatta | 371 | 656 |
| Medicago truncatula | 0 | 177 |
| Mus musculus | 20,286 | 47,883 |
| Oryza sativa | 26,768 | 28,778 |
| Plasmodium falciparum | 0 | 185 |
| Pseudomonas aeruginosa | 0 | 1 |
| Rattus norvegicus | 10,171 | 15,522 |
| Saccharomyces cerevisiae | 0 | 77 |
| Saccharum officinarum | 0 | 176 |
| Staphylococcus aureus | 0 | 2 |
| Sus scrofa | 1,182 | 1,953 |
| Triticum aestivum | 0 | 879 |
| Vitis Vinifera | 0 | 152 |
| Xenopus laevis | 0 | 12,004 |
| Zea mays | 0 | 953 |

gested that the original Affymetrix probe set definitions are imperfect, that conclusions derived from past Affymetrix GeneChips may be somewhat inaccurate, and that researchers should re-analyze their data.

Probe sequence information was also used to address the problem of poor reproducibility across different platforms or different generations of the same platform. For example, Carter (Carter, Eklund et al. 2005) redefined Affymetrix probe sets by sequence overlap with cDNA microarray probes to reduce cross-platform inconsistencies in cancer-associated gene expression measurements. Their study implied that “probes targeting identical transcript sequence regions give substantially stronger concordance than probes that target identical contiguous transcript molecules at different sequence regions.” It also suggested that discrepancies among different platforms are caused by improper cross-platform probe matching. Kong (Kong, Hwang et al. 2005) used sequence information to increase the compatibility among different generations of Affymetrix arrays. They filtered probes that were not consistent with the Affymetrix annotation.

Despite those positive developments, no computationally-efficient, broadly-capable tools for remapping have been made available to date. The only publicly-available tool, the BioConductor package *altcdfenvs*, is written in R. Unfortunately, R suffers from poor memory management (Hornik 2007): all objects are kept in memory until garbage collection is automatically invoked. For example, the *matchprobes* package (used in *altcdfenvs*) took several days to align probes with mRNA sequences (Elo, Lahti et al. 2005). We tried to use the *altcdfenvs* package to obtain remapped Chip Definition Files (CDFs) for HG-U133A. The package could be executed successfully on a Unix server with a test set of 500 human RefSeqs. However, when we tested the complete set of human RefSeqs, no results were obtained even after 10 days of processing.

To address such problems, we have developed AffyProbeMiner, a computationally-efficient and platform-independent tool. It uses mRNA RefSeqs and validated complete coding sequences in GenBank to (1) regroup the individual probes into consistent probe sets and (2) remap the probe sets to the correct sets of mRNA transcripts. AffyProbeMiner uses a local implementation of the UCSC BLAT server (Kent 2002) to circumvent the mapping bottleneck of *altcdfenvs*.

2 METHODS

We define a *transcript-consistent* probe set as one in which each probe maps to the same set of transcripts and a *transcript-unique* probe set as a transcript-consistent probe set in which each probe maps to a single transcript. A transcript-consistent probe set has the advantage of eliminating the inaccuracy that would have resulted from mixing probes that match disjoint sets of transcripts. A transcript-unique probe set has the additional advantage of eliminating the ambiguity inherent in lumping together the measurements of multiple transcripts. Ideally, we would eliminate from consideration probe sets that were not transcript-unique. Unfortunately, that degree of stringency would give up too much of the potential information. We are not willing, however, to accept probe sets that are not transcript-consistent because they produce estimates that are not merely ambiguous, but are also incorrect.

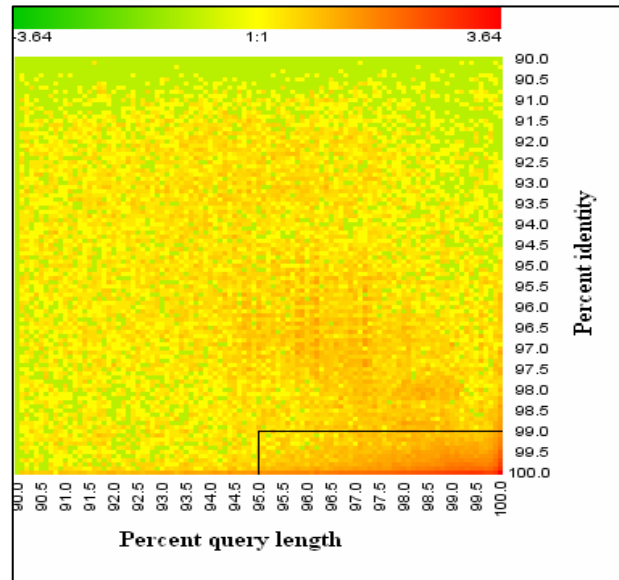


Figure 1. Two-dimensional distribution as a function of percent query length and percent identity. All human complete coding sequences in GenBank were aligned to the human genome. The output data of BLAT were parsed to determine the percent of the query length that aligned to the genome and the percent identity of the aligned portion of the query sequence. The rectangular box in the lower right corner encloses the region of acceptance. 94.6% of the human complete coding sequences met the acceptance criteria. The distribution values are given as log 10. Original distribution values were incremented by 0.10 to avoid attempting to compute log (0).

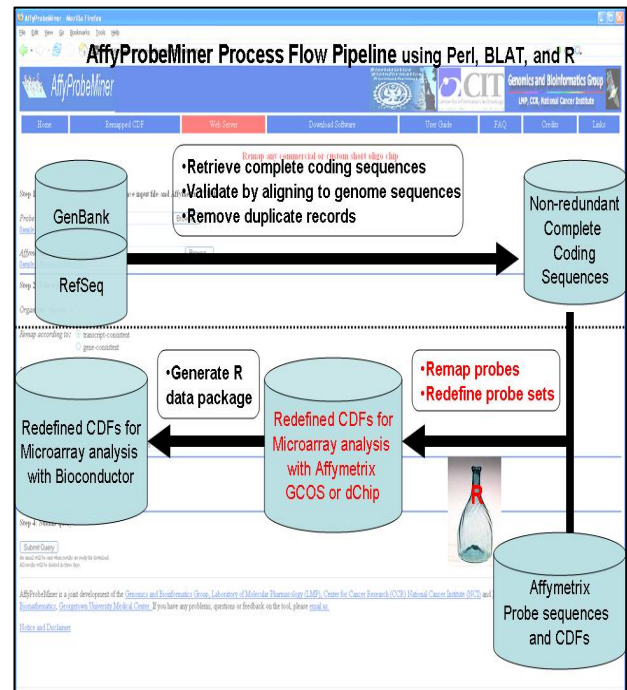


Figure 2. AffyProbeMiner process flow pipeline. Red font indicates novel features of AffyProbeMiner, which uses Perl and BLAT to circumvent the computational bottleneck (indicated by **R** in the bottle) of the R program package *altcdfenvs*. The R computation generates CDFs for only the Bioconductor environment. In contrast, AffyProbeMiner generates CDFs for (1) the Bioconductor environment, (2) Affymetrix GCOS, and (3) dChip. The background of Figure 2 is a screenshot of the web interface.

We also define the weaker concepts of gene-consistent and gene-unique probe sets. A gene-consistent probe set is one in which each probe maps to transcripts of the same set of genes, and a gene-unique probe set is one in which each probe maps to transcripts of a single gene. That is, the individual probes in gene-consistent and gene-unique probe sets may target different splice variants of the gene(s) in question. We reluctantly define and support resources in AffyProbeMiner for those users who choose to ignore the implications of splice variation. However, we suggest that, when possible, microarray data analysis be performed at the transcript level to avoid inconsistency in the results and their interpretation.

After AffyProbeMiner processing, the data are returned to the normal Bioconductor or Affymetrix stream for completion of the analysis using MAS5, RMA, or some other algorithm (see (Breitling 2006)).

The following section provides a brief description of the methods.

Construction of a database containing validated complete coding sequences from GenBank and RefSeq - We obtained a list of the raw complete coding sequences in GenBank. Those sequences could be detected because their GenBank records contained “complete CDS” or “complete sequences.” We excluded sequences whose GenBank records contained “intronic transcript” or “BAC clone”. All RefSeq records having accessions starting with “NM_” were included in the list.

Because of the variable reliability of records in GenBank, we determined their validity by alignment to the current genome builds. We considered a record to be valid if $\geq 95\%$ of the sequence can be aligned with the genome with $\geq 99\%$ identity. Figure 1 shows a typical two-dimensional distribution representing those two criteria applied to human GenBank sequences. Based on those two criteria, 94.6% (46,546 out of 49,189) complete coding sequences in GenBank were considered to be valid.

Sometimes the ends of a valid sequence contain vector contamination or a poly(A) tail. We detected those during BLAT alignment to the genome sequences and removed the unmatched ends. The cleaned, validated records were then de-duplicated. Records were considered to be replicates if (1) NCBI annotated them as mapping to the same gene, (2) the coding regions have the same length, and (3) one sequence can be subsumed by the other sequence with differences in nucleotides identity $\leq 1\%$. If GenBank and RefSeq records were found to be duplicates, we retained the RefSeq record. In the case of two duplicate GenBank records, we kept the longer one. The result was a validated non-redundant complete coding sequence database.

Generation of redefined chip definition files (CDFs) containing re-mapped and re-grouped probes - We downloaded all of the probe sequence files from the Affymetrix website. The mappings between probe sequences and complete coding sequences were determined by BLAT. The redefinition of CDFs was based on the mapping results. AffyProbeMiner circumvents the bottleneck in the remapping process that occurs when the computations are performed in the R language. The processing flow diagram (Figure 2) indicates how it does so (red font), to provide remapped CDF files that are suitable for various microarray data analyses.

AffyProbeMiner’s suite of Perl programs for generating CDFs is downloadable at the project website, <http://discover.nci.nih.gov/AffyProbeMiner>. The public availability of the programs allows a user to obtain CDFs with her/his own parameter settings. Advanced users can customize the parameter settings or program code.

The FAQ and User Guide at the AffyProbeMiner website are applicable to both transcript-level and gene-level analyses. The User Guide provides detailed instructions for using our remapped CDF files in the Affymetrix GCOS, dCHIP, or BioConductor environments.

In addition to examining all probe sets (“condition A”), we also constrained the study to those probe sets having at least five probes (“condition F”). We believe that, as a heuristic rule of thumb, a minimum of five probes should be required in a probe set to produce reliable measurements. The statistical characterization is similar whether no threshold or the five-probe threshold is used. For the human HG-U133A chip, the probes that match a subsequence of a complete coding sequence comprise a significant fraction ($206,624/247,966 \times 100\% = 83.3\%$ and $187,232/247,966 \times 100\% = 75.5\%$ under conditions A and F, respectively) of the probes on the chip. Additionally, the majority of the probes ($197,249/247,966 \times 100\% = 79.5\%$ and $196,695/247,966 \times 100\% = 79.3\%$) are gene-unique, whereas only a small fraction ($69,344/247,966 \times 100\% = 27.9\%$ and $66,213/247,966 \times 100\% = 26.7\%$) are transcript-unique.

3 RESULTS AND DISCUSSION

Affymetrix microarrays are widely used to measure global expression of mRNA transcripts. Overall analysis of an Affymetrix microarray experiment requires a combination of specific probe expression results from the experiment and a pre-defined CDF that groups probes into probe sets. Popular software systems for analysis of the expression data within the context of a given CDF are Bioconductor R (Gentleman, et al., 2004), Affymetrix GCOS, and dChip. All of the individual probes within a given probe set were originally selected by Affymetrix to hybridize with the same unique mRNA transcript. However, because of increasing accuracy in our knowledge of genomic sequences (particularly for the human genome), a substantial number of the manufacturer’s original probe groupings and mappings need to be corrected. Analysis and interpretation of an Affymetrix microarray experiment will be in error both qualitatively and quantitatively in the absence of corrected regrouping and remapping.

For the remapping process (Figure 2), we have developed AffyProbeMiner, a computationally-efficient platform-independent tool that uses all mRNA RefSeqs and complete coding sequences in GenBank to (1) regroup the individual probes into consistent probe sets and (2) remap the probe sets to the correct sets of mRNA transcripts.

There are, in total, 48 Affymetrix chips for 25 organisms. Table 2 shows the number of raw complete coding sequences extracted for each organism when considering RefSeq only and when considering both RefSeq and GenBank. We constructed a complete coding sequence database for every organism represented by an Affymetrix chip as of Jan, 1, 2006 and having $\geq 5,000$ raw complete coding sequences (RefSeq or GenBank). We generated redefined CDFs for each chip associated with those organisms. We considered only perfect matches in the pre-computed results, and remapped only probe sets for which at least five probes were retained. All of those processing steps are fully automated using Perl and R scripts, and it takes approximately 20 hours to build, from scratch, the complete release of AffyProbeMiner’s remapped and redefined CDFs for 64 microarray sets representing 9 species. The processing step that relieves the bottleneck in R takes approxi-

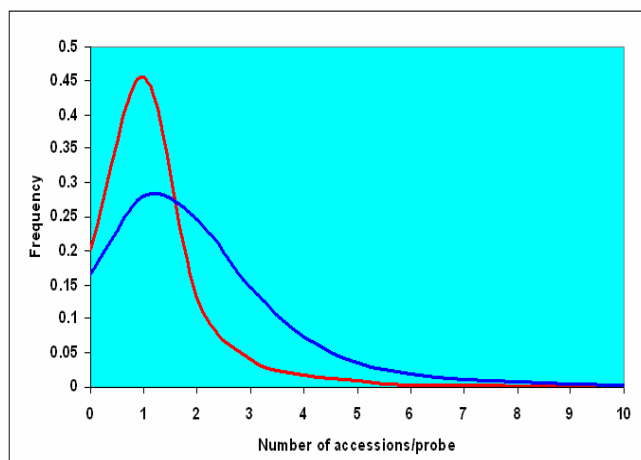


Figure 3. The distribution of HG-U133A probes that are mapped to multiple records. Two mapping strategies are considered: (1) RefSeqs only, and (2) RefSeqs and GenBank complete coding sequences. The relative number of transcript-unique probe sets is given by the ratio of frequencies evaluated at abscissa = 1. Neglecting the GenBank records will lead to an overly optimistic estimate of the number of transcript-unique probe sets.

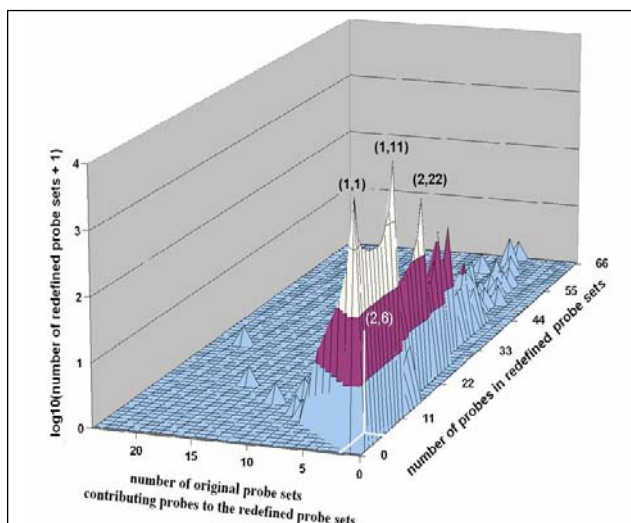


Figure 4. Three-dimensional histogram of the number of probes in redefined probe sets and number of original probe sets contributing probes to the redefined probe sets. A few examples will help in interpreting this figure:

- i) The large peak at (1,11) indicates the number of instances in which the redefined probe set was identical to an Affymetrix-defined probe set. That is, the Affymetrix-defined probe set, containing 11 probes, gave rise to a redefined probe set containing 11 probes (i.e., all 11 probes in the redefined set arose from one Affymetrix-defined set.)
- ii) The large peak at (1,1) indicates the number of instances in which the redefined probe set was a singleton, and of necessity arose from a probe from one Affymetrix-defined probe set.
- iii) Another special case is the point (2,22). All 22 probes in two Affymetrix-defined probe sets were merged into a single redefined probe set.
- iv) A somewhat more general case is the point (2, 6). The six probes in the redefined set arose from 2 Affymetrix-defined probe sets: p probes from one Affymetrix probe set and $(6 - p)$ probes from a second Affymetrix probe set.

mately three minutes/chip (PowerBook G4 with 1G memory). The pre-computed, remapped probe sets are freely available to the scientific community via download from our web site. Alterna-

Table 3. System comparison.

| Attributes | Authors/Tools | | |
|--|---------------|-----|----------------|
| | Gautier | Dai | AffyProbeMiner |
| Remapped CDFs down-loadable | No | Yes | Yes |
| All Affymetrix chips included | No | Yes | Yes |
| Annotation Package | No | No | Yes |
| Web server for custom computation or custom chips | No | No | Yes |
| Software down-loadable | Yes | No | Yes |
| Computationally efficient (i.e., does not require multimode cluster) | No | NA | Yes |
| Uses only complete coding sequences (i.e., no ESTs) | No | No | Yes |
| Option to keep 'consistent' sets or 'unique' sets* | No | No | Yes |
| Option to set threshold for number of probes per probe set | No | No | Yes |
| User-friendly integrated web interface | No | Yes | Yes |

NA, not applicable.

* We define a *transcript-consistent* probe set as one in which each probe maps to the same set of transcripts, and a *transcript-unique* probe set as a transcript-consistent probe set in which each probe maps to a single transcript. This table is intended for comparing Gautier with AffyProbeMiner and Dai with AffyProbeMiner, not for comparing Dai with Gautier.

tively, the computation can be performed as a web service on our server either for Affymetrix arrays or for any other short-oligo that use Affymetrix-format CDFs. The web service allows users to generate their own customized probe sets with options for the number of mismatches allowed during sequence alignment and minimum number of probes per probe set. AffyProbeMiner is computationally-efficient and also user-friendly.

Figure 3 shows the distribution of probes on one human chip, HG-U133A, that map to multiple records. The results shown in Figure 3 indicate that neglecting the GenBank complete coding sequences will lead to an overly optimistic estimate of the number of transcript-unique probe sets.

There are a total of 25,582 redefined probe sets for the human U133A chip. Of those, 6,878 map to single complete coding sequences, 16,426 map to multiple complete coding sequences of a single gene, and the rest map to multiple genes. Figure 4 summarizes the results of the remapping process. The peak at (1, 11) represents the relatively small percentage (27.6%) of original probe sets as designated by Affymetrix that were not affected by the remapping process; the majority (72.4%) were, in fact, affected. Strikingly, 70.5% of the redefined probe sets that map to a single gene are not transcript-unique (i.e., the probes in them map to multiple splice variants). The most common reason, in fact, that redefined probe sets end up non-unique is mapping of probes to multiple splice forms, rather than to different genes.

AffyProbeMiner offers a number of significant advantages over other packages (Table 3). Included are coverage of all Affymetrix chips except exon chips (see below), an annotation package, a web server for custom computation or remapping of custom chips, remappings based on high quality complete coding sequences (i.e., excluding ESTs but including both GenBank and RefSeq entries), an option for the user to select either “consistent” (less stringent) or “unique” (more stringent) sets, and an option for the user to select the minimum number of probes required in the remapped probe sets. Those features, along with the computational tractability and user-friendly, integrated web interface, make AffyProbeMiner a comprehensive solution available for the remapping problem.

4 FUTURE DIRECTION AND CONCLUSION

A relatively new technology, the exon chip, is available from Affymetrix, currently for human, mouse, and rat. Conceptually, it is similar to the traditional Affymetrix chips in that there are “probe sets.” However, the probe sets are directed toward coverage of exons, rather than the 3' ends of genes. AffyProbeMiner does not currently support the new exon chips, but we are in the process of extending its processing logic to do so. We expect to integrate exon chip features into the website in the near future.

In separate studies that are beyond the scope of this paper, we are using AffyProbeMiner to study several biologically and biomedically important research problems, including enhanced characterization of the NCI-60 cancer cell line (Weinstein, 2006).

In conclusion, the AffyProbeMiner web site (<http://discover.nci.nih.gov/AffyProbeMiner>)

- Enables the user to download pre-computed, redefined CDFs that are compatible in format with most microarray data analysis platforms, including Bioconductor (Gentleman, et al., 2004), Affymetrix GCOS, and dCHIP.
- Provides programs with which the user can construct redefined CDFs for any short oligo custom array having Affymetrix-format CDFs.
- Deploys a server that can upload probe sequence files in FASTA format for generating redefined CDFs.
- Includes a number of important algorithmic, computational, and usability features that differentiate AffyProbeMiner from other available approaches to the remapping problem.

ACKNOWLEDGEMENTS

This research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research, by NSF grant IIS-0430743 to HL, and by the Intramural Research Program of the NIH, Center for Information Technology and the NIH Clinical Center.

REFERENCES

Biotechnology, N. (2006). "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements." *Nature Biotechnology* **24**: 1151-1161.

Boue, S., I. Letunic, et al. (2003). "Alternative splicing and evolution." *Bioessays* **25**(11): 1031-4.

Breitling, R. (2006). "Biological microarray interpretation: The rules of engagement." *Biochim Biophys Acta*.

Carter, S. L., A. C. Eklund, et al. (2005). "Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements." *BMC Bioinformatics* **6**(1): 107.

Dai, M., P. Wang, et al. (2005). "Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data." *Nucleic Acids Res* **33**(20): e175.

Elo, L. L., L. Lahti, et al. (2005). "Integrating probe-level expression changes across generations of Affymetrix arrays." *Nucleic Acids Research* **33**(22): e193.

Gautier, L., M. Moller, et al. (2004). "Alternative mapping of probes to genes for Affymetrix chips." *BMC Bioinformatics* **5**: 111.

Harbig, J., R. Sprinkle, et al. (2005). "A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array." *Nucleic Acids Res* **33**(3): e31.

Hornik, K. (2007). *Frequently Asked Questions on R*.

Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." *Genome Res* **12**(4): 656-64.

Kong, S. W., K. B. Hwang, et al. (2005). "CrossChip: a system supporting comparative analysis of different generations of Affymetrix arrays." *Bioinformatics* **21**(9): 2116-7.

Lee, C. and M. Roy (2004). "Analysis of alternative splicing with microarrays: successes and challenges." *Genome Biol* **5**(7): 231.

Stoughton, R. B. (2005). "Applications of DNA microarrays in biology." *Annu Rev Biochem* **74**: 53-82.