

## A global assessment of microarray probe set quality using genome and transcriptome data verification

Zayed Albertyn, Ali Reza Zamli, Poh Yang Ming, Ching Soo Meng and Colin Hercus

*SXProbedb™ is a rapid sequence search tool designed to efficiently map short sequences with mismatches at their terminal ends to SynaBASE. SXProbedb mapped 604,260 probes from the Affymetrix HG-U133 Plus 2.0™ array in only 38 seconds against the Human genome and 17 minutes against Unigene. Performance results show a remarkable improvement of 105- and 2272.5 fold respectively when compared to miBLAST and NCBI BLASTN. When partial matches with up to 5 mismatches were allowed in a probe sequence, SXProbedb was able to accurately map 99.5- and 96.8 % of probe sets to the Human genome and Unigene respectively. Analysis of the mapping results against the Human genome demonstrated that matching microarray probes map on average to 9 unique locations in the Human genome. Probe-to-Unigene mappings and mRNA spliced alignments were used to confirm cases of HG-U133 Plus 2.0 probe sequences mapping across exon boundaries. The low specificity of these probes was confirmed with Unigene results showing that a significant portion of the probe sets mapped to multiple Unigene loci. Therefore sequence analysis of probe specificity using SynaBASE could motivate improvement of the probe design process.*

### Introduction

Microarray development has been critical in the study of holistic disease pathways. However, as genome sequence data and gene prediction improve, probes developed for a given microarray experiment should be continuously re-evaluated for their specificity for given genes. Hence the need for ultra-fast oligonucleotide probe mapping becomes ever more essential as annotations of genomes continue to evolve. This is required as it affects the interpretation of experimental data from a particular given probe set. Current applications that are widely available for mapping sequences include BLAST [1], BLAT [2], WU-BLAST, Gapped BLAST and PSI-BLAST [3] and most recently miBLAST [4].

Sequence database searches suffer very significant performance degradation when a very large number of query sequences need to be compared to a database [4].

New applications exhibiting improvements over traditional methods, such as BLAST++ and miBLAST are currently being used to reduce the time taken for this type of analysis.

SXProbedb™ is a new tool developed by Synamatix to map a set of query probe sequences to any genome built using the SynaBASE™ database technology. SXProbedb uses the SynaBASE engine to execute ultra-fast matching of oligonucleotides in a fraction of the time taken by the tools mentioned above. SXProbedb is an example of one of the many types of applications that interrogate SynaBASE. SynaBASE is a proprietary structured network database platform, which is designed to manage sequence data by efficiently storing unique subsequences or patterns for rapid and sensitive sequence analysis [5].

In this study the utilisation of SXProbedb for ultra-fast probe mapping using the SynaBASE SynaAPI™ is outlined. SXProbedb was used to map 604,258 probes from Affymetrix’s HG-U133 Plus 2.0™ set against a SynaBASE of the human genome NCBI35 and Unigene build 187 in 38 seconds and 17.5 minutes respectively [6-7].

## Materials and Methods

SXProbedb allows mismatches at the terminal ends of a query sequence, where users can choose a maximum number of mismatches that are allowed. The final results report sequence mapping positions as well as a distribution of matches for each allowed number of mismatches.

SXProbedb was used to map 604,258 unique probes from Affymetrix’s HG-U133 plus 2 array sets onto a SynaBASE of Unigene build no. 187, containing 3,332,152,639bp of EST/mRNA DNA in 5,160,228 sequences [6]. The Affymetrix dataset downloaded from [www.affymetrix.com](http://www.affymetrix.com) contained 604,260 unique probe sequences in 54,675 probe sets [8]. SXProbedb searches the forward and reverse strands of a probe sequence against a target SynaBASE. Two resultant datasets were generated for 5- and 1-mismatch(es) allowed in each case.

A smaller set of Affymetrix probes from human chromosome 22 were also mapped to the NCBI53 version of the human genome to confirm mapping accuracy. The human genome build of SynaBASE contains 3,076,781,888 base pairs of DNA in 24 chromosome sequences. The positions of the probe sequence mappings were compared to publicly available mappings from Ensembl ([www.ensembl.org](http://www.ensembl.org)) as well as results from the MegaBLAST program [9 - 10]. Mappings in Ensembl were produced using the Exonerate alignment tool [11]. A word size of 20 was used to map 25mer probe sequences to the genome using MegaBLAST because the default of 28 was not appropriate for the query length.

SynaMap™ is a SynaBASE application used to rapidly find spliced alignments of mRNA sequences to genome databases. By using the mRNA exon positions based on SynaMap transcript-to-genome mappings in tandem with probe-to-mRNA locations from Unigene, it was possible to find instances of probe sequences spanning exon boundaries.

Mismatch parameters	SynaBASE	No. Probe Sequences Mapped (% Total)	No. of Probe Sets Mapped (% Total)	Time taken
5	Human Genome	577,118 (95.5)	54403 (99.5)	38s
5	Unigene # 187	532,369 (88.1)	52940 (96.8)	17min 30 s
1	Unigene # 187	519,700 (86.0)	52395 (95.8)	21min 43s

Table 1: Summary of SXProbedb HG-U133 Array data mapped to SynaBASE of Unigene 187 and the NCBI35 build of the Human Genome.

## Sensitivity Comparison

The time required to map the Affymetrix dataset to the 24 chromosomes in the human genome using SXProbedb was 38 seconds (see Table 1). 27,140 sequences in the input dataset did not match the human genome at a threshold of 5 mismatches. The total numbers of matches to the human genome for the remaining 577,118 probe sequences at various probe mismatch lengths are summarized in Figure 1. Figure 1 shows that 5,160,995 exact matches are recorded for 577,118 Affymetrix probes that matched i.e. 0 mismatches to the human genome. Based on these results an Affymetrix probe from the HG-U133 plus 2.0 array maps on average to 9 different locations in the human genome. Probe matching to mRNA databases has been used in previous studies to globally assess the specificity of microarray probes [12]. Currently it is practical to scan a SynaBASE of one or many genomes databases in seconds for rapid assessment of microarray probe set quality.

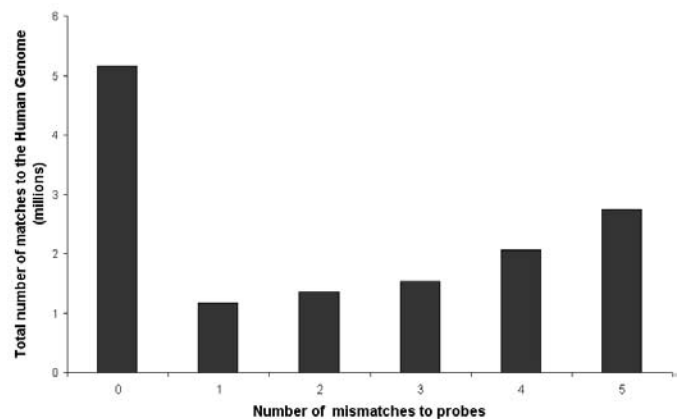


Figure 1: SXProbedb statistics for mapping the Affymetrix HG-U133™ Array probe sequences to the NCBI35 build of the human genome [7].



Figure 2: A comparison of SXProbedb results to Ensembl probe set mappings by Exonerate and Megablast [9-11]. SXProbedb-mapped probe sets produce consistent results with publicly available annotation data on Affymetrix HG-U133 probe data. Note SXProbedb identifying the correct positions as compared to the “Affy HG-U133\_+2” track and Megablast missing 1 out of 4 alignments.

A comparison of SXProbedb results to published Ensembl mappings and MegaBLAST alignments confirm the accuracy of these results [9-10]. The example in Figure 2 shows a mapping result where SXProbedb finds the same results as Exonerate’s “Affy HG-U133\_+2” track in Ensembl. Note that MegaBLAST using a word size of 20 misses 1 of the probe set alignments in this Figure 2.

### Performance and Mapping Statistics Using *Unigene Build 187*

Recent work on probe mapping to sequence databases has surfaced due to an interest in global assessment of probe set quality. Mapping of all the Affymetrix probe sequences to Unigene build 187 with SynaBASE and SXProbedb took 17m30s utilising a single CPU running linux. miBLAST required 1.26 days to perform the same task on a 2.2 GHz AMD Opteron processor and 4 GB RAM running the Linux 2.6.9 kernel. NCBI BLASTN took 27.27days to complete the same analysis (see Figure 3). These results show a 105 and 2272.5 fold improvement in performance by SXProbedb when compared to miBLAST and NCBI BLASTN, respectively [3, 4].

The performance differential in Table 1 between using a SynaBASE of the human genome and the Unigene database is due to the difference in source sequence numbers in each database. There are approximately 215,009 more sequences in Unigene than in the Human Genome SynaBASE and more processing time is needed for retrieving sequence information.

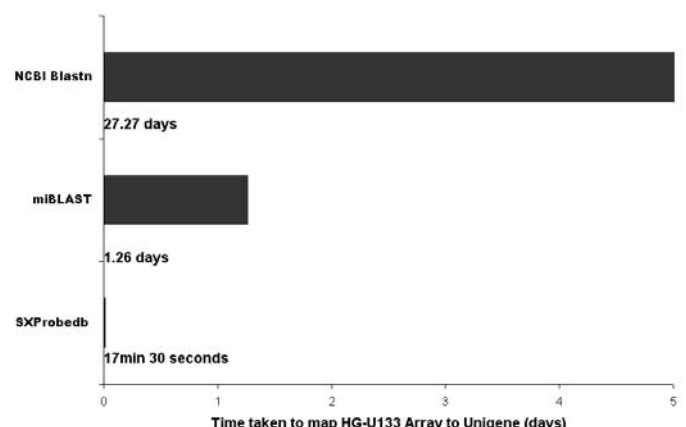


Figure 3: Relative performance of SXProbedb to published information on performance of NCBI BLASTN and miBLASTN used to map the same Affymetrix probe sequences to UniGene Build no. 187 [miBLAST ref, BLAST ref, Unigene Ref].

Sequence based approaches have the potential to improve microarray sequence quality by cross referencing probe sequences against external databases. Approximately 14% (71,889 out of the 604,260 probe sequences) of the probes sequences on the HG-U133 array did not match the Unigene data. These unmapped sequences represent either poorly designed probes or novel gene data missed by the probe design process. A comparison of the percent hit frequency for 1- and 5 mismatches is shown in Figure 4. Affymetrix probe sequences are designed from Unigene data and these results show evidence for low specificity of probe sequences in the microarray chip set.

Figure 5 provides a summary of the HG-U133 Plus 2.0 - Unigene 187 mapping in that the number of Unigene clusters matching a given probe have been enumerated. Probe sets are specifically designed to target particular genes in a microarray experiment. Unigene clusters are meant to provide a comprehensive representation of expressed genes within a genome. Therefore an individual probe set that is specific and measures gene expression should target at most, a single Unigene cluster for that gene. Cases where more than one Unigene cluster match a probe set, indicate that the result could be instances of gene or non-coding sequence duplications occurring within the genome. The distribution in Figure 5 confirms

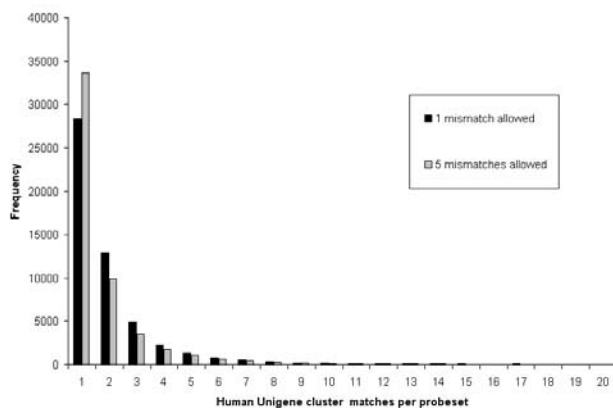


Figure 4: Statistics of SXProbedb HG-U133 Plus 2.0 probe set sequence mappings to Unigene. Results are shown for allowing up to one or five mismatches in a probe query sequence. Increasing the number of mismatches allowed increases the percent of unmapped sequences by almost 2 percent.

the one probe set - one gene theory with some minor exceptions in multiple mapping of Unigene clusters to Affymetrix probes.

Initial mapping of the probe set sequences to the genome revealed that 4.5% of Affymetrix probes did not map to any genomic location (see Table 1). A possible cause for these observations was that probe sequences were spanning exon boundaries due to spliced EST/mRNA clusters from Unigene being used as the source dataset to design these probe sequences. The probe set to Unigene mappings could therefore be used to compare the positions of the probe on the mRNA to the mRNA exon positions on the genome found by the SynaMap application. Figure 6 shows evidence of these probe sequences spanning exon boundaries. A search of the 27,140 unmapped probe sequence against Unigene yielded one or more hits per query to 20,283 (approximately 75%) probe sequences from the unmapped dataset that potentially spans exon boundaries.

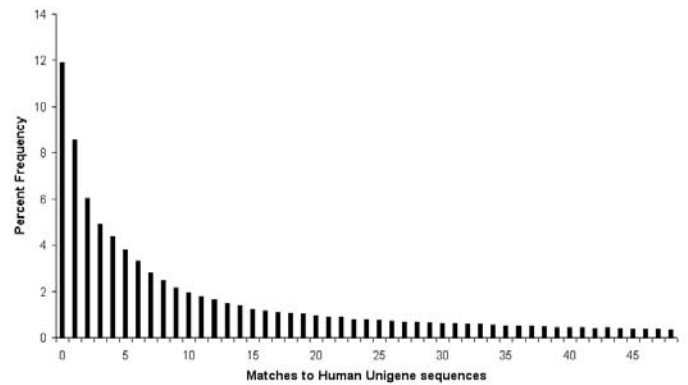


Figure 5: Specificity of Affymetrix probes matching unigene clusters. 5 Mismatches to probes were allowed to enumerate the number of Unigene clusters matching a probe set.

#### SXProbedb Result

Probe	mRNA	mRNA Start	mRNA Stop	Strand
1552322	at.615 NM_138819	834	858	+

#### SynaMap Result

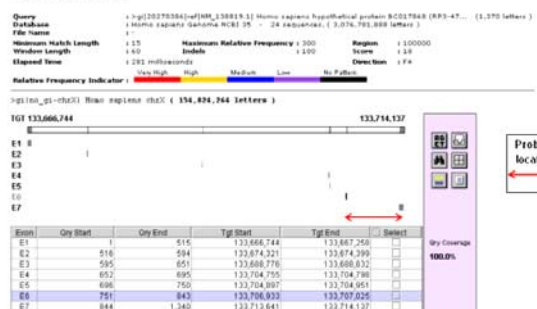


Figure 6: Example of an Affymetrix HG\_U133 Plus 2.0 array probe sequence mapping across an exon boundary. The SXProbedb probe-to-Unigene mappings were compared to SynaMap mRNA exon locations in the human genome to detect instances of probes spanning exon boundaries.

## Conclusion

The SXProbedb application, built upon the SynaBASE platform, has been used to rapidly map Affymetrix microarray probe sets to the human genome and Unigene in 38 seconds and 17.5 minutes respectively. The relative distribution of the HG-U133 Plus 2.0 probe set has been analysed according to gene specificity within the human genome and expression data from Unigene. SXProbedb results provide a motivation for improving the design process used to generate Affymetrix whole genome probes with the aid of ultra-rapid, high-throughput sequence comparison.

In addition to these results, SynaBASE and the SXProbedb mapping application can provide up to a 2272.5 fold increases in software speed performance as compared to published data by NCBI BLASTn. Another important aspect is the gains in scalability where SynaBASE performance advantages are noteworthy in comparing a large number of queries to numerous database entries e.g. 604,260 probe sequences to a database of 5,160,228 mRNA reads.

## References

1. Altschul S. F., Gish W., Miller W., Myers E. W., and Lipman D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215: 401-410.
2. Kent W. J. (2002) BLAT-the BLAST-like alignment tool. *Genome Res.* 12: 656-664.
3. Altschul S. F., Madden T. L., Schaffer A. A., Zhang J., Zhang Z., Miller W. and Lipman D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research.* 25: 3389-3402
4. Kim Y. J., Boyd A., Athey B. D., and Patel J. M. (2005) miBLAST: scalable evaluation of a batch of nucleotide sequence queries with BLAST. *Nucleic Acids Research.* 33(13): 4355-4344.
5. Albertyn Z. I., Ka Ju T., Chee San W., Ramachandran M., Iskandar J. (2004) Reconstructing Alignments Based on Patterns Inherent in Biological Data: A Qualitative Comparison. Available from [www.synamatix.com](http://www.synamatix.com).
6. Pontius JU, Wagner L, Schuler GD (2003). UniGene: a unified view of the transcriptome. In: *The NCBI Handbook*. Bethesda (MD): National Center for Biotechnology Information.
7. Kent, W.J. and Haussler, D. (2001). Assembly of the Working Draft of the Human Genome with GigAssembler. *Genome Res.* 11, 1541-1548.
8. Liu G, Loraine AE, Shigeta R, Cline M, Cheng J, Valmeekam V, Sun S, Kulp D, Siani-Rose MA (2003). NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res.* 31:82-6.
9. Zhang Z, Schwartz S, Wagner L, Miller W (2000). A greedy algorithm for aligning DNA sequences. *J Comput Biol.*7: 203-14.
10. Hubbard et al.(2002). The Ensembl genome database project. *Nucleic Acids Res.* 30:38-41.
11. Guy St. C. Slater and Ewan Birney (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31
12. Brigham H. Mecham, Gregory T. Klus,<sup>1</sup> Jeffrey Strovel,<sup>2</sup> Meena Augustus,<sup>2</sup> David Byrne,<sup>3</sup> Peter Bozso,<sup>3</sup> Daniel Z. Wetmore, Thomas J. Mariani, Isaac S. Kohane,<sup>3</sup> and Zoltan Szallas Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res.* 2004; 32(9): e74.



*Utilising Synamatix technologies to power its online bioinformatics application services.*

***CLICK HERE FOR YOUR 3-MONTH FREE TRIAL***

---

This Application Note is for distribution to Synamatix members, associate members and mailing list subscribers only. The contents are provided for personal, non-commercial purposes only and are protected by various national and international intellectual property laws, conventions and treaties. All title and intellectual property rights in and to Synamatix, SynaBASE, SynaMine, and SynaSuite and the accompanying printed materials are owned by Synamatix sdn bhd. Other trademarks or names are used only in an editorial fashion and to the benefit of the respective trademark owner with no intention of the infringement of the trademark. All trademarks or service marks are the property of their respective owners.