

SynaProbe™ - An Ultra High Speed Application for Whole Genome Microarray Probe Design

Zayed I. Albertyn, Wong Chee San, New Teong Chuan, Tay Liang Chung, M. Ramachandran, Ching Soo Meng, Arif Anwar, and Robert G. Hercus

The use of the world's 1st structured database, [SynaBASE™](#), that efficiently stores biological sequence patterns has seen the completion of a hyper-fast and accurate microarray probe design application: [SynaProbe™](#). SynaProbe was designed to accommodate the entire genome as a reference for designing probe sequences in addition to conventional references and parameters. Comparison of SynaProbe with current technologies using a breast cancer biomarker revealed that SynaProbe is able to find accurate probes thirty-six thousand times faster than published work. Probe specificity and sensitivity were verified using the Affymetrix HG-U133 Plus 2™ sequences and annotations as a control set. A separate analysis of the human genome showed that SynaProbe can be used to design probe sets for all known human genes in under 6 hours; this compares to the previous best benchmark which would take over 3 months. Furthermore, we were able to indicate that SynaProbe and SynaBASE technology can detect the same microarray probes in silico that have been experimentally validated by tissue-specific gene expression analyses. With unrivalled speed and scalability of SynaBASE technology the system can be applied to enhance the quality of probe design through enabling iteration and optimization for scientific researchers, which was previously improbable at whole or multiple genome scales.

Introduction

An organism's genome contains the central genetic blueprint outlining its architecture in terms of cellular structure, tissue organization and system biology. However, complexities lie in answering where, how and why genes are expressed as well as identifying factors that influence this process. DNA microarray technology addresses these questions by probing gene expression at the messenger RNA (mRNA) level [1]. By enabling the possibility of monitoring gene expression across an entire genome, both commercial and academic institutions are investing massive resources into implementing, validating and optimizing this technology. However, the promise of microarray-based discovery is plagued with inconsistencies centred on reproducibility, reliability and validity. In this paper we address one of the key variables that may affect some of the issues described above: microarray probe design.

Microarray probes are either synthetic DNA oligonucleotides or complementary DNA (cDNAs) routinely immobilized as a regular lattice network on a glass slide. These immobilized probes are gene

specific sequences that hybridize via Watson-Crick base pairing, to a solution containing a sample mRNA population. Quantitative and gene specific hybridization is usually detected via a fluorescent chemical reaction that is readily detected and quantified by laser excitation of the fluorophores. Therefore the activity of thousands of genes within an organism's genome can be measured in a single experiment to construct complex expression profiles. These techniques form the basis for understanding patterns in gene expression and protein function relating the molecular basis of diseases [2] e.g. detection of prostate cancer biomarkers for diagnostics [3].

There are inherent inefficiencies with microarray procedures encountered at most steps of experimental design, including, but not limited to, inaccurate reference sequence information; poor probe sensitivity; experimental reproducibility and statistical significance of expression results. Probes on the array need to be sufficiently sensitive yet specific to the gene for which they are designed [4].

The aim of designing accurate, high quality probes that are specific to their target genes has seen the development of a variety of computational algorithms for probe design [4]. Probe sequences need to be unique to enable specificity; display minimal or no secondary structure at hybridization temperatures and reside near the 3' terminal end of a gene. In this work we describe an ultra fast and highly accurate alternative approach to designing microarray probes based on structured pattern data stored in SynaBASE. The nature of SynaBASE's proprietary architecture allows comparison of a query sequence against all unique patterns derived from the reference database. An application of this concept in microarray probe selection would be the identification of the most unique and hence specific patterns from a gene by referencing against its entire genomic complement. We have successfully applied a new algorithm, SynaProbe, which accesses SynaBASE to design microarray probes for human genes and have assessed its accuracy by comparison to known standards from previously published benchmark studies [5].

Materials and Methods

Oligonucleotide Probes were designed using the Homo sapiens homeobox A9 (HOXA9) mRNA as input. This gene is located on Chromosome 7 and plays a role in breast cancer progression [4]. Probes were designed using two different tools; SynaProbe and Find Probe [4]. Both sets of probes were analysed and compared in parallel using the same metrics wherever possible. Affymetrix HG-U133 Plus 2 probe sets were also used as a control to verify probe specificity. To verify accuracy and probe uniqueness, SynaProbe and FindProbe probes were searched against SynaBASE versions of the human genome (NCBI33), Refseq mRNA and Affymetrix HG-U133 Plus 2 array probes. [SynaSearch™](#) results display sequence alignment matches as database frequency coloured patterns ranging from low to high frequency. BLAST was also used to conduct comparative analyses (data not shown).

SynaProbe, which is part of the [SynaSuite™](#) of applications, was run on Linux HP® Intel Itanium™ architecture using a single 1.3 GHz CPU accessing 64GB RAM.

Results and Discussion

In Silico Validation

SynaProbe finds the best probe candidates by selecting those patterns that are most unique within a SynaBASE of the human genome. Patterns may also be screened against other types of SynaBASE databases for SynaProbe e.g. human mRNA, expressed sequence tags (EST), all predicted open reading frames (ORFs), etc, depending on the aim and objectives of the user. We chose the human genome SynaBASE because it contains all possible unique sequence patterns that can be derived from the human genome, rather than focusing on the known transcriptome. Therefore our approach is unique in that any reference sequence database may be used for screening and selecting the best microarray probe set.

In addition to screening patterns for uniqueness and significance, SynaProbe calculates the weights from different parameters that are commonly used for probe selection thereby increasing specificity e.g. secondary structure, 3' position discrimination, low complexity, melting temperature, etc. (see Figure 1B on the next page) [4].

A single unique probe matching the NCBI33-Human Genome SynaBASE confirmed specificity, whilst matches with the correct gene from human Refseq mRNA SynaBASE showed probe sensitivity [6-7]. Searching Affymetrix HG-U133 Plus 2 SynaBASE with top-scoring probes was also used to confirm probe accuracy by referring to the corresponding gene details using the NetAffx Analysis Center at www.affymetrix.com [8]. SynaProbe found accurate and equivalent probe sets thirty-six thousand times faster than the computation time taken by FindProbe [4] (see Table 1 below).

Tool	Probe Sequence	Probe Start Position	SynaProbe Rank	SynaSearch Matches for Probe Query			Elapsed Time
				Human Genome	RefSeq mRNA	Affymetrix HG-U133 Plus 2	
SynaProbe	CCGCCAT TGGCCTA CTGTAGA TTTGAT CCTTGAT GAATCTG GGGTTTC CATCAGA CTGAAC TACACTG	1890	1	1(Chr7)	2(HOXA9)	5(209905_at)	531 milliseconds
FindProbe [5]	TGAAAC CGCCAT TGGGCT ACTGTAG ATTTGTA TCCTTGA TGAATCT GGGGTTT CCATCAG ACTGAAC TTA	1885	—	1(Chr7)	2(HOXA9)	5(209905_at)	5 hours

Table 1. SynaProbe and FindProbe result summary for the HOXA9 breast cancer progression gene.

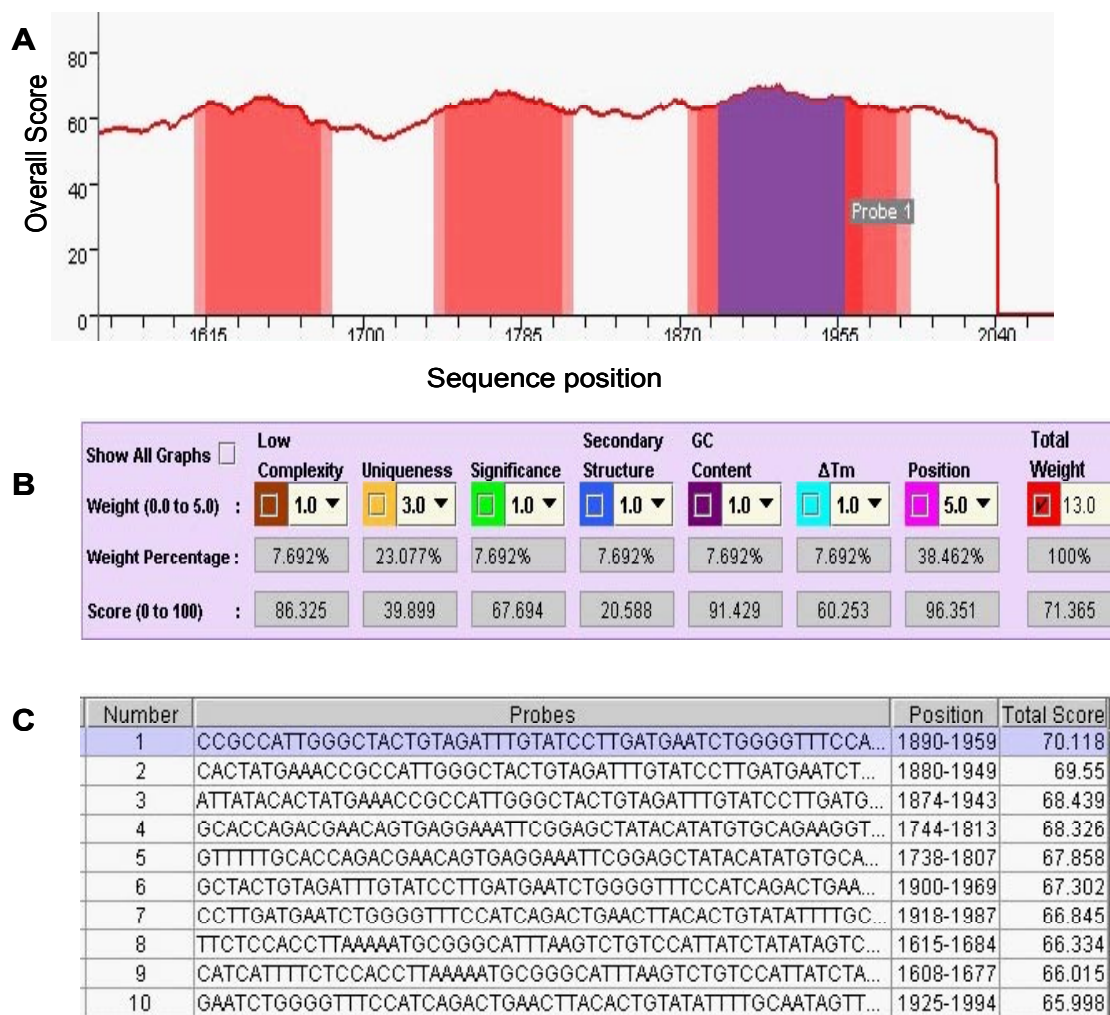


Figure 1. SynaProbe results for the HOXA9 breast cancer progression gene. The red highlighted regions correspond to probe locations on the sequence. Probes for the NM_152739 mRNA sequence are ranked (C) and scored according to the weighted parameters in B. Parameter weights run from a scale of 0 (no weight) to 5 (most weight). The highest ranking 70mer probe highlighted in A is the best probe that meets the criteria in B. SynaSearch may be launched from the SynaProbe results page to find unique genes, genome sequences or known probes.

SynaProbe typically finds a probe set with mean processing time ranging between 500 and 800 milliseconds. Therefore the technology permits researchers the ability to design, verify and parameterize the SynaProbe algorithm in matters of seconds per query. The advantages are that the process allows for iteration and refinement of probe quality for microarray expression experiments for all types of biological datasets where sequence information is available.

Single or multiple query sequences may be screened against all SynaBASE patterns for a genome, transcriptome or proteome. Probes were designed for an 11Kbp ubiquitin thiolesterase transcript, the largest characterized human cDNA according to the H-Invitational human cDNA database [9]. SynaProbe constructed a candidate probe set in 3.45 seconds. Using this example as the upper limit and the HOXA9 2Kbp gene as a further reference point (500ms), we extrapolate that at an average of 1 second per query SynaProbe would be able to construct probe sets for all 21037 human genes

in 5-6 hours of processing time on a single server. Therefore such scalability also permits probe quality optimization in a production environment for pharmaceutical and agricultural applications and new genomes.

Comparison to Experimental Probe Validation Methods

Further tests were conducted on an experimentally validated example to determine whether SynaProbe probes correlated with results from an actual microarray experiment. Shoemaker et al. devised a method for verifying gene prediction using microarray expression of genome-derived tiling arrays [10]. Based on their findings a novel transcript involved in testis specific expression on Chromosome 22 was discovered, EMBL AF324466 [10-11]. The transcript sequence corresponds to the Ensembl gene ENSG00000128346 and has the HG-U133 Plus 2 223706_at mapped to it [12]. SynaProbe selected a 25mer oligonucleotide for this transcript and a subsequent SynaSearch against the

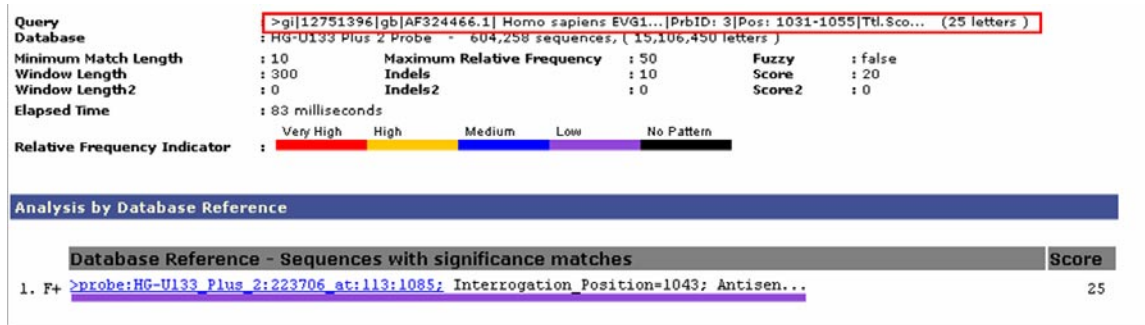


Figure 2. SynaSearch results showing the exact match of a SynaProbe sequence to the Affymetrix HG-U133 Plus 2 probe 223706_at [8]. The SynaProbe query and result header are highlighted in the red box. A SynaProbe result was searched against all Affymetrix HG-U133 Plus 2 sequences built in SynaBASE [8]. The database pattern frequencies are an indication of pattern uniqueness and can be used to guide analysis of results within SynaSuite.

HG-U133 Plus 2 SynaBASE revealed an exact match to HG-U133 Plus 2 223706_at (see Figure 2 above). Therefore SynaProbe finds probes that have been both experimentally validated and also incorporated into one of the world's most widely used microarray expression platforms.

Conclusions

The ability to store raw sequence information ranging from genomes to all human transcripts in SynaBASE provides a fast and accurate means to enhancing the quality of microarray probes. Although SynaBASE is able to filter low quality data it is still necessary to have a well curated reference set of genes and genomes for probe design. We have demonstrated that experimental evidence such as curated cDNA and existing probe annotations may be combined within SynaBASE for quality assessment of microarray probes. Pattern information derived from a whole genome version of SynaBASE is used as a reference to ensure probe specificity calculated by proprietary algorithms housed within the SynaBASE architecture.

The work presented in this paper shows that SynaProbe is a microarray probe design tool that produces results consistent with published *in silico* and experimental microarray benchmarks. With the added value of referencing patterns from multiple whole-genome databases and unrivalled speed, SynaProbe is an ideal tool to leverage design and optimization of probes for whole genome arrays, for both well validated and new genomes.

REFERENCES

1. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT and Solas D (1991). Light-directed, spatially addressable parallel chemical synthesis. *Science* 251, 767-73
2. Petricoin EF 3rd et al(2002). Medical applications of microarray technologies: a regulatory science perspective. *Nat Genet.* 32 Suppl,474-9. Review.

3. Luo J, Duggan DJ, Chen Y, Sauvageot J, Ewing CM, Bittner ML, Trent JM and Isaacs WB (2001). Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res.* 61, 4683-4688.
4. Tomiuk S and Hofmann K (2001). Microarray probe selection strategies. *Brief Bioinform.* 2001 4, 329-40.
5. Sung WK and Lee WH (2003) Fast and Accurate Probe Selection Algorithm for Large Genomes. *Proc. Comp. Syst. Bionform.* 65-74
6. Pruitt KD, Maglott DR (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* 29: 137-140.
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 and Siani-Rose Man (2003). NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res* 31, 82-6.
9. Gojobori et al. (2004). Integrative Annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones. *PLoS Biol.* [Epub ahead of print].
10. Shoemaker et al (2001). Experimental annotation of the human genome using microarray technology. *Nature* 409, 922-7.
11. Stoesser G et al (2002). The EMBL Nucleotide Sequence Database. *Nucleic Acids Res* 30, 21-6.
12. Clamp M et al (2003) Ensembl 2002: Accommodating comparative genomics. *Nucleic Acids Res* 31, 38-42.

To request detailed technical information or feedback please send an email to tech@synamatix.com

This Research Newsletter is for distribution to Synamatix members, associate members and mailing list subscribers only. The contents are provided for personal, non-commercial purposes 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, SynaSearch, SynaProbe 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.