Chapter 2
Mapping Affymetrix Microarray Probes to the Rat Genome via a Persistent Index

Susan Fairley
University of Glasgow, UK

John D. McClure
University of Glasgow, UK

Neil Hanlon
University of Glasgow, UK

Rob Irving
University of Glasgow, UK

Martin W. McBride
University of Glasgow, UK

Anna F. Dominiczak
University of Glasgow UK

Ela Hunt
University of Strathclyde, UK

ABSTRACT

A probe mapping technique using a novel implementation of a persistent q-gram index was developed. It guarantees to find all matches that meet certain definitions. These include exact matching of the central 19 bases of 25 base probes, matching the central 19 bases with at most one or three mismatches and exact matching of any 16 bases. In comparison with BLAST and BLAT, the new methods were either significantly faster or identified matches missed by the heuristics. The 16 bp method was used to map the 342,410 perfect match probes from the Affymetrix GeneChip Rat Genome 230 2.0 Array to the genome. When compared with the mapping from Ensembl, the new mapping included over seven million novel matches, providing additional evidence for researchers wishing to further investigate the sources of signals measured in microarray experiments. The results demonstrate the practicality of the index, which could support other q-gram based algorithms.

DOI: 10.4018/978-1-4666-1785-8.ch002
**INTRODUCTION**

In this paper two things are addressed. The first is the provision of a novel implementation of a persistent $q$-gram index using a relational database management system (RDBMS). This implementation allows a greater value of $q$ than has previously been reported. Further, it demonstrates the use of the RDBMS as a means of accessing the information stored on disk, the traditional bottleneck in the use of persistent indices.

Secondly, the index is applied to the task of mapping Affymetrix microarray probes, thereby demonstrating the practical use of the index implementation and providing a data set of use to those using the RAE 230 microarray. The advantage of our mapping approach is that it provides computational guarantees about the returned matches, making it possible to be certain that all matches meeting the chosen definition have been returned. Further, it is demonstrated that this can be achieved in reasonable time. The problems with the mappings provided by Affymetrix and the benefits of re-mapping are described below.

**Probe Mapping Background**

Probe mapping is essential to the interpretation of microarray experiments. A mapping should be current, accurate, and allow for some genetic variation, as in rat genomics various strains diverge from the published *Rattus norvegicus* sequence (Rat Genome Sequencing Consortium [RGSC], 2004). Ideally, a mapping should utilise criteria sufficient for a hybridisation signal to arise. These requirements have led to several investigations of the available mappings and to the creation of alternative mappings, summarised in Table 1.

Re-mapping has changed associations of probes to transcripts (Harbig, Sprinkle, & Enkemann, 2005; Leong, Yates, Wilson, & Miller, 2005; Stalteri & Harrison, 2007) and altered experimental interpretation (Dai et al., 2005; Lu & Zhang, 2006; Gautier, Mooller, Friis-Hansen, & Knudsen, 2004), with sequence-verification also improving measurement accuracy (Mecham et al., 2004b) and concordance between array platforms (Mecham et al., 2004a; Carter, Eklund, Mecham, Kohane & Szallasi, 2005; Severgnini et al., 2006; Barnes, Freudenberg, Thompson, Aronow, & Pavlidis, 2005) and generations (Elo et al., 2005). In addition, temporal inconsistencies (Perez-Iratxeta & Andrade, 2005) in NetAffx (Liu et al., 2003) have been highlighted. In NetAffx information centres on the representative sequence of each probeset, although problems with ‘Representative Public IDs’ have been identified (Dai et al., 2005), which makes it difficult to link ac-

<table>
<thead>
<tr>
<th>Group</th>
<th>Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affymetrix</td>
<td>Map target sequence to genome using BLAT. Keep only best alignment.</td>
</tr>
<tr>
<td>Ensembl (Hubbard et al., 2007)</td>
<td>Step 1 – Map probes to genome with Exonerate, allowing at most one mismatch. Step 2 – Map probesets to transcripts, requiring at least half the probes in a set to map to a transcript.</td>
</tr>
<tr>
<td>Harbig et al. (2005)</td>
<td>Step 1 - BLAST target sequences against RefSeqs. Step 2 - Map probes to the identified subset of RefSeq. Step 3 – Apply filters to link probesets to transcripts.</td>
</tr>
<tr>
<td>Dai et al. (2005)</td>
<td>Map probes to UniGenes and other sequences. Perfect matches only. One mapping to genome. Also, must match mRNA/EST sequence and genome perfectly. “Novel” probesets.</td>
</tr>
<tr>
<td>Mecham et al. (2004a, 2004b), Carter et al. (2005)</td>
<td>Mapping probes to UniGenes and to the cDNA sequences on cDNA probes for cross-platform comparison.</td>
</tr>
<tr>
<td>Kim et al. (2005)</td>
<td>Mapping probes to UniGenes using a $q$-gram supported batch implementation of BLAST.</td>
</tr>
</tbody>
</table>
Related Content

Introduction to Classification Error Estimation
www.igi-global.com/chapter/introduction-classification-error-estimation/53912?camid=4v1a

Personalized Disease Phenotypes from Massive OMICs Data
www.igi-global.com/chapter/personalized-disease-phenotypes-from-massive-omics-data/121465?camid=4v1a

Current Study Designs, Methods, and Future Directions of Genetic Association Mapping
www.igi-global.com/chapter/current-study-designs-methods-and-future-directions-of-genetic-association-mapping/121464?camid=4v1a

Bioinformatics-Driven Big Data Analytics in Microbial Research
www.igi-global.com/chapter/bioinformatics-driven-big-data-analytics-in-microbial-research/121462?camid=4v1a