Machine Learning in Low-level Microarray Analysis

Benjamin I. P. Rubinstein[1,2], Jon McAuliffe[3], Simon Cawley[4],
Marimuthu Palaniswami[5], Kotagiri Ramamohanarao[1], Terence P. Speed[2,3]

benr@ieee.org, jon@stat.berkeley.edu, simon_cawley@affymetrix.com,
swami@ee.mu.oz.au, rao@cs.mu.oz.au, terry@stat.berkeley.edu

[1] Department of Computer Science & Software Engineering, the University of Melbourne, Australia
[2] Division of Genetics & Bioinformatics, the Walter & Eliza Hall Institute of Medical Research, Australia
[3] Department of Statistics, University of California at Berkeley, CA
[4] Data Analysis Group, Affymetrix, Inc., Santa Clara, CA
[5] Department of Electrical & Electronic Engineering, the University of Melbourne, Australia

ABSTRACT

Machine learning and data mining have found a multitude of successful applications in microarray analysis, with gene clustering and classification of tissue samples being widely cited examples. Low-level microarray analysis – often associated with the pre-processing stage within the microarray life-cycle – has increasingly become an area of active research, traditionally involving techniques from classical statistics. This paper explores opportunities for the application of machine learning and data mining methods to several important low-level microarray analysis problems: monitoring gene expression, transcript discovery, genotyping and re-sequencing. Relevant methods and ideas from the machine learning community include semi-supervised learning, learning from heterogeneous data, and incremental learning.

Keywords

Low-level microarray analysis, gene expression estimation, genotyping, re-sequencing, transcript discovery, transductive learning, semi-supervised learning, learning from heterogeneous data, incremental learning

1. INTRODUCTION

DNA microarrays have revolutionized biological research over the short time since their inception [2; 27; 28; 29]. Although most widely used for parallel measurement of gene expression [27; 28], microarrays are starting to find common application in other areas of genomics and transcriptomics, including genomic re-sequencing [30; 31], genotyping [32; 33], and transcript discovery [34].

Research labs armed with microarrays have been able to partake in a range of studies, including finding gene function [35; 36; 37]; correcting mistaken database annotations [36; 7]; performing linkage analyses; determining specific genes involved in biological pathways; identifying genes that are important at certain times of development (or that are turned on/off over a course of treatment); elucidating gene regulatory networks [13]; diagnosing disease in tissue samples [38; 39; 40; 41]; and even identifying medical practitioners' misdiagnoses [38]. The common thread among these *high-level microarray analysis* problems is that they answer sophisticated questions of direct biological interest to medical researchers (such as "which genes are being co-expressed under treatment X?"), where the raw data used are estimates of biologically meaningful parameters (such as the expression level estimates for thousands of genes).

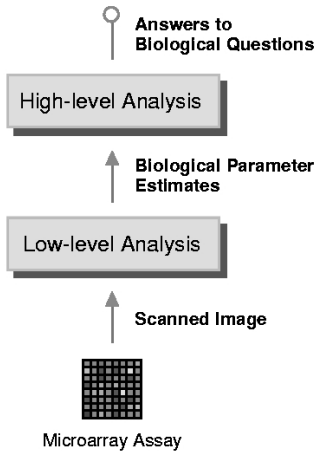

Figure 1: The relationship between low-level and high-level microarray analysis.

In contrast to these so-called high-level problems, *low-level microarray analysis* [19] is concerned with the preceding step in the microarray assay cycle (Figure 1) – given raw data straight from a scanner which has no direct biological interpretation, clean and summarize this data to produce the biologically meaningful parameter estimates (such as expression level estimates) that are later used in high-level analyses.

In low-level analysis, more consideration is generally given to the behavior of the underlying molecular biology, microarray technology, and experimental design than in high-level analysis. This makes generative methods readily applicable in low-level problems, facilitating the formulation of confidence

statements such as p-values in gene expression calls. Hence, while high-level problems have been tackled with discriminative approaches, such as those found in machine learning and data mining, in addition to classical statistical methods, the low-level analysis community has traditionally called upon only the latter.

In this paper we argue that low-level microarray analysis poses a number of interesting problems for the data mining and machine learning community, distinct to the traditional high-level microarray problems. These problems are relevant to the long-term success of DNA microarrays and are already topics of active research in the low-level microarray analysis community. It is our hope that this position paper motivates and enables further machine learning research in the area. Although we will focus on high density oligonucleotide microarrays, particularly those of the Affymetrix GeneChip variety, the underlying concepts and opportunities remain the same for related technologies. Throughout the paper, we distinguish machine learning from statistics. While these disciplines are closely related and serve as foundations for inference in microarray analysis, the distinction does have content. In our view, *classical statistics* is generative, dealing with relatively low-dimensional data and parameter spaces, while *machine learning* is often discriminative in nature and explicitly addresses computational issues in high-dimensional data analysis.

Section 2 reviews relevant background ideas from machine learning. For an overview of the background molecular biology and microarray technology, see the guest editorial elsewhere in this issue. The low-level problems of absolute and differential expression level summarization, expression detection, and transcript discovery are reviewed in Section 3, along with suggested applications of machine learning approaches to these problems. Sections 4 and 5 similarly cover microarray-based genotyping and re-sequencing. Finally, Section 6 concludes the paper.

2. BACKGROUND MACHINE LEARNING

We assume familiarity with the notions of unsupervised learning (clustering) and supervised learning (classification and regression). As many of the low-level analysis problems discussed below are amenable to learning from partially labeled data, learning from heterogeneous data, and incremental learning, we briefly review these paradigms here.

2.1 Learning from Partially Labeled Data

Given an i.i.d. *labeled sample* $\{(x_i, y_i)\}_{i=1}^{n}$ drawn from the unknown and fixed joint distribution $F(x, y)$, and an i.i.d. *unlabeled sample* $\{x_i\}_{i=n+1}^{m}$ drawn from the marginal distribution $F(x)$, the problem of *learning from partially labeled data* [22; 20] is to use the data in choosing a function $\hat{g}_m(X)$ approximating $\mathbb{E}(Y|X)$ where $(X, Y) \sim F$. This problem has been motivated by a number of applications where only limited labeled data is present, say due to expense, while unlabeled data is plentiful [16]. This is particularly the case in the areas of text classification, medical research, and computer vision [42], within which much of the research into learning from partially labeled data has occurred.

This problem, also called the *labeled-unlabeled data problem* [42], has been explored under a number of closely-related guises. Some of the earliest approaches used so-called *hybrid learners* [6], where an unsupervised learning algorithm assigns labels to the unlabeled data, thereby expanding the labeled dataset for subsequent supervised learning. The term *multimodal learning* is sometimes used to refer to partially labeled learning in the computer vision literature [17]. *Co-training* is a form of partially labeled learning where the two datasets may be of different types and one proceeds by using the unlabeled data to bootstrap weak learners trained on the labeled data [16].

More recently, *semi-supervised learning* [25] and *transductive learning* [26] have gained popularity. Equivalent to partially labeled learning, semi-supervised learning includes a number of successful algorithms, such as those based on the support vector machine (SVM) [25; 8]. Transductive learners, on the other hand, aim to predict labels for just the unlabeled data at hand, without producing the inductive approximation $\hat{g}_m$. This approach can be used to generalize the aforementioned hybrid learners, whose unsupervised step typically ignores the labeled data. In particular, it is shown in [26] that direct transduction is more effective than the traditional two-step approach of induction followed by deduction. A number of transductive schemes have been proposed, such as those based on the SVM [4; 25], a graph-based transductive learner [9], and a leave-one-out error ridge regression method [26]. Joachims [25] describes an approximate solver for the semi-supervised SVM which utilizes a fast SVM optimizer as an inner loop.

The story is not all good. [10] tells us that while unlabeled data may be useful, labeled examples are exponentially more valuable in a suitable sense. [43] tells us that unlabeled data may lead the transductive SVM to maximize the wrong margin, and in [42] it is shown that unlabeled data may in fact degrade classifier performance under certain conditions relating the risk and empirical risk. Nonetheless, learning from partially labeled data has enjoyed great success in many theoretical and empirical studies [16; 42; 44; 43].

We are especially interested in partially labeled learning as an approach to the low-level microarray analysis problems discussed in Sections 3–5, where we have relatively few labeled examples but an abundant source of unlabeled data. [45] is a recent example of partially labeled learning applied to high-level microarray analysis. There, the problem of predicting gene function is tackled using a semi-supervised scheme trained on a two-component dataset of DNA microarray expression profiles and phylogenetic profiles from whole-genome sequence comparisons. This leads us to the next relevant idea from machine learning.

2.2 Learning from Heterogeneous Data

Learning from heterogeneous data is the process of learning from training data, labeled or not, that can be partitioned into subsets, each of which contains a different type of data structure or originates from a different source. This notion is equivalent to the methods of *data fusion* [5].

Research into learning from heterogeneous data tends to be quite domain-specific and has enjoyed increasing interest from the bioinformatics community in particular (e.g., [18]). [46] presents a kernel-based framework for learning from heterogeneous descriptions of a collection of genes, proteins or other entities. The authors demonstrate the method's superiority to the homogeneous case on the problem of predicting yeast protein function using knowledge of amino acid sequence, protein complex data, gene expression data, and known protein-protein interactions.

[37] proposes an SVM method for classifying gene function from microarray expression estimates and phylogenetic profiles. This is achieved through the construction of an explicitly heterogeneous kernel: first separate kernels are constructed for each data type, taking into account high-order within-type correlations, then these kernels are combined, ignoring high-order across-type correlations.

Our interest in learning from heterogeneous data arises because several sources of knowledge relevant to low-level microarray analysis are available, and incorporating such problem domain knowledge has been shown to improve the performance of learning algorithms in the past.

2.3 Incremental Learning

Incremental learning is focused on learning from data presented sequentially, where the model may be required to make predictions on unseen data during training. This is in contrast to cases where all training occurs before any predictions are made (*batch learning*), and is similar to *online learning* [24].

A number of incremental learning algorithms have been proposed and applied in the literature. For example, several incremental support vector machines have been studied [24; 21; 47]. In [48], incremental learning is applied to distributed video surveillance. SVM algorithm parameter selection is investigated in [47]. [21] applies an incremental SVM to detecting *concept drift* – the problem of varying distributions over long periods of data gathering – and to adaptive classification of documents with respect to user interest. An exact incremental SVM is proposed in [24], where decremental unlearning of incremental training data is possible. This can be used to efficiently evaluate the computationally-expensive leave-one-out error measure.

Due to the relatively small sizes of datasets typically available in low-level microarray analysis, there is great potential for learners that can incrementally incorporate new data gathered in the lab, thereby improving estimator performance specific to that lab's patterns of microarray assay.

3. EXPRESSION ANALYSIS

The most successful application of DNA microarray technology to date has been to gene expression analysis. Traditionally, this has involved estimating gene expression levels (Section 3.1), an area that is being addressed through successful statistical methods and active statistics research. However, the task of determining transcription activity over entire chromosomes (Section 3.2) is less well developed and offers serious opportunities for machine learning.

3.1 Gene Expression Monitoring

3.1.1 The Problem

Traditional microarrays measure mRNA target abundance using the scanned intensities of fluorescence from tagged molecules hybridized to substrate-attached probes [29]. The brighter the intensity within a cell of identical probes, the more hybridization there has been to those probes (Figure 2a). The scanned intensity, then, roughly corresponds to target abundance.

Since probes are limited in length while targets may be thousands of bases long, the GeneChip uses a set of probes to detect each target nucleic acid. The probes are spread out along a 600 base pair region close to the 3' end of the transcript. To measure the effects of *cross-hybridization*, or unintended hybridization of target A to the probes intended for target B, a system of *probe pairs* is used. In each pair, a *perfect match* (PM) probe contains the target's exact complementary sequence, while a *mismatch* (MM) probe replaces the middle base of the perfect match probe with its Watson-Crick complement. In this way, a target is probed by a *probe set* of 11-20 PM-MM probe pairs. The aim is roughly for the PMs to measure signal plus noise and for the MMs to measure just noise, so that the signal is revealed using some function of (PM - MM). Figure 2b depicts the probe set arrangement, while Figure 2c gives an example of the scanned intensities. We may now define the expression level summarization problem.

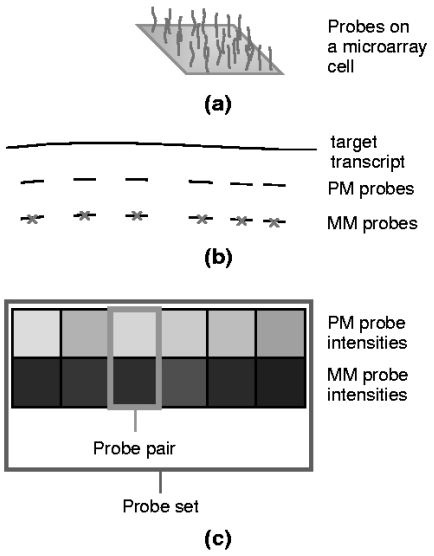

Figure 2: Probe-level features for expression level summarization: (a) a cell of probes; (b) target transcript, perfect match probe and mis-match probe sequences; and (c) scanned and image-analyzed probe-level intensities.

Low-level Problem 1. Given a probe set's intensities (possibly after background correction and normalization), the *expression level summarization problem* is to estimate the amount of target transcript present in the sample.

While the expression level summary aims to estimate gene expression level from the features of Figure 2, expression detection is concerned with determining the presence of any gene expression at all.

Low-level Problem 2. Given a probe set's intensities, possibly normalized, the *expression detection problem* is to predict whether the target transcript is present (P) or absent (A) in the sample, or otherwise call marginal (M) if it is too difficult to tell. In addition to the P/M/A *detection call*, we wish to state a confidence level in the call, such as a p-value.

Detection calls are not as widely utilized as expression level estimates. They are often used, for example, to filter out genes with negligible expression before performing computationally-expensive high-level analyses, such as clustering on gene expression profiles.

The previous two problems dealt with estimates based on a single probe-set read from a single array. *Comparative studies*, on the other hand, involve assaying two arrays, one the *baseline* and the other the *experiment*, followed by computation of a single comparative estimate.

Low-level Problem 3. Given two sets of intensities, possibly normalized, for the same probe set on two arrays:

a. The *differential expression level summarization problem* is to estimate the relative abundance of target transcript on each array.

b. The *comparison call problem* is to predict whether the expression of the target has increased, not changed, or decreased from one chip to the other. As in Low-level Problem 2, a statement of confidence in the call should be supplied.

The log-ratio of expression levels for a target is sometimes known as the *relative expression level* [3] and is closely related to the notion of *fold change* (which is sign(log-ratio) $\times$ $2^{\text{log-ratio}}$). Comparison calls are sometimes referred to as *change calls*. An advantage of working with these comparative estimates is that probe-specific affinities (one cause of undesired variation) are approximately cancelled out by taking ratios [3].

All of these problems are complicated by exogenous sources of variation which cloud the quantities we are interested in. [49] proposes a breakdown of the sources of variation in microarray experiments into intrinsic noise (variation inherent in the experiment's subjects), intermediate noise (arising for example from laboratory procedures), and measurement error (variation due to the instrumentation, such as array manufacture, scanning, or in silico processing).

3.1.2 Current Approaches

At the level of microarray design, sophisticated probe modeling and combinatorial techniques are used to reduce probe-specific effects and cross-hybridization. However, much of the unwanted variation identified above must still be tackled during low-level analysis. This means that care must be taken with the relevant statistical issues. For example, in experimental design, we must trade off between *biological replicates* (across samples) and *technical replicates* (one sample across chips). *Background correction* and *normalization*, for reducing systematic variation within and across replicate arrays, also surface as major considerations [19; 11].

Three popular approaches to Low-level Problem 1 [11] are the Affymetrix microarray suite (MAS) 5.0 signal measure [14; 3; 1], the robust multi-array average (RMA) [50; 11] and the model-based expression index (MBEI) [51].

MAS5 first performs background correction by subtracting a background estimate for each cell, computed by partitioning the array into rectangular zones and setting the background of each zone to that zone's second-percentile intensity. Next MAS5 subtracts an "ideal mismatch value" from each PM intensity and log-transforms the adjusted PMs to stabilize the variance. A robust mean is computed for the resulting values using a biweight estimator, and finally this value is scaled using a trimmed mean to produce the signal estimate.

RMA proceeds by first performing quantile normalization [52], which puts the probe intensity distributions across replicate arrays on the same scale. RMA then models the PMs as background plus signal, where the signal is exponentially and the background normally distributed – MM intensities are not used in RMA. A robust additive model is used to model the PM signal (in log-space) as the sum of the log scale expression level, a probe affinity effect, and an i.i.d. error term. Finally, median polish estimates the model parameters and produces the log-scale expression level summary.

MBEI fits $PM_{i,j} - MM_{i,j} = \theta_i \phi_j + \epsilon_{i,j}$, using maximum likelihood to estimate the per-gene expression levels θ_i. Here the ϕ_j are probe-specific affinities and the $\epsilon_{i,j}$ are i.i.d. normal errors.

Although it may seem that expression detection is just a matter of thresholding expression level estimates, this has proven not to be the case [53]. It is known that expression level estimators often have difficulty at low levels of expression, while detection algorithms are designed with this setting in mind.

The most widely used detection algorithm for the GeneChip is a method based on a Wilcoxon signed-rank test [54; 3; 55]. This algorithm corresponds to a hypothesis test of $H_0 : median(\frac{PM_i - MM_i}{PM_i + MM_i}) = \tau$ versus $H_1 : median(\frac{PM_i - MM_i}{PM_i + MM_i}) > \tau$, where τ is a small positive constant. These hypotheses correspond to absence and presence of expression, respectively. The test is conducted using a p-value for a sum of signed ranks $R_i = \frac{PM_i - MM_i}{PM_i + MM_i} - \tau$. The p-value is thresholded so that values in $[0, \alpha_1)$, $[\alpha_1, \alpha_2)$, and $[\alpha_2, 1]$ result in present, marginal, and absent calls, respectively. Here $0 < \alpha_1 < \alpha_2 < 0.5$ control the trade-off between false positives (FP) and true positives (TP).

Recently, a number of alternate rank sum-based algorithms have been proposed [53]. One in particular – a variant on the MAS5 method where scores are set to $R_i = \log \frac{PM_i}{MM_i}$ – has been shown to outperform MAS5 detection in a range of real-world situations. One aspect of the study in [53] of particular interest is the use of the Receiver Operating Characteristic (ROC) Convex Hull method [56] for comparing competing classifiers on a spike-in test set.

ROC curves (see Figure 3) characterize the classification performance of a family of classifiers parameterized by a tun-

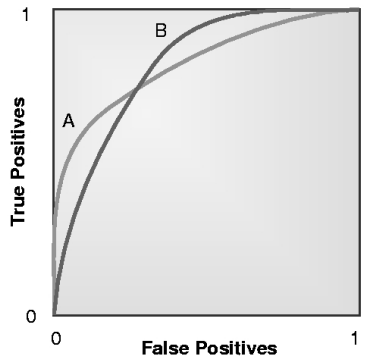

Figure 3: An ROC curve: $(0, 0)$ and $(1, 1)$ correspond to the "always negative" and "always positive" classifiers respectively. The closer to the ideal point $(0, 1)$ the better. Neither of the two families A or B dominates the other. Instead, one or the other is better according to the desired trade-off between FP and TP.

able parameter that controls the FP-TP trade-off. For example, as the level of a hypothesis test is decreased, the rate of false positive rejections decreases (by definition), while the rate of false negative acceptances will typically go up. An ROC curve encodes this trade-off, extending the notion of contingency table to an entire curve. It is a more expressive object than accuracy, which boils performance down to one number [56; 57].

Comparing ROC curves has traditionally been achieved by either choosing the "clear winner" (in the rare case of domination [57]), or choosing the maximizer of the Area Under Curve (AUC). Although AUC works in some cases, it gives equal credit to performance over all misclassification cost and class size settings – usually an undesirable strategy if any domain knowledge is available. The ROC Convex Hull method, on the other hand, relates expected-cost optimality to conditions on relative misclassification cost and class size, so that the typical case of semi-dominance (as in Figure 3) can be handled in a principled way – rather than selecting p-value thresholds by hand, end-users are provided with the right classifier and thresholds by the method. This use of the ROCCH method demonstrates a surprising application of machine learning to low-level microarray analysis.

Many of these *absolute expression* algorithms have their comparative analogues. For example, MAS5 produces the signal log ratio with an associated confidence interval, using a biweight algorithm [14; 3]. MAS5 also implements a comparison call based on the Wilcoxon signed-rank sum test, just as in the absolute MAS5 detection algorithm above [55].

While the Affymetrix microarray suite is the software package bundled with the GeneChip, the Bioconductor project [15] – an open-source set of R [12] packages for bioinformatics data analysis – has been gaining popularity and implements most of the methods discussed here.

3.1.3 Open Problems

While Low-level Problem 1 involves prediction of continuous expression levels (non-negative real values) given a vector of (non-negative real) perfect match and mismatch intensities, with total length between 22 and 40, Low-level Problem 2 is a 3-class classification problem with call confidence levels.

OPEN PROBLEM 1. *In the respective settings of Low-level Problems 1–3:*

a. *What machine learning techniques are competitive with algorithms based on classical statistical methods for expression level estimation?*

b. *Which machine learning classifiers are competitive for expression detection?*

c. *What machine learning methods achieve high performance on the comparative analogues of the previous two problems, posed on the appropriate product space of microarray measurements?*

Comparisons for expression level estimators might be made based on bias and variance, computational efficiency, and biological relevance of learned models. The ROCCH method is ideal for detector comparison. Issues of background correction and normalization across multiple arrays must likely also be addressed to enable competitiveness with the state of the art.

Research into applying semi-supervised, heterogeneous data and incremental learners to gene expression monitoring is directly motivated by the proportion of labeled to unlabeled data available, the existence of GeneChip domain knowledge, and the endemic nature of microarray assays that are continually performed in individual research labs. Biologists could augment the limited labeled probe-level data available with relatively abundant unlabeled data. Labeled data can be procured, for example, from bacterial control experiments with known concentrations, called *spike-in assays*, and bacterial control probe sets that are present in some GeneChips for calibration purposes. The former source of labeled data is the more useful for this problem, as it provides examples with a range of labels. Unfortunately, spike-in studies are rare because they are not of independent scientific interest: they are only performed for low-level microarray research. For the few spike-in assays that are available, only a small number of targets are spiked in at an equally small number of concentrations (typically $\approx$10). Unlabeled data, in contrast, could be taken from the large collection of available biologically relevant assays; each one providing tens of thousands of data points. Beyond probe intensities, other data sources could include probe sequences and probe-affinity information derived from probe models. Such information is closely related to the hybridization process and might be of use in expression level estimation: both target and non-specific hybridization are known to be probe-dependent. Although labeled data from spike-in studies are of greatest utility for learning [10], the quantity of unlabeled data produced by a series of biologically interesting microarray assays in any given lab suggests a semi-supervised incremental approach.

Since the ROCCH involves taking a pointwise maximum over the individual noisy ROC curves, it incorporates a possibly large degree of uncertainty. It should be possible to extend the results of [53] to quantify this property.

OPEN PROBLEM 2. *Can the ROC Convex Hull method of [56] be extended to provide confidence intervals for its conditions on expected-cost optimality?*

3.2 Transcript Discovery

3.2.1 The Problem

The applications to expression monitoring described above are all related to addressing questions about pre-defined transcripts. More precisely, the vast majority of expression analysis is performed using probes interrogating only a small sub-sequence of each transcript. This has clearly been a useful approach, but there are at least two potential drawbacks. One is that we can only monitor the expression of genes known to exist at the time of the array's design. Even in a genome as well-studied as that of the human, new transcripts are routinely discovered. Another is that in directly monitoring only a sub-sequence of the transcript, it will often be impossible to distinguish between alternatively spliced forms of the same gene (which may have very different functional roles).

An alternative approach is to use arrays with probes tiled uniformly across genomic sequence, without regard to current knowledge of transcription. Such *genome tiling arrays* have been used to monitor expression in all the non-repetitive sequence of human chromosomes 21 and 22 [34], and more widespread use is underway.

The problems arising in the analysis of data from genome tiling arrays are essentially the same as those for the expression monitoring arrays described above: estimation of expression level, detection of presence, and detection of differential expression. There is, however, the additional challenge of determining the number of distinct transcripts and their location within the tiled genomic region.

Low-level Problem 4. The problem of *transcript discovery* can be viewed in two steps:

a. Determining the exon structure of genes within a tiled region; and

b. Determining which exons should be classified together as part of a single gene's transcript.

3.2.2 Current Approaches

A simple heuristic approach is taken in [34], in which PM-MM probe pairs are classified as positive or negative based on thresholds applied to the difference and ratio of the PM and MM values. Positions classified as positive and located close to other positive positions are grouped together to form predicted exons.

A more effective approach [58] is based on the application of a Wilcoxon signed-rank test in a sliding window along the genomic sequence, using the associated Hodges-Lehmann estimator for estimation of expression level. Grouping into exons is achieved by thresholding on present call p-values or estimated expression level, then defining groups of probes exceeding the threshold to be exons.

3.2.3 Open Problems

The problem of detecting exons based on probe intensities (Low-level Problem 4a) is very similar to the problem of absolute expression detection (Low-level Problem 2). For example, the exon detection method of [58] and the MAS5 expression detection algorithm [55] are both built around the Wilcoxon signed-rank test. The problem of finding exons has been addressed as described, but the methods are heuristic and there is plenty of room for improvement. Associating exons to form transcripts (Low-level Problem 4b) has been addressed in a large experiment across almost 70 experimental pairs using a heuristic correlation-based method; again, this presents an opportunity for research into more effective methods.

Open Problem 3. *Are there machine learning methods that are able to out-perform current classical statistical methods in transcript discovery as defined in Low-Level Problem 4?*

One possibility which appears well suited to the problem is the use of hidden Markov models where the underlying unobserved Markov chain is over states representing expressed versus non-expressed sequence. The distribution of the observed probe intensities would depend on the underlying hidden state. Another possible approach, considering the success which has been demonstrated in predicting genes from sequence data alone, would also be to integrate array-derived data with sequence information in prediction of transcripts.

4. GENOTYPING

4.1 The Problem

Descriptions of genome sequencing efforts such as the human genome project often lend the impression that there is a unique genomic sequence associated with each species. This is a useful and approximately correct abstraction. But in fact, any two individuals picked at random from a species population will have differing nucleotides at a small fraction of the corresponding positions in their genomes. Such *single-nucleotide polymorphisms*, or SNPs, help form the basis of genetically-determined variation across individuals. Biologists estimate that about one position in 1,000 in the human genome is a SNP. With over 3 billion bases of genomic DNA, we see that SNPs number in the several millions. Although there are other kinds of individual genomic variation, such as insertions, deletions, and duplications of DNA segments, our focus here is SNPs.

Further complicating the picture is the fact that humans are *diploid* organisms—each person possesses two complete but different copies of the human genome, one inherited from the mother and one from the father. Now consider a polymorphic position, or *locus*, at which two different bases occur in the population, say G and T. These variants are called the *alleles* at the locus, so in this case we are describing a *biallelic* SNP. A given individual will have inherited either a G or T in the paternal genome, and the same is true of the maternal genome. Thus there are three possible *genotypes*, or individual genetic signatures, at this SNP: they are denoted GG, TT, and GT. We do not distinguish the last case from TG, since there is no inherent ordering of the paternal and maternal genomes at a given polymorphic position.

We refer generically to the alleles of a biallelic SNP as A and B. Biological evidence suggests that essentially all SNPs are biallelic in humans. The *genotyping problem*, then, is to establish an individual's genotype as AA, BB, or AB for as many SNPs as possible in the human genome. The completion of the human genome project means that one has recourse to the full genomic sequence surrounding a SNP to help solve the genotyping problem. Furthermore, various large-scale public projects to locate SNPs and identify their alleles exist, notably The SNP Consortium (TSC); the data they generate may also be utilized for genotyping.

The major drawback to traditional genotyping protocols are their lack of parallelism, with consequent expense in terms of material and labor. In contrast, Kennedy et al. [33] describe whole-genome sampling analysis (WGSA), which enables massively parallel genotyping via *genotyping microarrays*.

For the Affymetrix Mapping 10k Array, which genotypes approximately 10,000 SNPs across the human genome, each SNP actually has 56 corresponding probes, collectively termed a *miniblock*. The miniblock has 7 *probe quartets* for the SNP's flanking region on the forward strand and another 7 probe quartets for the reverse complement strand, so $4 \times 7 \times 2$ yields 56 probes. Each probe quartet in turn corresponds to a 25-mer in which the SNP is at one of 7 offsets from the central position. The four probes within a probe quartet differ in the base they put at the SNP: a perfect match to the A allele, a perfect match to the B allele, and mismatches for each.

Low-level Problem 5. Given a SNP's 56-vector of miniblo-

ck probe intensities, the *genotype calling* problem is to predict the individual's corresponding alleles as AA, BB or AB.

Write PM(A), PM(B), MM(A), and MM(B) for the probe intensities within a quartet. We would then hope that an AA individual has PM(A) > MM(A) but PM(B) ≈ MM(B), for all probe quartets on both strands. For a BB individual, we hope to find just the opposite effect, and an AB individual should have both PM(A) > MM(A) and PM(B) > MM(B). The mismatch probes in each quartet act as controls, establishing the level of nonspecific hybridization for their corresponding perfect match probes. The presence of multiple probe quartets allows for the determination of genotype even when one strand and/or some offsets do not yield reliable hybridization, say for biochemical reasons.

4.2 Current Approaches

Low-level Problem 5 is a three-class classification problem. In many machine learning applications, the metric of interest for competing classifiers is predictive accuracy, in this case the probability of correctly genotyping a new individual's SNP based on the miniblock vector. However, in the kinds of genetic studies which take large numbers of genotypes as input, there is usually an explicit requirement that genotype predictions have a prespecified accuracy, often 99%. To attain such accuracy, it is usually permissible for some fraction of genotypable SNPs to be *no-calls*; that is, the classifier can refuse to predict a genotype for some miniblocks. When comparing genotypers, our interest therefore lies in the trade-off between the rate of no-calls and the accuracy attained on those SNPs which are called. For example, some studies consider the *punt rate*, or lowest no-call rate which yields a prespecified accuracy level on the called SNPs.

A simple unsupervised approach to training a genotyper is to ignore available labels during training, instead using these labels to subsequently assess the trade-off between accuracy and no-call rate for the trained model. This is the strategy pursued by MPAM (*modified partitioning around medoids*) [59], the discriminative clustering genotyper used for the Affymetrix 10k array. An alternative approach, using a parametric generative model for the clustering, will be described elsewhere. It resembles ABACUS, a model studied in the context of re-sequencing microarrays [31] (see Section 5).

4.3 Open Problems

Open Problem 4. *Are there machine learning methods that are able to meet typical accuracy and punt-rate specifications on the genotype calling problem?*

In order to choose a genotyper using supervised learning, we need labels (true genotypes) along with corresponding miniblock reads from genotyping arrays. Unfortunately, there is no large-scale set of publicly available genotypes. Instead, one makes do with modestly-sized sets of genotypes available commercially from companies using smaller-scale techniques. Of course, no genotyping method is error-free, so in practice one measures *concordance* with reference genotypes. If the concordance is high enough, the remaining cases of disagreement between a candidate genotyper and the reference genotypes can be resolved via the older labor-intensive methods. The incomplete nature of reference genotype data leads naturally to the setting of semi-supervised learning. Rather than falling back to unsupervised methods such as those described above, we may consider employing more general semi-supervised learners as described in Section 2.1. Additionally, the methods of [23] could be used to incorporate low-level physical parametric models of hybridization into a kernel-based classifier.

5. RE-SEQUENCING

5.1 The Problem

As explained in Section 4, within a single species genomic sequence will vary slightly from one individual to the next. While Low-level Problem 5 focuses on the determination of genotype at a position known in advance to be polymorphic, the problem described in this section concerns locating such polymorphic sites in the first place.

The usual starting point is a newly-sequenced genome, such as the recently-finished human genome. It is often the case that, based on previous research, an investigator will be interested in detailed study of variation in a particular genomic region (say on the order of tens or hundreds of kilobases) and wants to *re-sequence* this region in a large number of individuals. Such re-sequencing allows for identification of the small subset of polymorphic locations. Here we consider the more recent challenges of microarray-based re-sequencing of diploid genomic DNA.

A typical re-sequencing array uses eight probes to interrogate each base of the monitored sequence. These eight probes comprise two quartets, one for the forward strand and one for the reverse. Each quartet is formed of 25-mer probes perfectly complementary to the 25 bases of the reference sequence centered on the interrogated base, but with all four possible bases used at the central position.

Low-level Problem 6. The goal of the *re-sequencing problem* is to start with a set of probe intensities and classify each position as being one of A, C, G, T, AC, AG, AT, CG, CT, GT, or N, where N represents a 'no call' (due to sample failure or ambiguous data).

The intuition is that for a *homozygous* position, one of the four probes should be much brighter relative to the others on each strand, and for a *heterozygous* position, two probes corresponding to the two bases of a SNP should be brighter on each strand. Of particular interest are positions in which the called base is heterozygous, or homozygous and different to the reference sequence, as such positions exhibit polymorphism and are candidate positions for explaining phenotypic differences between individuals.

At face value, this classification problem is much harder than the genotyping problem. There are fewer probes to start with (a miniblock of 8 rather than 40 or more) and more categories (11 as opposed to 3 or 4) into which to classify.

5.2 Current Approaches

The most recent analysis of the kind of re-sequencing array discussed here [31] is based on modeling pixel intensities within each probe as independent random variables with a common mean and variance. The model for a homozygous base is that, on each strand, the probe corresponding to the base has one mean and variance, and the other three probes have another. The means and variance are estimated by maximum likelihood, and the likelihood of the

model is evaluated. The model for each of the six heterozygous possibilities is similar, except two probes correspond to each heterozygote model and the other two are background. The likelihoods (overall and for each strand) are converted to scores and, provided the maximum score exceeds some threshold, the best-scoring model is chosen as the base call. A number of other filters that deal with the signal absence, signal saturation, sample failure, and so on are applied, as is an iterative procedure to account for bias in the background probes. This method, called ABACUS, was found to make base calls at over 80% of all bases, with an estimate accuracy in excess of 99% at the bases which were called.

5.3 Open Problems

A good base-calling method for re-sequencing arrays already exists in ABACUS, but there remains room for improvement. A recent and improved implementation [60] of the ABACUS method on a new genomic region found the overall sequencing accuracy to be on the order of 99.998%, but the accuracy on heterozygote calls to be about 96.7%. Biologists would value highly an improvement in heterozygote call accuracy.

Open Problem 5. *Can a supervised learning method be used to call bases in re-sequencing arrays with accuracy, in particular heterozygote accuracy, in excess of the accuracies achieved by the more classic statistical approaches used to date?*

Considering the ongoing efforts of SNP detection projects, there is an abundance of labeled data available, so the problem seems quite amenable to machine learning approaches. As with the genotyping problem, it would be desirable to have a measure of confidence associated with base calls. It may also be useful to take into account the sequences of the 25-mer probes, as there are known sequence-specific effects on the probe intensities.

6. CONCLUSIONS

We have described a variety of low-level problems in microarray data analysis and suggested the applicability of methods from several areas of machine learning. Some properties of these problems which should be familiar to machine learning researchers include high-dimensional observations with complicated joint dependencies (probe intensities), partially labeled data sets (expression levels, genotypes), data from disparate domains (microarray assays, probe sequences, phylogenetic information), and sequential observations (ongoing experimental work at individual labs). We pointed out the suitability of semi-supervised, heterogeneous, and incremental learning in these settings. It is worth remarking that analogous problems arise with other high-throughput technologies, such as cDNA and long oligonucleotide microarrays, mass spectrometry, and fluorescence-activated cell sorting.

There are other issues in low-level analysis we did not cover. Here we mention two of these. Image analysis is the problem of going from raw pixel values in the scanned image of a microarray to a set of pixel intensities for each feature placed on the probe, and then to single-number probe intensities. The surface of the GeneChip contains detectable grid points which facilitate rotation and translation of the image to a canonical alignment; subsequent mapping of each pixel to a feature is semi- or fully automated and has not previously raised major analysis issues. However, work is being done on aggressive reduction of feature sizes to a scale where this mapping procedure could become a central concern.

On the more theoretical side, probe models based on the physics of polymer hybridization have recently been the focus of considerable interest. These models reflect a significant increase in the use of biological knowledge for estimating target abundance and present an opportunity for application of machine learning techniques which can exploit parametric distributions in high-dimensional data analysis, such as graphical models.

We close by observing that a fuller awareness of low-level microarray analysis issues will also benefit machine learning researchers involved with high-level problems: the inevitable information reduction from earlier stage to later could well conceal too much of what the unfiltered array data reveal about the biological issue at hand. Familiarity with initial normalization and analysis methods will allow the high-level analyst to account for such a possibility when drawing scientific conclusions.

7. ACKNOWLEDGMENTS

We thank Rafael Irizarry, Ben Bolstad, Francois Collin and Ken Simpson for many useful discussions and collaboration on low-level microarray analysis.

8. REFERENCES

[1] Affymetrix. *Affymetrix Microarray Suite Guide*. Affymetrix Inc., Santa Clara, CA, 2001. version 5.0.

[2] M. Schena. *DNA Microarrays: A Practical Approach*. Oxford University Press, 1999.

[3] Affymetrix. Statistical algorithms description document. Whitepaper, Affymetrix Inc., Santa Clara, CA, 2002.

[4] A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In *Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 148–155. Morgan Kaufmann Publishers, 1998.

[5] P. K. Varshney. Scanning the issue: Special issue on data fusion. *Proceedings of the IEEE*, 85:3–5, 1997.

[6] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, New York, 1973.

[7] T. Gaasterland and S. Bekiranov. Making the most of microarray data. *Nature Genetics*, 24:204–206, 2000.

[8] K. P. Bennett and A. Demiriz. Semi-supervised support vector machines. In *Advances in Neural Information Processing Systems 11*, pages 368–374, Cambridge, MA, 1999. MIT Press.

[9] T. Joachims. Transductive learning via spectral graph partitioning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2003.

[10] V. Castelli and T. Cover. On the exponential value of labeled samples. *Pattern Recognition Letters*, 16:105–111, 1995.

[11] R. A. Irizarry. *Science and Statistics: A Festschrift for Terry Speed*, volume 40 of *Lecture Notes–Monograph Series*, chapter Measures of gene expression for Affymetrix high density oligonucleotide arrays, pages 391–402. Institute of Mathematical Statistics, 2003.

[12] R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

[13] N. Friedman. Probabilistic models for identifying regulation networks. *Bioinformatics*, 19:II57, October 2003.

[14] E. Hubbell, W. M. Liu, and R. Mei. Robust estimators for expression analysis. *Bioinformatics*, 18:1585–1592, 2002.

[15] Bioconductor Core. An overview of projects in computing for genomic analysis. Technical report, The Bioconductor Project, 2002.

[16] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Workshop on Computational Learning Theory*. Morgan Kaufmann Publishers, 1998.

[17] L. Wu, S. L. Oviatt, and P. R. Cohen. Multimodal integration - a statistical view. *IEEE Transactions on Multimedia*, 1:334 –341, 1999.

[18] A. J. Hartemink and E. Segal. Joint learning from multiple types of genomic data. In *Proceedings of the Pacific Symposium on Biocomputing 2004*, 2004.

[19] G. K. Smyth, Y. H. Yang, and T. P. Speed. *Functional Genomics: Methods and Protocols*, volume 224 of *Methods in Molecular Biology*, chapter Statistical issues in cDNA microarray data analysis, pages 111–136. Humana Press, Totowa, NJ, 2003.

[20] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *International Conference on Machine Learning (ICML)*, 2001.

[21] R. Klinkenberg and T. Joachims. Detecting concept drift with support vector machines. In P. Langley, editor, *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pages 487–494, Stanford, CA, 2000. Morgan Kaufmann Publishers.

[22] M. Szummer and T. Jaakkola. Partially labeled classification with Markov random walks. In *Neural Information Processing Systems (NIPS)*, 2001.

[23] T. S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11: Proceedings of the 1998 Conference*, pages 487–493. MIT Press, 1998.

[24] G. Cauwenberghs and T. Poggio. Incremental and decremental support vector machine learning. In *NIPS*, pages 409–415, 2000.

[25] T. Joachims. Transductive inference for text classification using support vector machines. In I. Bratko and S. Dzeroski, editors, *Proceedings of the 16th Annual Conference on Machine Learning*, pages 200–209. Morgan Kaufmann, 1999.

[26] O. Chapelle, V. Vapnik, and J. Weston. *Advances in Neural Information Processing Systems 12*, chapter Transductive inference for estimating values of functions. MIT Press, 2000.

[27] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.

[28] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

[29] R. J. Lipshutz, S. P. A. Fodor, T. R. Gingeras, and D. H. Lockhart. High density synthetic oligonucleotide arrays. *Nature Genetics*, 21:20–24, 1999. Supplement.

[30] J. B. Fan, D. Gehl, L. Hsie, K. Lindblad-Toh, J. P. Laviolette, E. Robinson, R. Lipshutz, D. Wang, T. J. Hudson, and D. Labuda. Assessing DNA sequence variations in human ests in a phylogenetic context using high-density oligonucleotide arrays. *Genomics*, 80:351–360, September 2002.

[31] D. J. Cutler, M. E. Zwick, M. M. Carrasquillo, C. T. Yohn, K. P. Tobin, C. Kashuk, D. J. Mathews, N. A. Shah, E. E. Eichler, J. A. Warrington, and A. Chakravarti. High-throughput variation detection and genotyping using microarrays. *Genome Research*, 11:1913–1925, November 2001.

[32] J. B. Fan, X. Chen, M. K. Halushka, A. Berno, X. Huang, T. Ryder, R. J. Lipshutz, D. J. Lockhart, and A. Chakravarti. Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome Research*, 10:853–860, June 2000.

[33] G. C. Kennedy, H. Matsuzaki, D. Dong, W. Liu, J. Huang, G. Liu, X. Su, M. Cao, W. Chen, J. Zhang, W. Liu, G. Yang, X. Di, T. Ryder, Z. He, U. Surti, M. S. Phillips, M. T. Boyce-Jacino, S. P. A. Fodor, and K. W. Jones. Large-scale genotyping of complex DNA. *Nature Biotechnology*, October 2003.

[34] P. Kapranov, S. E. Cawley, J. Drenkow, S. Bekiranov, R. L. Strausberg, S. P. A. Fodor, and T. R. Gingeras. Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, 296:916–919, 2002.

[35] M. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. Sugnet, M. Ares Jr, and D. Haussler. Support vector machine classification of microarray gene expression data. Technical Report UCSC-CRL-99-09, Department of Computer Science, University of California at Santa Cruz, 1999.

[36] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares Jr, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97:262–267, 1997.

[37] P. Pavlidis, J. Weston, J. Cai, and W. N. Grundy. Gene functional classification from heterogeneous data. In *Proceedings of the Fifth International Conference on Computational Molecular Biology*, pages 242–248, 2001.

[38] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16:906–914, 2000.

[39] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J. P. Mesirov, and T. Poggio. Support vector machine classification of microarray data. Technical Report 182, Center for Biological and Computational Learning Massachusetts Institute of Technology, 1998.

[40] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98, 2001.

[41] C. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. M. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, and T. Golub. Molecular classification of multiple tumor types. *Bioinformatics*, 1:1–7, 2001.

[42] F. G. Cozman and I. Cohen. Unlabeled data can degrade classification performance of generative classifiers. In *Fifteenth International Florida Artificial Intelligence Society Conference*, pages 327–331, 2002.

[43] T. Zhang and F. J. Oles. A probability analysis on the value of unlabeled data for classification problems. In *Proceedings of the International Conference on Machine Learning*, pages 1191–1198, 2000.

[44] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39:103–134, 2000.

[45] T. Li, S. Zhu, Q. Li, and M. Ogihara. Gene functional classification by semi-supervised learning from heterogeneous data. In *Proceedings of the ACM Symposium on Applied Computing*, 2003.

[46] G. R. G. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble. Kernel-based data fusion and its application to protein function prediction in yeast. In *Proceedings of the Pacific Symposium on Biocomputing 2004*, 2004.

[47] A. Shilton, M. Palaniswami, D. Ralph, and A. C. Tsoi. Incremental training in support vector machines. In *Proceedings of the International Joint Conference on Neural Networks*, 2001.

[48] C. P. Diehl. *Toward Efficient Collaborative Classification for Distributed Video Surveillance*. PhD thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, 2000.

[49] J. H. Maindonald, Y. E. Pittelkow, and S. R. Wilson. *Science and Statistics: A Festschrift for Terry Speed*, volume 40 of *IMS Lecture Notes–Monograph Series*, chapter Some Considerations for the Design of Microarray Experiments, pages 367–390. Institute of Mathematical Statistics, 2003.

[50] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264, 2003.

[51] C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Science*, 98:31–36, 2001.

[52] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, 19:185–193, 2003.

[53] B. I. P. Rubinstein and T. P. Speed. Detecting gene expression with oligonucleotide microarrays, 2003. manuscript in preparation.

[54] W. Liu, R. Mei, D. M. Bartell, X. Di, T. A. Webster, and T. Ryder. Rank-based algorithms for analysis of microarrays. *Proceedings of SPIE, Microarrays: Optical Technologies and Informatics*, 4266, 2001.

[55] W. M. Liu, R. Mei, X. Di, T. B. Ryder, E. Hubbell, S. Dee, T. A. Webster, C. A. Harrington, M. H. Ho, J. Baid, and S. P. Smeekens. Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, 18:1593–1599, 2002.

[56] F. Provost and T. Fawcett. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Third International Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA, 1997.

[57] T. Fawcett, F. Provost, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Fifteenth International Conference on Machine Learning*, 1998.

[58] D. Kampa and et al. Novel RNAs identified from a comprehensive analysis of the transcriptome of human chromosomes 21 and 22. Manuscript in preparation.

[59] W.-M. Liu, X. Di, G. Yang, H. Matsuzaki, J. Huang, R. Mei, T. B. Ryder, T. A. Webster, S. Dong, G. Liu, K. W. Jones, G. C. Kennedy, and D. Kulp. Algorithms for large scale genotyping microarrays. *Bioinformatics*, 2003. In press.

[60] Affymetrix. GeneChip CustomSeq resequencing array: Performance data for base calling algorithm in GeneChip DNA analysis software. Technical note, Affymetrix Inc., Santa Clara, CA, 2003. http://www.affymetrix.com/support/technical/technotes/customseq_technote.pdf.