Probabilistic analysis of probe performance on short nucleotide arrays

L. Lahti[1], L.L. Elo[2,3], T. Aittokallio[2,3,4] and S. Kaski[1]

[1] Helsinki Institute for Information Technology and Adaptive Informatics Research Centre, Laboratory of Computer and Information Science, Helsinki University of Technology, Finland. [2] Department of Mathematics, University of Turku, Finland. [3] Turku Centre for Biotechnology, Finland. [4] Systems Biology Unit, Institut Pasteur, Paris, France

Introduction

Gene expression arrays are used to monitor gene activity in a given sample. Varying performance of individual probes is a main source of noise for these arrays. While modeling of probe effects has been shown to improve the estimation of gene expression, tools for analyzing probe performance have been missing. We propose a probabilistic model for this task.

Probe-level problems

Short oligonucleotide arrays of Affymetrix utilize multiple probes per gene (Fig. 1).

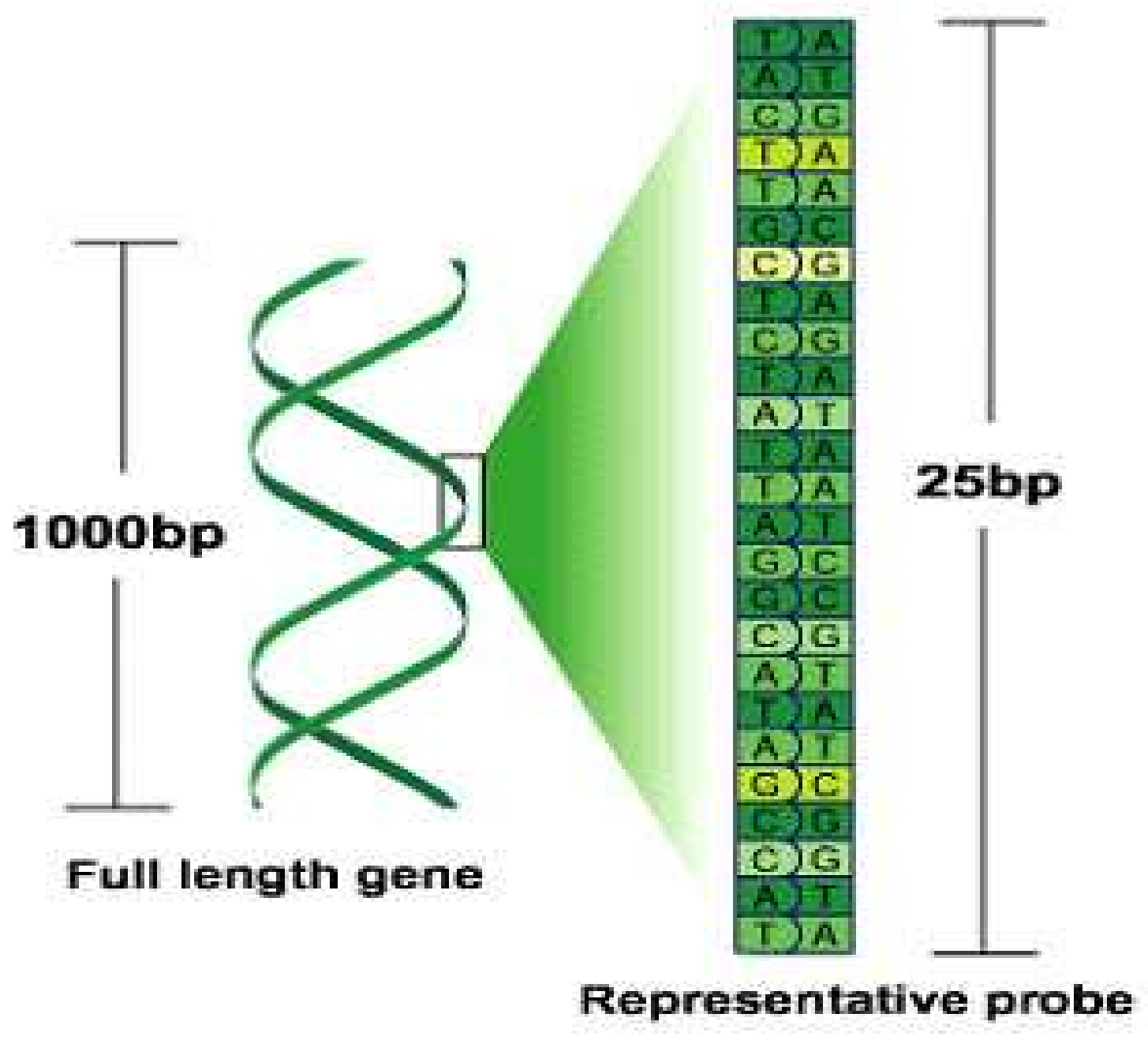

Figure 1 Principle of microarray technology (www.affymetrix.com). Each probe is complementary to a specific position on the target sequence. Probes are designed to uniquely bind with the target mRNA and measure its expression.

Bad probes add noise to data (Fig. 2). Understanding probe-related noise can help to improve array design and preprocessing, and to avoid misleading interpretations.

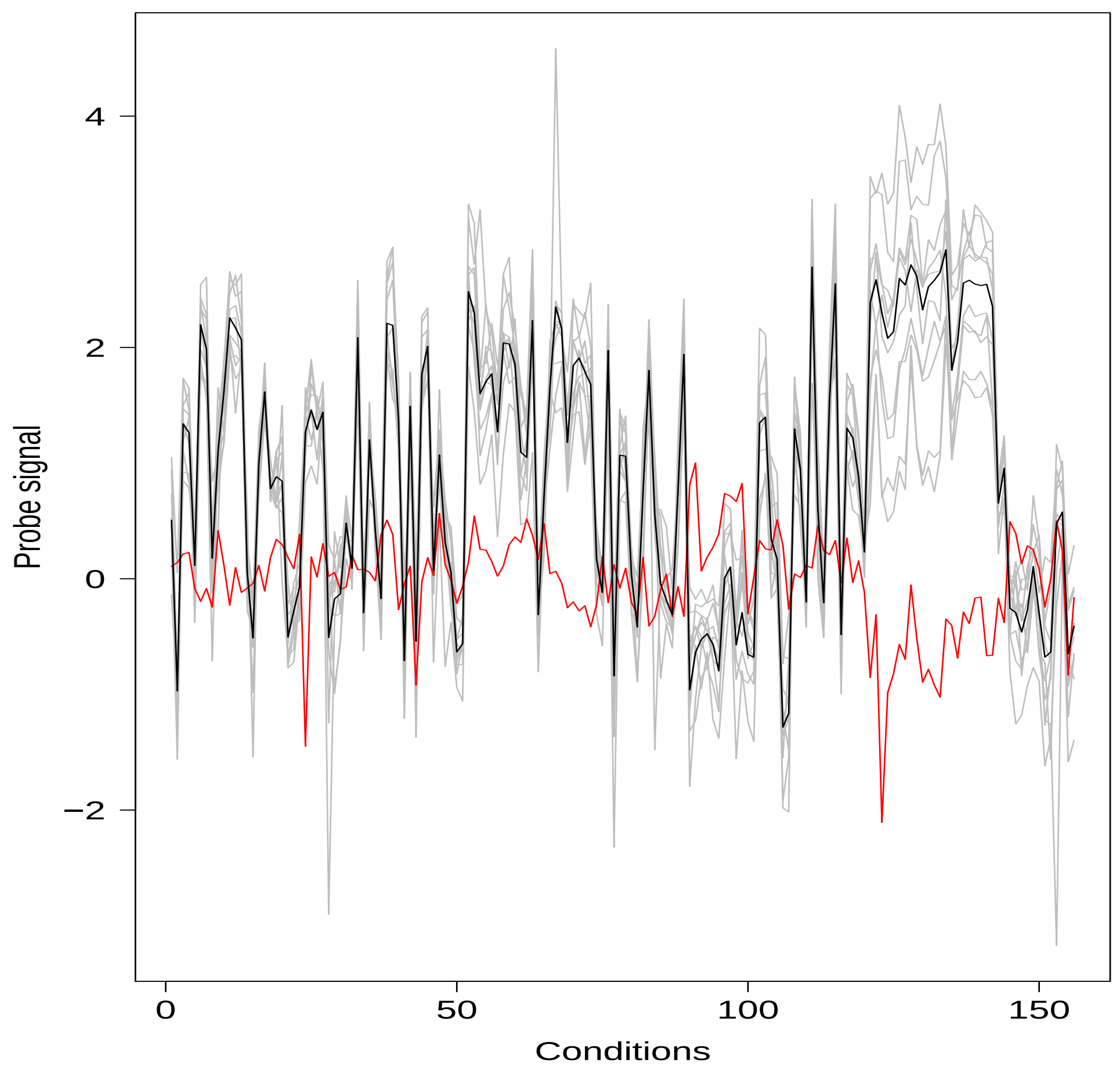

Figure 2 Probe-level observations (grey) of differential gene expression in various conditions for probe set '215157_x_at' (GEA-133). Black curve is an estimate of the **probe-set level signal**, and red curve highlights a highly noisy probe.

Probe noise model

Probe-level measurements of differential gene expression (treatment vs. control) are used in the current study. The observed multivariate expression signal s_p for probe p in a given probe set is modeled as a sum of

- real (probe set-level) expression profile g;
- variation in the control samples $N(\mathbf{0}, \sigma_c I)$;
- probe-related noise $N(\mathbf{0}, \sigma_p I)$.

Noise is assumed Gaussian with standard deviations σ_p and σ_c, respectively. This leads to the following model:

$$\mathbf{s}_p \sim \mathbf{g} + N(0, \sigma_p I) + N(0, \sigma_c I)$$

An estimate of probe-related noise (σ_p) can be used as a measure of probe performance, and is the focus of the current study. The model is implemented with Gibbs sampling and applied on two popular human genome arrays (Table 1).

Name	Conditions	Controls	Author
ALL-95	31	6	Yeoh et al. 2002
GEA-95	83	2	Su et al. 2002
ALL-133	31	6	Ross et al. 2003
GEA-133	156	2	Su et al. 2004

Table 1 Data sets used to analyze probes on HG-U95A/Av2 and HG-U133A arrays. The number of probes in a probeset is 16 and 11 on these arrays, respectively.

Results

Many probe sets are known to contain probes that bind to erroneous targets. The estimated noise σ_p is often high in such probes (Fig. 3).

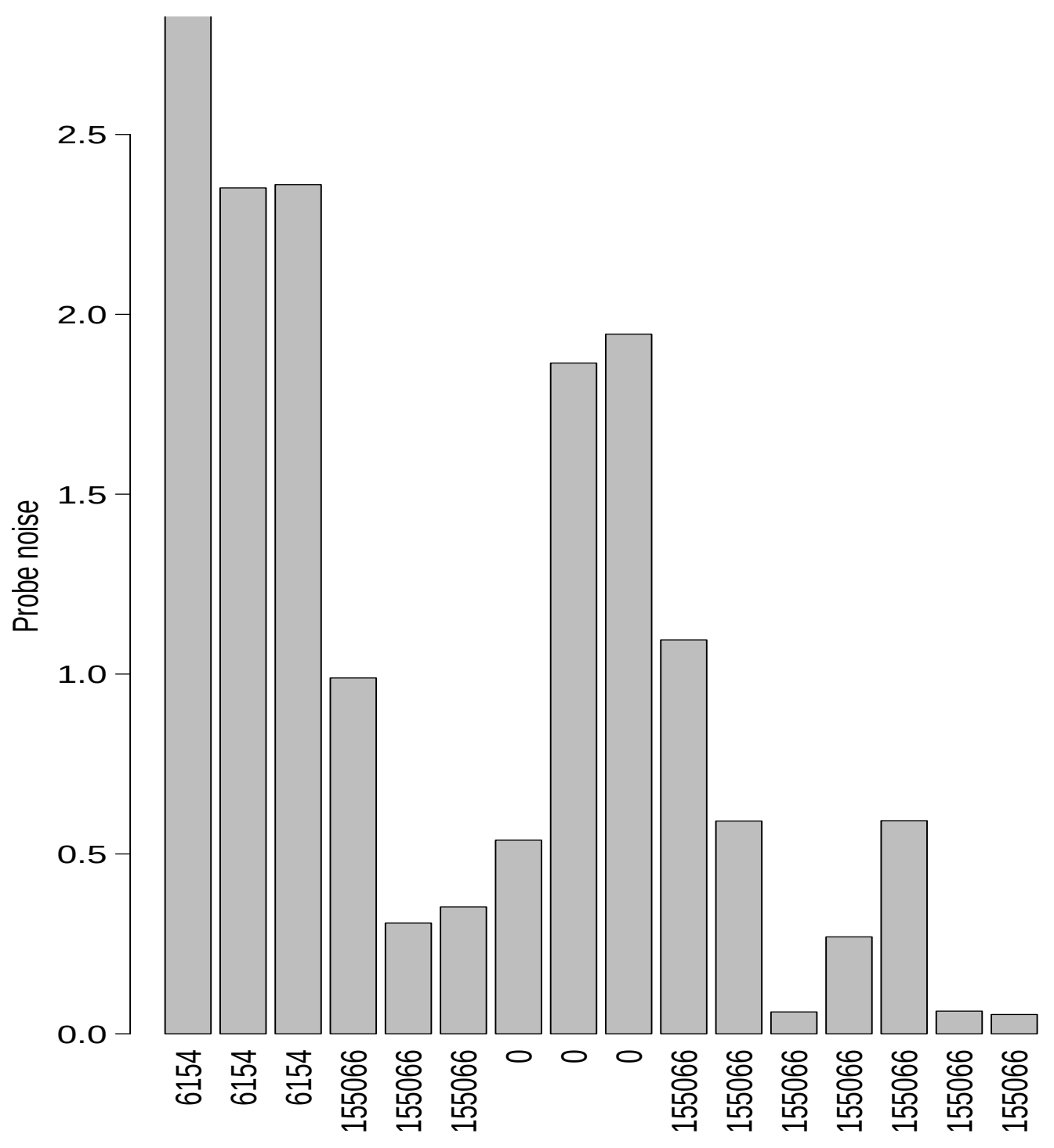

Figure 3 GeneID targets and estimated probe-related noise σ_p for the probes in the probe set '32444_at' (ALL-95). In this example, probes with an erroneous genomic match (6154), or no detected match (0) exhibit a higher level of probe-related noise than the correctly matched (155066) probes in this probe set.

More generally, many noisy probes detected by the model are explained by known errors in genomic alignment (Fig. 4).

The preliminary results also confirm that the binding position on the target sequence, and the GC-content of a probe are related to probe-level noise. In addition to identifying known erroneous probes, the model reveals consistently unreliable probes with no explanation known so far.

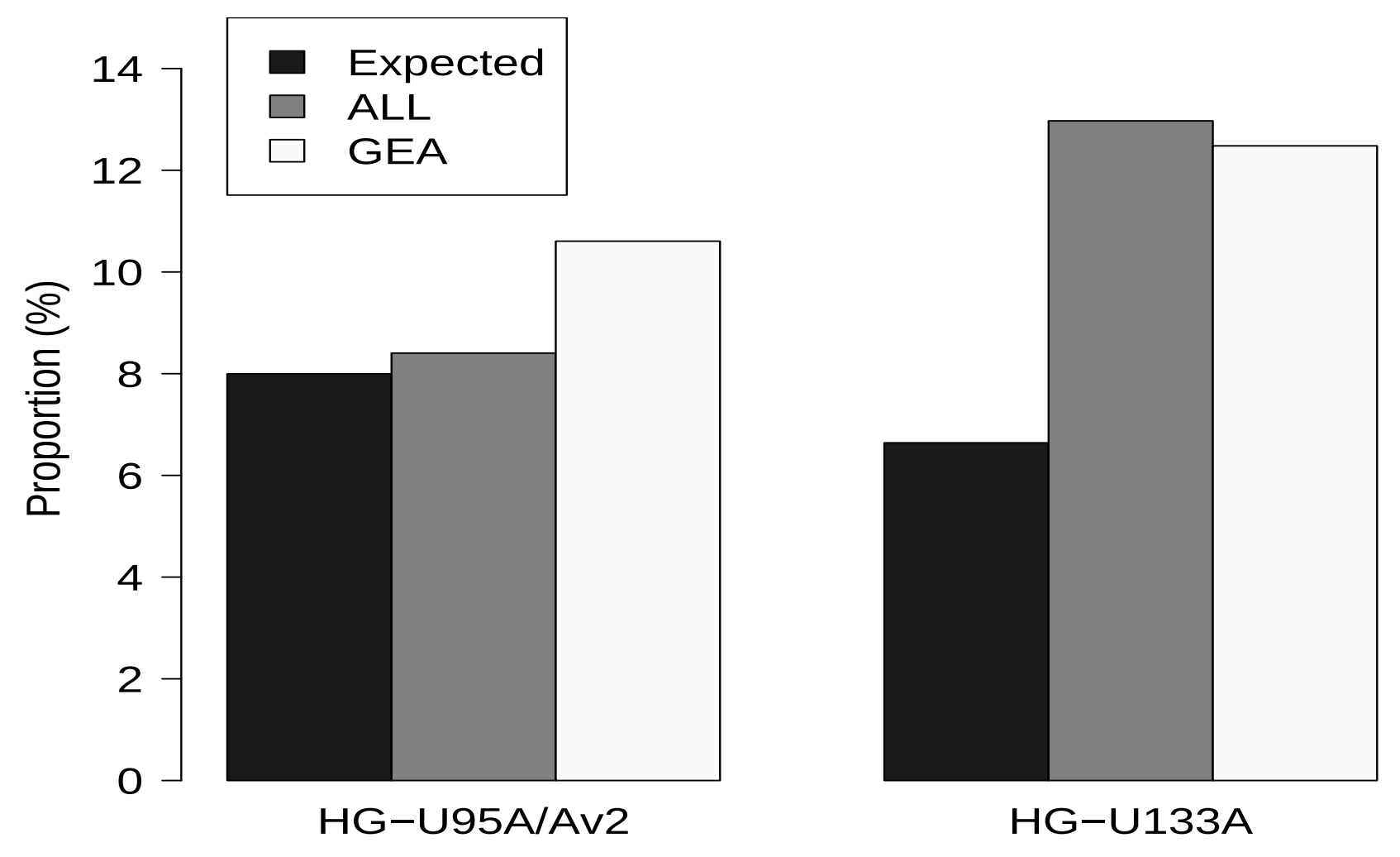

Figure 4 Proportion of probes with known errors in genomic alignment among the most noisy 1% of the probes on the two arrays in our model. Black bars show the proportion among all probes. Probes with known errors in genomic alignment explain part of the noisy probes in these data sets.

Conclusions

We have analyzed the performance of individual probes on short oligonucleotide arrays using a probabilistic model. While existing probe detection methods rely on external information such as genomic alignment of the probe sequences, the current model relies completely on expression data and can detect noisy probes independently of the error source. This helps to obtain better understanding of the factors that affect probe performance, and gives tools to guide microarray design and preprocessing.

References

1. Yeoh et al. Cancer Cell 1, 2002.
2. Su et al. PNAS 99, 2002.
3. Ross et al. Blood 102, 2003.
4. Su et al. PNAS 101, 2004.

TEKNILLINEN KORKEAKOULU
HELSINKI UNIVERSITY OF TECHNOLOGY

Contact: leo.lahti@tkk.fi
http://www.cis.hut.fi/projects/mi/