Consiglio Nazionale delle Ricerche
INFM

Unita' di Ricerca di Brescia
Via Valotti, 9 - 25133 Brescia
Tel.Direttore 030 3715771
Tel Segreteria 030 3715677
Fax 030 3715677

Dr. Matteo Pardo
Unità CNR-INFM Brescia
c/o Dipartimento di Chimica e Fisica
V. Valotti 9, 25133 Brescia
Tel 030-3715709
pardo@ing.unibs.it

Brescia, 8.9.2006

Scientific report of the visit

The 2 month visit to the Molecular Genetics Institute of the Max Planck Society served two distinct and related purposes:

1. Learn software and computational procedures, which are typical in bioinformatics (in particular in the analysis of DNA microarray data), and which can be usefully transferred to the sensor array arena.
2. Individuate a computational biology problem of interest and apply pattern recognition in order to solve it.

The two research lines brought to two conference papers, which we describe in turn below.

1. **Random Forests, Nearest Shrunken Centroids and Support Vector Machines for the Classification of Diverse E-Nose Datasets.** M. Pardo, G. Sberveglieri. 5th IEEE international Conference on Sensors, Daegu (Korea), October 2006 (accepted).

While feature plots (e.g. responses of single sensors over time) and descriptive statistics (e.g. calibration tables) may be sufficient for the analysis of small, low dimensional sensor data (e.g. for materials development), proper pattern recognition (PR) methods are needed to evaluate sensor *systems* (such as E-Noses) performance in practical tasks. The potential of state-of-the-art PR algorithms like Random Forests (RF), Nearest Shrunken Centroids (NSC) and SVM is exploited in diverse application areas, as in postgenomics (e.g. for DNA microarrays data analysis). RF is an ensamble of classification trees. It uses both bootstrap aggregation (bagging), a successful approach for combining unstable learners, and random variable selection for tree building. NSC classification makes one important modification to standard nearest centroid classification. The shrinkage consists of moving the centroid towards zero by threshold, setting it equal to zero if it hits zero. The shrinkage has two advantages: 1) it can make the classifier more accurate by reducing the effect of noisy features, 2) it does automatic feature selection. Finally, SVM are nowadays between the most used learning machines.

The computations we carried out rely on a number of R packages. R is a programming language used in statistics with a number of libraries for data analysis. The R package *MCRestimate* implements parameter optimization and error estimation by two nested cross–validation loops.
The E-Nose datasets we analyzed have been produced by the commercial EOS^{835} E-Nose, manufactured by the Italian company Sacmi s.c.a.r.l.. In the 1st experiment we determined the ripening levels of a roasted coffee blend inside the production chain of an Italian company; in the 2nd we investigated the detection of toxigenic strains of the deleterious fungus Fusarium verticillioides in corn; the 3rd dataset presents a commercially relevant problem, namely the distinction between extra virgin olive oil defects.
We applied each of the three types of classifier to each of the three datasets. We always used 5-fold external CV (error estimation) and 4-fold internal CV (parameter optimization). Folds were stratified and we further averaged on 50 repetitions of the whole procedure to get stable error determination. Parameter optimization was done over a grid of parameters, which depend on the classifier. As also recognized in the literature we found that the performance of RF does not depend much on the actual value of its parameters inside a large interval.
For space constraints we report the detailed results for the most difficult problem, the distinction between fusarium verticilloides producers and *non*-producers. In tables 1-3 we report the confusion matrices for the three classifiers. We see that SVM has a significant lower error, while RF and NSC perform similarly (though RF is better in estimating the fus. v. producers samples). The detail of the SVM correct classification frequency for each sample is given in Fig 1. In this way the hard samples can be singled out.

Table 1 Confusion table for Nearest Shrunken Centroids

	fus. v. non-producer	fus. v. producer	**Classification error**
fus. v. non-producer	31	24	**0.44**
fus. v. producer	20	38	**0.35**

Table 2 Confusion table for Random Forests

	fus. v. non-producer	fus. v. producer	**Classification error**
fus. v. non-producer	30	25	**0.46**
fus. v. producer	16	42	**0.28**

Table 3 Confusion table for SVM

	fus. v. non-producer	fus. v. producer	**Classification error**
fus. v. non-producer	37	18	**0.33**
fus. v. producer	12	46	**0.21**

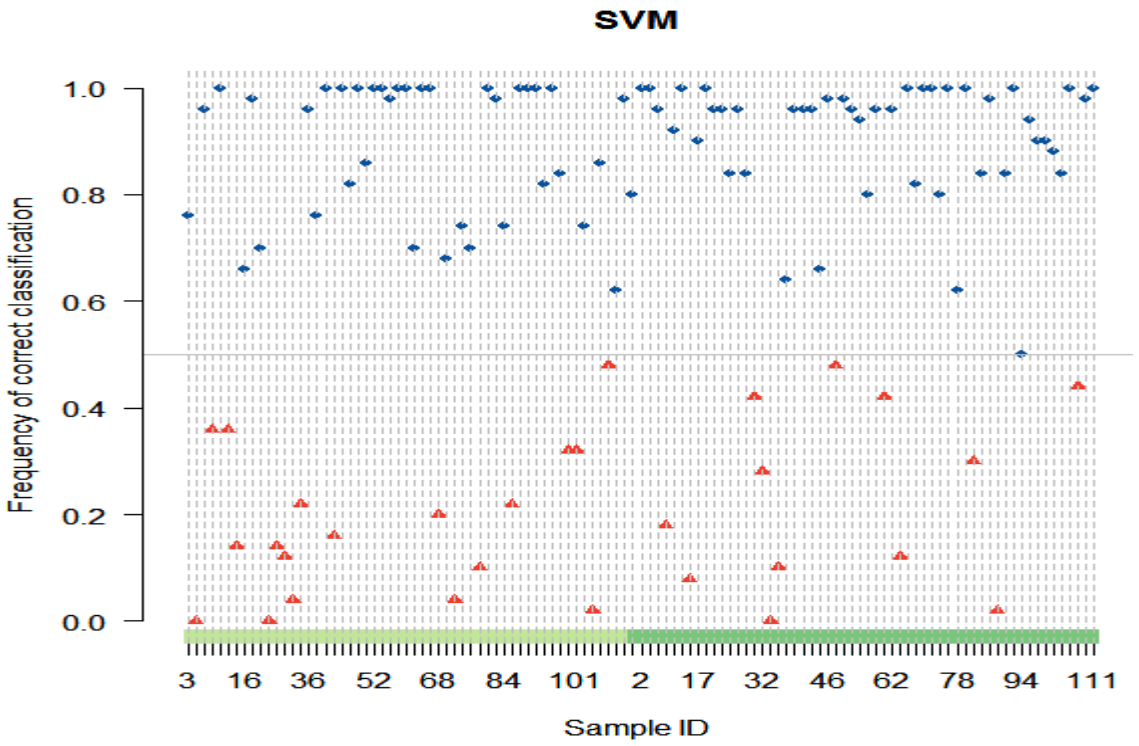

Figure 1 SVM correct classification frequency (over the 50 repetition) for each sample (triangles are mistakes). Left fus. v. non-producer, right fus. v. producer

2. **Translating a microarray signature into an equivalent but smaller signature applicable to paraffin material using rtPCR or immunophenotyping.** M. Pardo, S. Bentink, R. Spang. International Meeting on Biomarkers Selection. Genova, September 2006.

Burkitt's lymphoma (BL) and diffuse large-B-cell lymphoma (non-BL) are mature aggressive B-cell lymphomas. It is important to distinguish between the two B-cell lymphomas since they have different prognosis and must undergo different treatments. In particular one would like to single out the smaller class of BL from the broader non-BL class.
In [1] gene-expression profiling was performed using Affymetrix U133A GeneChips with RNA from 220 mature aggressive B-cell lymphomas, including a core group of 8 BLs that met all World Health Organization (WHO) criteria. An expression based disease entity was defined called the molecular Burkitt lymphoma (mBL) using a signature consisting of 74 probe sets pertaining to 57 unique genes.
In order to characterize clinical properties of this entity like survival or response to treatment one needs larger studies, which is only possible when including patients from the pre microarray era. For these patients no shock frozen tissue is available, hence a microarray analysis is not possible. However, there is paraffin material, which practically allows for expression measurement of 5-10 genes using rtPCR or immunohistochemistry.
The challenge is to find a smaller but equivalent mBL signature. In addition one has to take in account that immunohistochemistry might fail for a subset of signature genes. Hence gene selection requires some redundancy.
We therefore select eight genes from the 74 probe sets derived from [1] with the constraint that any four of them have a high discrimination power between BL and non-BL.

The procedure used to arrive to a gene list consists of subsequent steps:

1. Filter (variance) probesets to reduce them to N1<74.
2. Exhaustively rank N1 pick 4 sets (N1=35 takes 4h on a PC).
3. Pick N2 < N1 probesets.
4. Evaluate N2 pick 8 sets according to the performance of the 8 pick 4 = 70 4-feature sets.

To rank the N1 pick 4 sets we:

- Use CV error of a QDC with cost matrix

0.75 0
0 0.25

- Evaluate the performance is 1-cost

To pick N2 < N1 probesets we considered three strategies:

- N2 highest univariate top scoring filtered features
- Union of the top scoring 4-feature sets until N2
- Feature ranking from subset ranking and choose N2 highest features

Finally, the performance of each of the N2 pick 8 sets can be evaluated with three different measures, which summarize in different ways the performances of the 8 pick 4 4-feature sets:

- Worst 4-feature set
- 5 percentile 4-feature set
- Median 4-feature set

Results are summarized in table 4. The unique gene names relative to the probesets indices in table 4 are shown in table 5 (i.e. some probesets are relative to the same gene).
These results have been shown to physician involved in the German Lymphoma Alliance. By biological insight the gene sets in bold have been singles out for further analyses with rtPCR.
A global scoring of genes according to the recall frequency of any gene across the nine selection methods is shown in figure 2.

Table 4. Probesets numbers for the 9 different simulations conditions.

	Best 8 gene indices								**criterion**	**Perfomance: best = 0.9956; worst = 0.8803**	**4 sets index: indices run from 1 (best) to 52360 (worst)**
univariate ranking	1	12	18	30	57	60	63	69	worst	0.9432	23233
	12	18	30	42	58	60	67	69	median	0.9827	636
	1	12	18	30	57	60	63	69	perc5	0.9467	19946
ranking after union of best 4 feature sets	1	12	23	30	45	54	60	69	worst	**0.9452**	**21172**
	12	18	19	30	58	60	67	69	median	0.9827	636
	12	18	23	30	45	54	67	69	perc5	**0.9594**	**11630**
weighted ranking (pardo)	12	16	18	23	30	60	69	70	worst	0.9432	23233
	7	12	16	18	30	60	67	69	median	0.9827	636
	12	18	23	30	60	66	67	69	perc5	0.9467	19946

Table 5. Unique genes

'ARHGAP25'	'CFLAR'	'FHOD3'	'HNRPA3'	'LHFP'	'PRDM10'	'SSBP2'	[ ]
'ARHGAP25'	'CFLAR'	'DLEU1'	'HNRPA3'	'RCBTB1'	'SSBP2'	[ ]	[ ]
'ARHGAP25'	'CFLAR'	'FHOD3'	'HNRPA3'	'LHFP'	'PRDM10'	'SSBP2'	[ ]
'ARHGAP25'	'C7orf10'	'CFLAR'	'FNBP1'	'PRDM10'	'SSBP2'	'STAT3'	[ ]
'ARHGAP25'	'CD44'	'CFLAR'	'DLEU1'	'HNRPA3'	'RCBTB1'	'SSBP2'	[ ]
'ARHGAP25'	'C7orf10'	'DLEU1'	'FNBP1'	'HNRPA3'	'SSBP2'	'STAT3'	[ ]
'ARHGAP25'	'CD44'	'CFLAR'	'GARNL4'	'HNRPA3'	'SSBP2'	'STAT3'	[ ]
'ARHGAP25'	**'CD44'**	**'CFLAR'**	**'DLEU1'**	**'HNRPA3'**	**'SMARCA4'**	**'SSBP2'**	**[ ]**
'ARHGAP25'	'BCL3'	'CFLAR'	'DLEU1'	'HNRPA3'	'SSBP2'	'STAT3'	[ ]

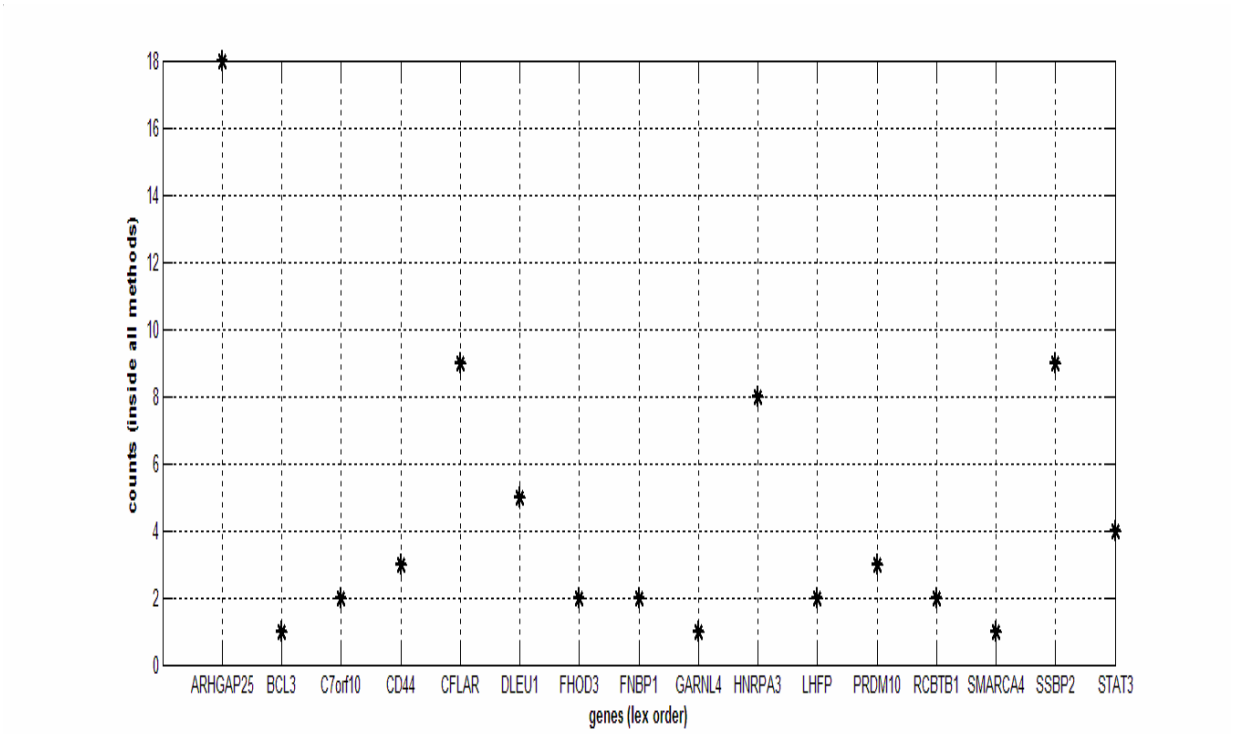

Figure 2. Frequency of gene recall inside Table 4.

[1] A Biologic Definition of Burkitt's Lymphoma from Transcriptional and Genomic Profiling. Hummel et al. N Engl J Med 2006;354:2419-30.