A Machine Learning Approach
to Profiling and Stratification
of Integrated -omics Data

Giuseppe Jurman, Samantha Riccadonna, Silvano Paoli, Stefano Merler, and Cesare Furlanello

FBK-irst, Trento, Italy

Brisbane, September 2007
High-throughput functional genomics: predictive data mining on wide data (p>>n, e.g. p=100K, n<100)

- Select stable ranked lists of markers in a supervised or semi-supervised process depending on target function
- Control Selection Bias in all steps, including preprocessing pipelines (e.g. Maldi-TOF spectra \rightarrow peaks \rightarrow features)

Toolset for Integrative –omics pattern recognition

- Combine markers from multiple genomics levels, integrate clinical information
FEATURES ➔ PATTERNS

Identify complex alteration patterns

Predict patient subtypes’ response as synthesis of integrated molecular profiles and of multifactorial patho-physiological data

Clinical features

Gene variation: SNPs
Copy number (CGH)
CpG Methylation

Expression
Exon array: Expr, Alt splicing

Mass Spectrometry
Maldi TOF / TOF TOF
Chromatography

Integrated molecular profile

Clinical features:
Tumor characteristics: histology, staging

Treatments, Environment (diet, drugs, habits)

Follow-up patient data

Other disease conditions, Family history

Experiment:
ALL THE RELEVANT FUNCTIONAL GENOMICS ANNOTATIONS
Two Issues: Bias and Stability

1. **Bias**: in data preparation, preprocessing (complex!), classification
   - **Microarrays**: since 2002 (Ambroise & McLachlan)
   - **Proteomics**: since 2004

2. **Instability** of biomarker lists:
   - “... very unstable signatures, with the slightest perturbation of the training set producing drastically different markers”
   - **Microarrays**: since 2005 (Michiels et al.)
     - Ioannidis 2005.
• externally, a stratified random partitioning
• internally, a model selection

The BioDCV platform

A complete validation setup

100K – 5ML models

Classification disease vs. control on microarray db

DATA SET

Feature values

CLASS labels

Model Selection

Biomarkers

Predictive Error

ATE

Run 1

Run 2

Run b

Run B

Mutual genelist distance
Canberra distance

200 300 400 500 600 700

Neural Networks 2003: DNA microarrays
BMC Bioinformatics 2003: E-RFE
Int. J. of Cancer 2005: Rhabdomyosarcoma
JAR, 2007: integration with clinical data

AIRC - IFOM BICG
Bioinformatics Platform
BioDCV’s Payoffs in 2007

**Characteristics:**
- Predictive error estimates and Feature Selection (RFE, E-RFE) up to 100K vars
- Feature selection in Python
- Machine Learning (SVM and data-driven kernels) in C
- Model metadata and results in SQLite database files
- Graphical output interface in R, GUI in TK/Python
- Linux and Windows implementation (GPL)

- HPC application: on clusters + grid implementation for the European EGEE BioMed VO

**Predictive error estimates and top-k lists for standard functional genomics tasks**

Samplewise analysis, outlier removal, subtyping, randomiz.

**Stability analysis of sets of ranked lists:** algebraic methods and software (ListsPy: 2007)
**STARTING POINT:**


**IDEA:**

- Consider as predictive classification task and use supervised machine learning in BioDCV to detect *patterns of different response* in genome copy number and gene expression profiles.

A study of association of expression profiling and CNA, endowed with relevant patient stratification data.

Indirect use of supervised classifiers to define subclasses of cases in which the genomic levels are *functionally different*.
DATA (from Chin et al 2006):

- **Task**: Predictive classification and feature ranking of Lum/ER+ vs Basal/ER- classes

- **Comparative genomics hybridization** (CGH) array data: 2149 BACs x 149 samples.
  - Used directly after kNN imputation* → **1590** BAC features
  - Encoded by Circular Binary Segmentation (includes imputation) with DNACopy package in R/BioConductor → **1674** feat

- **Expression**: Affy U133A, **22215** probes from original paper x 118 cases

- **Clinical data**: 136 fields for 174 patients

- **INTERSECTION FOR INTEGRATIVE GENOMIC ANALYSIS**: **65**
  - **42**: LumA or LumB with ER status > 0
  - **23**: Basal ER-
Steps 1-3: profile, shave, profile

- **Affy:** Predictive Error ATE < 1% with 3 genes
  - Trivially: 205225 (ESR1) best feature on 400 list sets.
- **BAC:** ATE < 12% for all models less than 300 features. But **two** samples are consistently misclassified (one is not univocally assigned as label) → removed from dataset
- **Also remove ESR1 probe, repeat profiling:**
  - Affy: ATE < 1% with 15 features
  - BAC: ATE < 10% with 15 features
    and min ATE = 6.9% with 50 BACs

Effective improvement in prediction and on list.

- Remove ESR1 probe, repeat profiling:
  - Affy: ATE < 1% ; 15+ features
  - BAC: ATE < 10% ; 15+ features
    min ATE = 6.9% ; 50+ BACs

Average Test Error on the copy number data (BACs) for the Lum/ER+ and Basal/ER- task, with 95% student’s bootstrap confidence intervals
Consider the OPTIMAL BORDA LISTS for BACs and for Affy ad align with: GenomeMAP, NetAffix, UCSC Genome Browser:

- **Given Affy list:** first 15 probes (<1% ATE) → 11 BACs, but only 2 usable due to imputation

- Given BAC list: first 15 BACs → 11 probes found for 5 BACS:
  - 201805_at is the best ranked for Affy (252)
  - all others are low or very low ranked!
Step 5: Sample-tracking

Copy-number data    Number of features
Sample-tracking: details

E(S): % error for BioDCV runs in which the sample is in test, averaged on models of the same feature size

- Error rate (ER+ Lum)
- Error rate (ER- Basal)
- No-information error rate

Estimated predictive error on all data

Copy-number data: Average error (for increasing number of features) as an aggregation of each sample-tracking profiles
**IN SUMMARY:** the class for Luminal/ER+ versus Basal/ER- for those 8 samples is
- often or mostly misclassified by BACs
- easily classified by gene expression

**CLINICAL FEATURES:**
- SBR grade is higher than median of all sample
- Recurrence time is shorter (3.3 vs 6.9y)

Gains and losses freq. for the 8 misclassified samples, compared to BACs for the remaining 55 samples: changes on 5, 6, 7, 10, 12, 14

On top-15 best gene expression signature, most probes separate within the subclass class better than for the other 55 samples
OTHER APPLICATIONS
supervised classification of

- Gene expressions and Copy-number alterations defined by 100K SNP Arrays on NCI60 cell lines (Garraway et al 2005)

- Gene expression and Oligonucleotide arrays Agilent Human 1A V2 on adenocarcinoma vs squamous-cell carcinoma (Tonon et al 2005).

CONCLUSIONS

1. Predictive error such 12.5% with 10 BACs supports the use of aCGH data in classification tasks

2. Absence of co-occurrence between genes located in top-ranked BACs and the location of the top ranked Affy probesets →indipendente sources for tumour classification

3. Patient stratification emerging from CNA profile
More subtypes