Machine learning for patient stratification from genomic information

Jean-Philippe Vert
jean-philippe.vert@ens.fr

Paris Sciences and Data, December 1, 2016
Big data for health
Molecular data
Opportunity

Personalized Cancer Therapy

1. Molecular Profiling
2. Prognostic Markers
   - Markers predictive of drug sensitivity/resistance
   - Markers predictive of adverse events

https://pct.mdanderson.org
Learning from data: supervised classification/regression

- Patients with VS without relapse in 5 years
- Case where $n (=19)$ patients $>> p (=2)$ markers
Learning from data: supervised classification/regression

- Patients with VS without relapse in 5 years
- Case where $n (=19)$ patients $>> p (=2)$ markers
Learning from data: supervised classification/regression

- Patients with VS without relapse in 5 years
- Case where $n (=19)$ patients $>> p (=2)$ markers
Patients with VS without relapse in 5 years
Case where \( n (=19) \) patients \( >> p (=2) \) markers
Real data: \( n << p \)

- **Gene expression**

- **Somatic mutations**

- \( n = 10^2 \sim 10^4 \) (patients)
- \( p = 10^4 \sim 10^7 \) (genes, mutations, copy number, ...)
- Data of various nature (continuous, discrete, structured, ...)
- Data of variable quality (technical/batch variations, noise, ...)
Consequence: limited accuracy


- C: 16 standard clinical data (age, tumor size, ...)
- M: 80k molecular features (gene expression, DNA copy number)
- P: incorporate prior knowledge
Example: survival prediction from somatic mutations

Figure 2 – Comparison of the survival predictive power of the raw mutation data, NSQN and NetNorM (with Pathway Commons as gene network) for 8 cancer types. For each cancer type, samples were split 20 times in training and test sets (4 times 5-fold cross-validation). Each time a sparse survival SVM was trained on the training set and the test set was used for performance evaluation. The presence of asterisks indicate when the test CI is significantly different between 2 conditions (Wilcoxon signed-rank test, $P < 5 \times 10^{-2}$ or $P < 1 \times 10^{-2}$).

Data from TCGA (3.3k samples, 8 cancer types, >10k genes)

Survival SVM on raw binary data, or processed by NSQN (Hofree et al., 2013) or NetNorm (Le Morvan et al., 2016).
Consequence: unstable biomarker selection

70 genes (Nature, 2002)  
76 genes (Lancet, 2005)

3 genes in common

van ’t Veer et al. (2002); Wang et al. (2005)
Some research directions

- Find a better representation
- Incorporate prior knowledge

---

**Figure 2.1:** Taxonomic constitution of the benchmark dataset. The tree shown on the upper part of the figure shows the taxonomic organisation of the bacterial panel considered in our benchmark. The leaves of the taxonomy correspond to the 20 species and their parent to the 9 genera. Internal nodes correspond to either phenotypic (e.g., Gram positive and negative at the top of the taxonomy) or evolutive attributes. Nodes shown in grey are those that can be pruned for computational efficiency, because they have a single child. Circles shown on the bottom represent the number of strains (dark grey) and spectra (light grey) available for each species.

We note finally that we have considered in this study a peak-list representation in which a mass spectrum is represented by a vector $x \in \mathbb{R}^p$, where $p$ is the numbers of bins considered to discretize the mass to charge range, and each entry of $x$ is derived from the intensity of the peak(s) found in corresponding bin. Figure 2.2 represents a clustered version of the MicroMass dataset, where the rows correspond to 571 mass-spectra ordered according to their genus label and the columns are the 1300 intensity peaks grouped by an unsupervised clustering step. Interestingly, we remark block structures suggesting that some features uniquely belong to one genus class.

While several schemes have been proposed to define such a peak-list representation (see [49] for instance), we have relied in this study on the approach embedded in the VITEK-MS system, which provides a peak-list representation of dimension $p=1300$, 1.

---

From prognostic to predictive models

**Predictive factors**
Determine *which* treatment is best

**Prognostic factors**
Determine *who* needs treatment

Therapeutic choices

1. Avoid under and over treatment
2. Personalised treatment

**Prognostic:**
- Predict outcome $Y$ of a disease on an untreated individual $X$
- Standard supervised learning: model $Y = f(X)$ from observations of $(X_i, Y_i)$ pairs

**Predictive:**
- Predict the benefit in outcome $Y$ of a treatment $A$ on an individual $X$
- We observe $(X_i, A_i, Y_i)$ but want to model $Y = f(X, A_1) - f(X, A_2)$
- For each $X$ we only observe the outcome $Y$ under one treatment $A$ (cf e-marketing)
Clinical trials for precision medicine?

1. Meta-analysis of clinical trials (typically *A/B testing*) to estimate predictive models

2. Dynamic trial to jointly optimize the predictive model and its performance (*contextual multi-armed bandit problem*)
Conclusion

- Lots of data
- $n << p$ is the rule, more and more...
- Limited impact so far for patients
- Active research
  - new representations $x \rightarrow \Phi(x)$
  - new learning techniques (structured sparsity, regularization, ...)
  - new experimental design strategies (contextual bandit)
