Machine learning for patient stratification from genomic data

Jean-Philippe Vert

Ghent University, March 7, 2017
Goal

Personalized Cancer Therapy

Molecular Profiling

1

Prognostic Markers

Markers predictive of drug sensitivity/resistance

Markers predictive of adverse events

2

https://pct.mdanderson.org
Mathematical model

- Patients with VS without relapse in 5 years
- \( n (=19) \) patients \( \gg p (=2) \) markers
Patients with VS without relapse in 5 years

$n (=19)$ patients $>> p (=2)$ markers
Mathematical model

- Patients with VS without relapse in 5 years
- $n (=19)$ patients $>> p (=2)$ markers
Mathematical model

- Patients with VS without relapse in 5 years
- $n (=19)$ 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

Breast cancer prognosis competition, $n = 2000$ (Bilal et al., 2013)

- C: 16 standard clinical data (age, tumor size, ...)
- M: 80k molecular features (gene expression, DNA copy number)
Consequence: unstable biomarker selection

Gene expression profiling predicts clinical outcome of breast cancer

Laura J. van ’t Veer†, Hongyue Dai‡, Marc J. van de Vijver†, Yadong He†, Augustinus A. M. Hart*, Mao Mao†, Hans L. Peterse*, Karin van der Kooy‡, Matthew J. Marton‡, Anke T. Witteveen‡, George J. Schreiber‡, Ron M. Kerkhoven‡, Chris Roberts‡, Peter S. Linsley‡, René Bernards* & Stephen H. Friend‡

* Divisions of Diagnostic Oncology, Radiotherapy and Molecular Carcinogenesis and Center for Biomedical Genetics, The Netherlands Cancer Institute, 121 Plesmanlaan, 1066 CX Amsterdam, The Netherlands
‡ Rosetta Inpharmatics, 17040 115th Avenue NE, Kirkland, Washington 98034.


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 organization 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. For intellectual property issues, it had to be anonymized: the name of the species are not given, and most of the nodes of the taxonomy are hidden. We note however that the considered taxonomy is polyphasic and involves both phenotypic and evolutionary traits. For instance, the uppermost level of the taxonomy separates species into Gram positive and Gram negative, and the two lowest levels provide the species and genus information. Such a hybrid taxonomic definition is common in the context of clinical microbiology, where manual identification involves a succession of tests meant to establish several phenotypic and metabolic properties of the micro-organism to identify (e.g., Gram +/- or aerobe/anaerobe). These properties correspond to the upper levels of the taxonomy, while the lower ones correspond to standard phylogenetic levels (e.g., family, genus and 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 $\mathbf{x} \in \mathbb{R}^p$, where $p$ is the number of bins considered to discretize the mass to charge range, and each entry of $\mathbf{x}$ 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 $p = 1300$. 

Outline

1. Learning from mutation data
2. Supervised quantile normalization
3. The Kendall and Mallows kernels
4. Conclusion
1. Learning from mutation data
2. Supervised quantile normalization
3. The Kendall and Mallows kernels
4. Conclusion
Somatic mutations in cancer

Stratton et al. (2009)
Large-scale efforts to collect somatic mutations

- **3,378 samples** with survival information from 8 cancer types
- Downloaded from the TCGA / cBioPortal portals.

<table>
<thead>
<tr>
<th>Cancer type</th>
<th>Patients</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUAD (Lung adenocarcinoma)</td>
<td>430</td>
<td>20,596</td>
</tr>
<tr>
<td>SKCM (Skin cutaneous melanoma)</td>
<td>307</td>
<td>17,463</td>
</tr>
<tr>
<td>GBM (Glioblastoma multiforme)</td>
<td>265</td>
<td>14,750</td>
</tr>
<tr>
<td>BRCA (Breast invasive carcinoma)</td>
<td>945</td>
<td>16,806</td>
</tr>
<tr>
<td>KIRC (Kidney renal clear cell carcinoma)</td>
<td>411</td>
<td>10,609</td>
</tr>
<tr>
<td>HNSC (Head and Neck squamous cell carcinoma)</td>
<td>388</td>
<td>17,022</td>
</tr>
<tr>
<td>LUSC (Lung squamous cell carcinoma)</td>
<td>169</td>
<td>13,590</td>
</tr>
<tr>
<td>OV (Ovarian serous cystadenocarcinoma)</td>
<td>363</td>
<td>10,195</td>
</tr>
</tbody>
</table>
Survival prediction from raw mutation profiles

- Each patient is a **binary vector**: each gene is mutated (1) or not (2)
- Silent mutations are removed
- Survival model estimated with sparse survival SVM
- Results on 5-fold cross-validation repeated 4 times
Can we replace

\[ x \in \{0, 1\}^p \text{ with } p \text{ very large, very sparse} \]

by a representation with more information shared between samples

\[ \Phi(x) \in \mathcal{H} \text{ ?} \]
NetNorm Overview (Le Morvan et al., 2016)

- **Modify** the binary vector \( x \in \{0, 1\}^p \) of each patient by adding or removing mutations, using a gene network as prior knowledge.
- **After Netnorm**, all patients \( \Phi(x) \in \{0, 1\}^p \) have the same number of (pseudo-)mutations.

**Results**

**2.1 Overview of NetNorM**

- **Number of mutated neighbours**
- **Degree of mutated genes**
- **Patient with less than \( k \) mutations**
- **Patient with more than \( k \) mutations**

**Application of NetNorM with \( k = 4 \)**

**Figure 1**

- **Overview of NetNorM.** (a) Using a gene network as background knowledge (lower left), NetNorM normalises each mutation profile in a collection of somatic mutation profiles (upper left) into a new, binary representation (right) which encodes additional information relative to patient mutation rates and hubs' neighbourhood mutational burden. This new representation allows performing patient stratification with unsupervised clustering techniques, or survival analysis. (b) NetNorM normalises every patient mutation profile to \( k \) mutations. Patients with less than \( k \) mutations get 'proxy' mutations in their genes with the highest number of mutated neighbours until they reach \( k \) mutations. Patients with more than \( k \) mutations have mutations 'removed' in their genes with lowest degree until they reach \( k \) mutations.
1. **Add** mutations for patients with **few** (less than $k$) mutations

2. **Remove** mutations for patients with **many** (more than $k$) mutations
Related work (Hofree et al., 2013)

Network-based stratification of tumor mutations

Matan Hofree¹, John P Shen², Hannah Carter², Andrew Gross³ & Trey Ideker¹–³

¹Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California, USA. ²Department of Medicine, University of California, San Diego, La Jolla, California, USA. ³Department of Bioengineering, University of California, San Diego, La Jolla, California, USA. Correspondence should be addressed to T.I. (tideker@ucsd.edu).

RECEIVED 14 FEBRUARY; ACCEPTED 12 AUGUST; PUBLISHED ONLINE 15 SEPTEMBER 2013; DOI:10.1038/NMETH.2651

Network smoothing:
- Gene
- Gene-gene interaction

Network-based stratification:
- Patients
- Genes
- Cluster prototypes
- Network influence constraint

Fig. 1 summarizes the number of genes and works. Our implementation of NBS is available as

1108 | VOL.10 NO.11 | NOVEMBER 2013 | NATURE METHODS
Use Pathway Commons as gene network.
NSQN = Network Smoothing / Quantile Normalization (Hofree et al., 2013)
NetNorM and NSQN benefit from biological information in the gene network

Comparison with 10 randomly permuted networks:
Selected genes represent "true" or "proxy" mutations

**Genes selected in at least 50% of the cross-validated sparse SVM model**
Proxy mutations encode local mutational burden

Figure 4 – Analysis of predictive genes. (a) Comparison of survival prediction performances according to patients' mutation rate for LUAD. Three different representations of the mutations are used to perform survival prediction using a ranking SVM: raw (the raw binary mutation data), NSQN (network smoothing with quantile normalisation) and NetNorM. NSQN and NetNorM are applied with Pathway Commons as gene-gene interaction network. Performances for half of the patients with fewer (resp. more) mutations are derived from the predictions made using the whole dataset. (b) Scatter plot of the correlation between the total number of mutations across patients and the number of mutated neighbours of a gene across patients (x-axis) against the degree of a gene (y-axis). This plot was generated using the raw mutation data for LUAD and Pathway Commons. (c) Scatter plot of the total number of mutations in a patient (x-axis) against the number of mutated neighbours of KHDRBS1 in a patient (y-axis). Only patients with less that $k_{med} = 295$ mutations are shown, where $k_{med}$ is the median value of $k$ learned across cross-validation folds. Red (resp. blue) indicate patients mutated (resp. non mutated) in KHDRBS1 after processing with NetNorM using $k = k_{med}$. The black line was fit by linear regression and by definition indicates the expected number of mutated neighbours of KHDRBS1 given the mutation rate of a patient. The plot was generated using the LUAD dataset with Pathway Commons.

KHDRBS1: a member of the K homology domain-containing, RNA-binding, signal transduction-associated protein family
Adding good old clinical factors

Combination by averaging predictions
Joint work with

Marine Le Morvan
Standard full quantile normalization

Typically followed by a predictive model on the normalized data
Choosing a "good" target distributions is important.

Cancer prognosis from somatic mutations.
How to choose a "good" target distribution?

- Gaussian distribution (mean=0, sd=1)
- Uniform distribution
- Bigaussian distribution

Quantile functions:
- Gaussian to Gaussian
- Uniform to Uniform
- Bigaussian to Bigaussian
Notations

- $x_1, \ldots, x_n \in \mathbb{R}^p$ a set of $p$-dimensional samples

- $f \in \mathbb{R}^p$ a non-decreasing target distribution (CDF)

- For $x \in \mathbb{R}^p$, let $\Phi_f(x) \in \mathbb{R}^p$ be the data after QN with target distribution $f$
From QN to supervised QN (SUQUAN)

Standard approaches: learn model after QN preprocessing:

1. Fix $f$ arbitrarily
2. QN all samples to get $\Phi_f(x_1), \ldots, \Phi_f(x_n)$
3. Learn a generalized linear model $(w, b)$ on normalized data:

$$
\min_{w, b} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell_i \left( w^\top \Phi_f(x_i) + b \right) + \lambda \Omega(w) \right\}
$$

SUQUAN: jointly learn $f$ and $(w, b)$:

$$
\min_{w, b, f} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell_i \left( w^\top \Phi_f(x_i) + b \right) + \lambda \Omega(w) + \gamma \Omega_2(f) \right\}
$$
SUQAN as matrix regression

For $x \in \mathbb{R}^p$, let $\Pi_x \in \mathbb{R}^{p \times p}$ the permutation matrix of $x$’s entries

$$x = \begin{pmatrix} 4.5 \\ 1.2 \\ 10.1 \\ 8.9 \end{pmatrix}, \quad \Pi_x = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad f = \begin{pmatrix} 0 \\ 1 \\ 3 \\ 4 \end{pmatrix}$$

Quantile normalized $x$ with target distribution $f$ is:

$$\Phi_f(x) = \Pi_x f$$

SUQUAN solves

$$\min_{w,b,f} \frac{1}{n} \sum_{i=1}^{n} \ell \left( w^\top \Pi_x f + b \right) + \lambda \Omega(w) + \gamma \Omega_2(f)$$

$$= \min_{w,b,f} \frac{1}{n} \sum_{i=1}^{n} \ell \left( <wf^\top, \Pi_x> + b \right) + \lambda \Omega(w) + \gamma \Omega_2(f)$$  \hspace{1cm} (1)

A particular rank-1 matrix optimization, $x$ is replaced by $\Pi_x$

Solved by alternatively optimizing $f$ and $w$
Experiments

\[
\min_{w, b, f} \frac{1}{n} \sum_{i=1}^{n} \ell_i \left( w^\top \Phi_f(x_i) + b \right) + \frac{\lambda}{2} ||w||_2^2 + \frac{\gamma}{2} \sum_{j=1}^{p-1} (f_{j+1} - f_j)^2
\]

- Breast cancer prognosis from gene expression data.
- Two classes of patients: those who relapsed within 6 years of diagnosis and those who did not.

<table>
<thead>
<tr>
<th>Dataset name</th>
<th># genes</th>
<th># patients</th>
<th># positives</th>
<th>% positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE7390</td>
<td>22283</td>
<td>189</td>
<td>58</td>
<td>0.31</td>
</tr>
<tr>
<td>GSE4922</td>
<td>22283</td>
<td>225</td>
<td>73</td>
<td>0.32</td>
</tr>
<tr>
<td>GSE2990</td>
<td>22283</td>
<td>106</td>
<td>32</td>
<td>0.30</td>
</tr>
<tr>
<td>GSE2034</td>
<td>22283</td>
<td>271</td>
<td>104</td>
<td>0.38</td>
</tr>
<tr>
<td>GSE1456</td>
<td>22283</td>
<td>141</td>
<td>37</td>
<td>0.26</td>
</tr>
</tbody>
</table>
Performance

average over all datasets

Number of genes
AUC
SUQ_logistic_med1
SUQ_logistic_gau1
SUQ_logistic_lin1
SUQ_logistic_bigau1
SUQ_logistic_med2_spav
glm_logistic_raw
glm_logistic_rma
Estimated distribution: iteration=0
Estimated distribution: iteration=1
Estimated distribution: iteration=2
Outline

1. Learning from mutation data
2. Supervised quantile normalization
3. The Kendall and Mallows kernels
4. Conclusion
Joint work with

Yunlong Jiao
An idea: all pairwise comparisons

Replace \( x \in \mathbb{R}^p \) by \( \Phi(x) \in \{0, 1\}^{p(p-1)/2} \):

\[
\Phi_{i,j}(x) = \begin{cases} 
1 & \text{if } x_i \leq x_j , \\
0 & \text{otherwise.} 
\end{cases}
\]

One sample \( x \) 
\( p \) features

Mapping \( f(x) \) 
\( p(p-1)/2 \) bits
Related work: Top scoring pairs (TSP)

(Geman et al., 2004; Tan et al., 2005; Leek, 2009)
Practical challenge

- Need to store $O(p^2)$ bits per sample
- Need to train a model in $O(p^2)$ dimensions
Kernel trick

Theorem (Wahba, Schölkopf, ...)

Training a linear model over a representation $\Phi(x) \in \mathbb{R}^Q$ of the form:

$$\min_{w \in \mathbb{R}^Q} \frac{1}{n} \sum_{i=1}^{n} \ell(w^\top \Phi(x_i), y_i) + \lambda \|w\|^2$$

can be done efficiently, independently of $Q$, if the kernel

$$K(x, x') = \Phi(x)^\top \Phi(x')$$

can be computed efficiently.

Ex: ridge regression, $O(Q^3 + nQ^2)$ becomes $O(n^3 + n^2 T)$
Other: SVM, logistic regression, Cox model, survival SVM, ...
Kernel trick for us: Kendall’s $\tau$

$$\Phi(x)\Phi(x') = \tau(x, x') \quad \text{(up to a scaling)}$$

Good news for SVM and kernel methods!
More formally

- For two permutations $\sigma, \sigma'$ let $n_c(\sigma, \sigma')$ (resp. $n_d(\sigma, \sigma')$) the number of concordant (resp. discordant) pairs.
- The Kendall kernel (a.k.a. Kendall tau coefficient) is defined as
  \[ K_\tau(\sigma, \sigma') = \frac{n_c(\sigma, \sigma') - n_d(\sigma, \sigma')}{\binom{p}{2}}. \]
- The Mallows kernel is defined for any $\lambda \geq 0$ by
  \[ K_\lambda^M(\sigma, \sigma') = e^{-\lambda n_d(\sigma, \sigma')} . \]

**Theorem (Jiao and V., 2015)**

*The Kendall and Mallows kernels are positive definite.*

**Theorem (Knight, 1966)**

*These two kernels for permutations can be evaluated in $O(p \log p)$ time.*
Kondor and Barbarosa (2010) proposed the diffusion kernel on the Cayley graph of the symmetric group generated by adjacent transpositions.

Computationally intensive ($O(p^p)$)

Mallows kernel is written as

$$K_M^\lambda(\sigma, \sigma') = e^{-\lambda n_d(\sigma, \sigma')} ,$$

where $n_d(\sigma, \sigma')$ is the shortest path distance on the Cayley graph.

It can be computed in $O(p \log p)$
Application: supervised classification

Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>No. of features</th>
<th>No. of samples (training/test)</th>
<th>$C_1$</th>
<th>$C_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer 1</td>
<td>23624</td>
<td>44/7 (Non-relapse)</td>
<td></td>
<td>32/12 (Relapse)</td>
</tr>
<tr>
<td>Breast Cancer 2</td>
<td>22283</td>
<td>142 (Non-relapse)</td>
<td></td>
<td>56 (Relapse)</td>
</tr>
<tr>
<td>Breast Cancer 3</td>
<td>22283</td>
<td>71 (Poor Prognosis)</td>
<td></td>
<td>138 (Good Prognosis)</td>
</tr>
<tr>
<td>Colon Tumor</td>
<td>2000</td>
<td>40 (Tumor)</td>
<td></td>
<td>22 (Normal)</td>
</tr>
<tr>
<td>Lung Cancer 1</td>
<td>7129</td>
<td>24 (Poor Prognosis)</td>
<td></td>
<td>62 (Good Prognosis)</td>
</tr>
<tr>
<td>Lung Cancer 2</td>
<td>12533</td>
<td>16/134 (ADCA)</td>
<td></td>
<td>16/15 (MPM)</td>
</tr>
<tr>
<td>Medulloblastoma</td>
<td>7129</td>
<td>39 (Failure)</td>
<td></td>
<td>21 (Survivor)</td>
</tr>
<tr>
<td>Ovarian Cancer</td>
<td>15154</td>
<td>162 (Cancer)</td>
<td></td>
<td>91 (Normal)</td>
</tr>
<tr>
<td>Prostate Cancer 1</td>
<td>12600</td>
<td>50/9 (Normal)</td>
<td></td>
<td>52/25 (Tumor)</td>
</tr>
<tr>
<td>Prostate Cancer 2</td>
<td>12600</td>
<td>13 (Non-relapse)</td>
<td></td>
<td>8 (Relapse)</td>
</tr>
</tbody>
</table>

Methods

- Kernel machines Support Vector Machines (SVM) and Kernel Fisher Discriminant (KFD) with Kendall kernel, linear kernel, Gaussian RBF kernel, polynomial kernel.
- Top Scoring Pairs (TSP) classifiers Tan et al. (2005).
- Hybrid scheme of SVM + TSP feature selection algorithm.
Results

Kendall kernel SVM

- **Competitive accuracy!**
- Less sensitive to regularization parameter!
- No need for feature selection!
Kendall kernel SVM

- Competitive accuracy!
- Less sensitive to regularization parameter!
- No need for feature selection!
Kendall kernel SVM
- Competitive accuracy!
- Less sensitive to regularization parameter!
- No need for feature selection!
Application: clustering

- APA data (full rankings)
- $n = 5738$, $p = 5$
- (new) Kernel k-means vs (standard) k-means in $S_5$
- Show silhouette as a function of number of clusters (higher better)
Two interesting types of partial rankings are interleaving partial ranking

\[ x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_k}, \quad k \leq n. \]

and top-\(k\) partial ranking

\[ x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_k} \succ X_{\text{rest}}, \quad k \leq n. \]

Partial rankings can be uniquely represented by a set of permutations compatible with all the observed partial orders.

Theorem

For these two particular types of partial rankings, the convolution kernel (Haussler, 1999) induced by Kendall kernel

\[
K^*(R, R') = \frac{1}{|R||R'|} \sum_{\sigma \in R} \sum_{\sigma' \in R'} K_\tau(\sigma, \sigma')
\]

can be evaluated in \(O(k \log k)\) time.
Extension to partial rankings

- Two interesting types of partial rankings are **interleaving partial ranking**
  \[ x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_k}, \quad k \leq n. \]
  and **top-\(k\) partial ranking**
  \[ x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_k} \succ X_{\text{rest}}, \quad k \leq n. \]

- Partial rankings can be **uniquely represented** by a set of permutations compatible with all the observed partial orders.

**Theorem**

*For these two particular types of partial rankings, the convolution kernel (Haussler, 1999) induced by Kendall kernel*

\[
K_T^*(R, R') = \frac{1}{|R||R'|} \sum_{\sigma \in R} \sum_{\sigma' \in R'} K_T(\sigma, \sigma')
\]

*can be evaluated in \(O(k \log k)\) time.*
Extension to smoother, continuous representations

Instead of \( \Phi : \mathbb{R}^p \rightarrow \{0, 1\}^{p(p-1)/2} \), consider the continuous mapping \( \Psi_a : \mathbb{R}^p \rightarrow \mathbb{R}^{p(p-1)/2} \):

\[
\Psi_a(x) = \mathbb{E}\Phi(x + \epsilon) \quad \text{with} \quad \epsilon \sim (\mathcal{U}[-\frac{a}{2}, \frac{a}{2}])^n
\]

Corresponding kernel \( G_a(x, x') = \Psi_a(x)^\top \Psi_a(x') \)
Computation of $G(x, x')$

- $G_a(x, x')$ can be computed exactly in $O(p^2)$ by explicit computation of $\Psi_a(x)$ in $\mathbb{R}^{p(p-1)/2}$

- $G_a(x, x')$ can be computed approximately in $O(D^2 p \log p)$ by Monte-Carlo approximation:

$$\tilde{G}_a(x, x') = \frac{1}{D^2} \sum_{i,j=1}^{D} K(x + \epsilon_i, x' + \epsilon_j)$$

- Theorem: for supervised learning, Monte-Carlo approximation is better\(^1\) than exact computation when $n = o(p^{1/3})$

\(^1\)faster for the same accuracy
Performance of $G_a(x, x)$

The figure shows the performance of $G_a(x, x)$ with varying noise window size $a$.

- **SVMkdtALLalt−−exact**
- **SVMkdtALLalt−−MCapprox (D=1)**
- **SVMkdtALLalt−−MCapprox (D=3)**
- **SVMkdtALLalt−−MCapprox (D=5)**
- **SVMkdtALLalt−−MCapprox (D=7)**
- **SVMkdtALLalt−−MCapprox (D=9)**
- **SVMkdtALL**

The plot compares the performance metrics across different noise window sizes, with a focus on the SVMkdtALL algorithm and its approximate versions with different parameters.
Outline

1. Learning from mutation data
2. Supervised quantile normalization
3. The Kendall and Mallows kernels
4. Conclusion
Many learning problem in precision medicine are **hard**, machine learning is no magic bullet with \( n \ll p \) and complex data.

Understanding the **benefits and cost** of different representations remains very heuristic and sometimes counterintuitive.

**NetNorm** is one way to use prior knowledge; why it "works" is not fully understood.

Representing omics data as **permutations** has some potential; the information lost about the gene expression values seems irrelevant (SUQUAN, Kendall and Mallow’s kernels).

**Learning representation** is worth investigating.
POINT & ÉTAPPE / 20 FÉVRIER – 1ER JUILLET 2014

C. SURIAM – F. LEQUEUX

1/ PROMOTION DU PROGRAMME

• Rencontre avec les directeurs, directeurs scientifiques et directeurs des études de l’ENS, Mines, ENSCP, ESPCI pour organiser la promotion du programme auprès des étudiants.

• Ecole des Mines : présentation d’ITI devant le conseil de l’enseignement ; 8 présentatrices d’élèves ; relance

• ENSCP : mailing de présentation d’ITI aux 3A via la direction des études.

• ENS : mailing ; présentation directe auprès des étudiants (2 élèves présents) ; diffusion des plaquettes et du syllabus

• ESPCI : mailing aux 3A + présentation d’ITI / rencontre avec les étudiants en présence de la directrice de la scolarité et de la directrice des relations entreprises ; 8 élèves présents

• Contacts en cours avec les Ponts et l’ENS Lyon (en attente de réponse)

• Rencontre à l’ANRT avec le délégué général et la chef du service CIFRE : accord de communication sur le programme ITI via le site internet et de l’ANRT (rubrique “zoom sur”)

• Promotion aux Rencontres Universités Entreprises (RUE)

• Echange avec Stéphane Mallat et réflexion sur la pertinence du programme tronc commun

• Rencontre avec le responsable des relations internationales de la National Taiwan University à l’ESPCI : présentation d’ITI

• Contact en cours pour visite d’entreprise

• Article dans les Échos

• Création du logo PSL / ITI et du Diplôme Supérieur de Recherche et d’Innovation de PSL / ITI (DSRI), dépôt INPI en cours.

2/ MISE EN ŒUVRE OPérationnelle Du PROGRAMME


