Cancer stratification from mutation profiles

Jean-Philippe Vert

Banff, March 28, 2017
Joint work with

Marine Le Morvan

Andrei Zinovyev
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
Patient stratification (unsupervised) from raw mutation profiles

✓ Non-Negative matrix factorisation (NMF)

✓ Desired behaviour:

✓ Observed behaviour:

Patients share very few mutated genes!
Can we replace

$$x \in \{0, 1\}^p$$

with \( p \) very large, very sparse

by a representation with more information shared between samples

$$\Phi(x) \in \mathcal{H}$$

that would allow better supervised and unsupervised classification?
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.

**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

In practice, $k$ is a free parameter optimized on the training set, typically a few 100's.
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

Related work (Hofree et al., 2013)
Performance on survival prediction

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 both total number of mutations and 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 than $k_{\text{med}} = 295$ mutations are shown, where $k_{\text{med}}$ 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 $k = k_{\text{med}}$. 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
Performance on unsupervised patient stratification

Log-rank statistic

Number of subtypes

HNSC
OV
BRCA
KIRC
GBM
SKCM
LUAD

0 50 100 150 200
time (months)

0.0
0.2
0.4
0.6
0.8
1.0
survival probability

HNSC
OV
KIRC
SKCM
Summary

- Somatic mutation profiles are challenging because
  - Little overlap between patients
  - Large variability in number of mutations
- Network smoothing / local averaging sometimes helps
  - but with current methods, looking at the direct neighbors is good enough
- Normalizing for total number of mutations is important
  - through QN or NetNorm, for example
  - this is not for biological reasons, but for mathematical reasons
  - Much room for improvement to find a good representation $\Phi(x)$
- References
  - https://hal.archives-ouvertes.fr/hal-01341856
  - https://github.com/marineLM/NetNorM
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 présentations 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 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 intern 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 les Echos

• 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

Point et étape ITI / 20 Février – 1er Juillet 2014 ! C.SURIAM – F.LEQUEUX

...


NBS representation helps to predict survival

- **NS** = Network Smoothing
- **QN** = Quantile normalization
- **NBS** = NS+QN
What is QN?
QN after network smoothing

- Patient A - NS
- Patient B - NS
- Patient A - NSQN
- Patient B - NSQN

Mutation values vs Sorted genes