Can big data cure cancer?

Jean-Philippe Vert

Miller Institute, UC Berkeley, Nov 24, 2015
Your body contains 100 trillions cells
Each cell contains a copy of the genome
The genome (DNA) differs:

- Between **species**
  - >1 nucleotide / 100

- Between **individuals**
  - 1 nucleotide / 1,000

- Between **cells**
  - 1 nucleotide / 100,000,000
  (
    \(~10 \text{ mutations per cell division}\)\)
DNA sequencing

Cost per Genome

NIH National Human Genome Research Institute
genome.gov/sequencingcosts
What is cancer(s)?

New view of cancer
Towards precision medicine

1) Human-designed « strategy », or
2) Computer-designed « strategy » ?
1) Human-designed strategy: Example of the SHIVA clinical trial

**Table 1** SHIVA treatment algorithm established to select molecularly targeted agents based on the molecular profile

<table>
<thead>
<tr>
<th>Targets</th>
<th>Targeted therapies</th>
<th>Molecular alterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>KIT, ABL1/2, RET</td>
<td>Imatinib</td>
<td>Activating mutations/amplification</td>
</tr>
<tr>
<td>PI3K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AKT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PTEK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INPP4B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRAF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDGFRB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EGFRTV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HER1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ER, I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Limits of human-designed strategies:**

- Limited to what we know (or believe)
- Limited to a few alterations, and a few drugs
- No combinatorial rule
- No weighting of evidences
- No combination of drugs
- ... and did not succeed in the clinical trial
2) Computer-designed strategy

1. Collect molecular data about many individuals

2. Collect the response to treatment

3. Let the computer figure out how to predict the response from the molecular data
Collecting data: ongoing

- http://aws.amazon.com/1000genomes/
Let the computer « learn » the rule

(a.k.a. machine learning)

n=15 samples  >>  p=2 descriptors  (easy)
Machine learning is hard when \( n << p \)

\[
n = 1e2 \sim 1e4 \\
\text{(patients)}
\]

\[
p = 1e4 \sim 1e7 \\
\text{(genes, mutations, copy numbers, ...)}
\]
One solution: reduce dimension
(a.k.a. « feature selection », « molecular signature »)
Example: Mammaprint, the 70-gene Breast cancer prognostic signature

(Van de Vijver et al 2002)
But…

Gene expression profiling predicts clinical outcome of breast cancer
Laura J. van 't Voor†, Hongyue Doli‡, Marc J. van de Vijver†, Yudong D. He†, Augustinus A. M. Hart†, Mao Mao†, Hans L. Petersen†, Karin van der Kooy†, Matthew J. Marton‡, Anke T. Witteveen†, George J. Schreiber†, Ron M. Kerkhoven†, Chris Robertst, 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 Heemsteedseweg, 1066 CX Amsterdam, The Netherlands
‡Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034.

Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer
Yixin Wang, Jan G M Klijn, Yi Zhang, Anieto M Sleuwers, Maxime P Looij, Fei Yang, Dmitri Tolantov, Mieke Timmermons, Marian E Meijer-van Gelder, Jack Yu, Tim Jatkoe, Els M J J Berns, David Atkins, John A Foekens


Only 3 genes in common

... and not really better than choosing 70 genes at random!
(Haury et al., PLoS One 2011)
Improving feature selection with prior knowledge

Can we « force » the signature to be « coherent » with a known gene network?
Example: the graph lasso

• **Step 1**: Using the network, define a subset of « candidate » signatures

• **Step 2**: Among the candidates, find the best signature to explain the data

\[
\Omega(\beta) = \sup_{\alpha \in \mathbb{R}^p; \forall i, \|\alpha_i^p + \alpha_i^q\| \leq 1} \alpha^T \beta
\]

(a convex body in \( p \) dimensions)

\[
\min_{\beta \in \mathbb{R}^p} R(f_\beta) + \lambda \Omega(\beta)
\]

(convex optimization)

(Jacob et al 2009)
Classical signature (accuracy = 0.61)
Graph lasso signature (accuracy=0.64)
Ongoing project: multiple drugs

Collaboration S. Dudoit (UC Berkeley), R. Bourgon (Genentech)
Our approach

[Diagram showing the relationship between cell line descriptors and drug descriptors through kernelized bilinear regression.]
Somehow it worked...

But the best performance is barely better than random
Conclusion

• New opportunities to exploit big data in precision medicine

• Challenging machine learning problems

• Still a long way to go before curing cancer...
Thanks!

Thanks
Alexandre d'Aspremont, Emmanuel Barillot, Anne-Claire Haury, Laurent Jacob, Pierre Mahé, Julien Mairal, Guillaume Obozinski, Franck Rapaport, Jean-Baptiste Veyrieras, Andrei Zynoviev... and all CBIO@Mines

C.SURIAM! – F.LEQUEUX

1/!
PROMOTION! DU! PROGRAMME


• Ecole! des! Mines!:! présentation! d'ITI! devant! le! conseil! de! l'enseignement;! 8! représentants! 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!


• 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


2/!
MISE! EN! ŒUVRE! OPERATIONNELLE!

The Adolph C. and Mary Sprague Miller Institute for Basic Research in Science
University of California, Berkeley

Genentech
A Member of the Roche Group