From bewilderment to enlightenment: Logic in cancer research

Wong Limsoon
Based on joint work with Wilson Goh and Sriganesh Srihari
About Limsoon

Position
Kwan-Im-Thong-Hood-Cho-Temple Chair Professor, Dept of Computer Science, NUS

Research
database systems & theory, knowledge discovery, bioinformatics & computational biology

Honours
• ACM Fellow
• FEER Asian Innovation Gold Award 2003
• ICDT Test of Time Award 2014
Two bewilderments

Breast cancer survival signatures are no better than random signatures

Mutation mutual exclusivity are not associated with synthetic lethality

And maybe some enlightenment at the end....
Some NUS numbers

- 3 campuses
  - Kent Ridge, Bukit Timah, & Outram
- 150 hectares
- 13 undergrad schools
- 4 graduate schools
- 28k undergrads
- 10k grad students
- 2.4k faculty
- 3.5k research staff
- 5.4k other staff
Story #1

BREAST CANCER-SURVIVAL SIGNATURES
A seemingly obvious conclusion

A multi-gene signature (social defeat in mice) is claimed as a good biomarker for breast cancer survival

- Cox’s survival model p-value << 0.05

A straightforward Cox’s analysis. Anything wrong?
In fact, almost all random signatures also have p-value < 0.05;

And the larger a random signature is, the more likely this happens.
Maybe significant random signatures share genes with reported signatures? Not quite…
Perhaps instead of asking whether a signature is significant, ask what makes a signature significant.
Proliferation is a hallmark of cancer

Hypothesis: Proliferation-associated genes make a signature significant

<table>
<thead>
<tr>
<th>Cutoffs</th>
<th>Counts</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NP</td>
<td>P</td>
<td>Marginals</td>
</tr>
<tr>
<td>Above 0.05</td>
<td>7043</td>
<td>19 043</td>
<td>26 086</td>
</tr>
<tr>
<td>Below 0.05</td>
<td>2766</td>
<td>19 148</td>
<td>21 914</td>
</tr>
<tr>
<td>Marginals</td>
<td>9809</td>
<td>38 191</td>
<td>48 000</td>
</tr>
</tbody>
</table>

# of random signatures w/ ≥1 prolif gene
Impact of proliferation genes on reported signatures

P-value of reported signatures, before removing proliferation genes

P-value of reported signatures, after removing proliferation genes
Many random signatures with proliferation genes are not significant;

Which proliferation genes make many random signatures significant?
Helpful analytical practices

Leverage existing data and knowledge

Careful and systematic evaluation of gene sets

Rigorous testing against as many published datasets as possible
Leverage background knowledge

Proliferation is a cancer hallmark

Good signatures with high diff in p-values before vs after removing proliferation genes:
GLINSKY, DAI, RHODES, ABBA, WHITFIELD

SPS = \{ \text{genes appearing in at least two of these good signatures} \}:
83 genes in total
81 of these are proliferation associated
Systematic evaluation

SPS genes show additive effect,

other proliferation genes don’t
Test on many datasets

For any independent dataset, a random signature has ~50% chance to be significant in it

<table>
<thead>
<tr>
<th>n</th>
<th>$(50%)^n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50.00%</td>
</tr>
<tr>
<td>2</td>
<td>25.00%</td>
</tr>
<tr>
<td>3</td>
<td>12.50%</td>
</tr>
<tr>
<td>4</td>
<td>6.25%</td>
</tr>
<tr>
<td>5</td>
<td>3.13%</td>
</tr>
<tr>
<td>6</td>
<td>1.60%</td>
</tr>
<tr>
<td>7</td>
<td>0.78%</td>
</tr>
</tbody>
</table>

How many independent datasets are needed to avoid reporting random signatures as significant?
Test on many datasets

SPS is universally significant on 7 breast cancer datasets

Random signatures (same size as SPS) are hardly universal, even though they get better p-values than known signatures on some datasets.
A theory-practice gap

~50% of random signatures are significant in 1 dataset

Red histogram is expected # of random signatures significant in n independent dataset (according to binomial distribution)

Blue histogram is observed distribution
## A related cautionary note

<table>
<thead>
<tr>
<th>NN</th>
<th>NN Acc. (%)</th>
<th>Acc. $t_1$-sparse (%)</th>
<th>Acc. $t_2$-sparse (%)</th>
<th>NPAQ r for $t_1$-sparse (%)</th>
<th>NPAQ r for $t_2$-sparse (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARCH1</td>
<td>74.00</td>
<td>78.00</td>
<td>81.00</td>
<td>20.31</td>
<td>62.50</td>
</tr>
<tr>
<td>ARCH2</td>
<td>62.00</td>
<td>73.00</td>
<td>78.00</td>
<td>12.50</td>
<td>65.62</td>
</tr>
<tr>
<td>ARCH3</td>
<td>76.00</td>
<td>82.00</td>
<td>83.00</td>
<td>45.31</td>
<td>52.34</td>
</tr>
<tr>
<td>ARCH4</td>
<td>50.00</td>
<td>64.00</td>
<td>72.00</td>
<td>17.19</td>
<td>93.75</td>
</tr>
<tr>
<td>ARCH5</td>
<td>78.00</td>
<td>82.00</td>
<td>83.00</td>
<td>74.22</td>
<td>24.22</td>
</tr>
<tr>
<td>ARCH6</td>
<td>80.00</td>
<td>11.00</td>
<td>87.00</td>
<td>37.50</td>
<td>55.47</td>
</tr>
<tr>
<td>ARCH7</td>
<td>87.00</td>
<td>89.00</td>
<td>89.00</td>
<td>6.25</td>
<td>79.69</td>
</tr>
</tbody>
</table>

Table 2: First and second column refer to the baseline model where we use BNNs with 7 different architectures. The third and fourth represent the accuracies of sparsified models with $t_1 = 0.03, t_2 = 0.05$ sparsification thresholds. The last 2 columns show NPAQ estimates for the difference between each sparsified model and the original model.

Credit: Teodora Baluta
Story #2

SYNTHETIC LETHALS
Synthetic lethality
Some models of synthetic lethality

Why interested in synthetic lethality?

Synthetic-lethal partners of frequently mutated genes in cancer are likely good treatment targets.
Synthetic lethality implies mutual exclusivity

Fact/postulate:
When a pair of genes is synthetic lethal, mutations of these two genes avoid each other

Observation:
Mutations in genes (A,B) are seldom observed in the same subjects

Conclusion by abduction:
Genes (A,B) are synthetic lethal

A seemingly obvious approach based on hypergeometric test

Mutations of genes (A,B) avoid each other if $P[X \leq |S_{AB}|] \leq 0.05$

Anything wrong with this?
Seems to work fine
Among top ME-genes, GARP score ranks correlate with mutual exclusion ranks.

But GARP scores of ME-genes (i.e. have mutually exclusive mutations to BRCA1) are similar to other genes.
The hypergeometric distribution does not reflect real-world mutations

The Hypergeometric distribution assumes:

- Mutations are independent
- Mutations have equal chance to appear in a subject

Real-life mutations are inherited in blocks; those close to each other are correlated

Some subjects have more mutations than others

⇒ Null distribution is not hypergeometric, binomial, etc.
Real-life example: Mutations of TP53 and its neighbours

(a) Genomic location of genes close to TP53

(b) CNA profile of genes close to TP53
A related cautionary note

Solution?

Group genes into genomic clusters

Test genes in far-apart genomic clusters for mutually exclusive mutations

Mutually exclusive clusters should contain synthetic-lethal & collateral-lethal gene pairs
Prediction of collateral / synthetic lethality partners of 7 DDR genes

<table>
<thead>
<tr>
<th>Gene</th>
<th>HDMI</th>
<th>ME Pairs</th>
<th>PW Confirmed</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTEN</td>
<td>0.040</td>
<td>71</td>
<td>53</td>
</tr>
<tr>
<td>TP53</td>
<td>0.025</td>
<td>51</td>
<td>36</td>
</tr>
<tr>
<td>BRCA2</td>
<td>0.030</td>
<td>34</td>
<td>22</td>
</tr>
<tr>
<td>ATM</td>
<td>0.020</td>
<td>26</td>
<td>19</td>
</tr>
<tr>
<td>CDH1</td>
<td>0.025</td>
<td>33</td>
<td>17</td>
</tr>
<tr>
<td>RB1</td>
<td>0.040</td>
<td>52</td>
<td>37</td>
</tr>
<tr>
<td>MSH3</td>
<td>0.025</td>
<td>18</td>
<td>11</td>
</tr>
</tbody>
</table>
A example of synthetic lethality: TP53-BCL2

(a) Pathway heatmaps of the TP53 group and the BCL2 group.

(b) GARP score distribution of the TP53 and BCL2 genes.
Another example: PTEN-WDR48

(a) Pathway heatmaps of the PTEN group and the WDR48 group.

(b) GARP score distribution of the PTEN and WDR48 genes.
An example of collateral lethality

FXR2 is located near TP53
FXR1 and FXR2 are paralogs that buffer each other’s function

Do FXR1 and TP53 deletions avoid each other?

Is FXR1 synthetic lethal to TP53?
Does inhibiting FXR1 lead to cell death for TP53-deleted cell lines?
Tumour bearing homozygous TP53/FXR2 co-deletion shrinks upon doxycycline-induced FXR1 knock down

Fan et al., eLife, 6:e26129, 2017
SUMMARY
Bewilderment: Breast cancer survival signatures are no better than random signatures
Enlightenment: SPS genes

Bewilderment: Mutation mutual exclusivity are not associated with synthetic lethality
Enlightenment: Collateral lethality

Happy families are all alike; every unhappy family is unhappy in its own way
References


PhD program at NUS Graduate School of Integrative Sciences and Engineering,
http://ngs.nus.edu.sg/graduate_programme.html

PhD program at NUS School of Computing,
http://comp.nus.edu.sg/programmes/pg/phdcs