Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer Omics profiles

Denis Bertrand, Kern Rei Chng, Faranak Ghazi Sherbaf, Niranjan Nagarajan et al., Genome Institution of Singapore.

Nucleic Acids Research, 2015, Vol. 43, No. 7

Introduction

• Cancer is a disease characterized with deregulated cell proliferation.

• Deregulated proliferation is caused (among others) by gene mutations effecting downstream expression of proteins that are active in the proliferation/cell cycle/cell survival pathways.

• Cancer is very heterogeneous, different genes can drive the canorous process in different patients with different types of cancer.

• In order to personalized treatment, the first step is nominate patient-specific driver candidates.

• The main challenge is to distinguish “driver genes” from “passengers” in the patient level.
Goals

• Revealing patient-specific cancer mutations in order to personalize therapeutic targets and treatment.
• Distinguishing driver mutations from passengers.
Suggested method

• “OncoIMPACT” - A “multiomics” approach, explaining frequently deregulated genes (“Phenotypes”) in cancer by association to mutated genes (SNV, CNV, indels) in a PPI network context, and assessing the Impact of these genes.
• “-Omics” involved: Genomic (SNP, Indels, CNV), Transcriptomic (RNA-seq), Interactomic.
Motivation

Map-Kinase Pathway

Proliferation
“Patient specific” Map-Kinase Pathway

p1

TF

HER2

Deregulated gene

p2

TF

HER2

Undetectable by OncoIMPACT

p3

TF

HER2

“Phenotype”

Non-deregulated gene

Driver mutation

Treatable!

No treatment

Undetectable
Inhibits heterodimerization with other receptors.

**Herceptin®**
**Potential Mechanism of Action**

- Downregulates HER2 receptor expression
- Arrests cell growth in HER2 protein-overexpressing tumors
- Potentiation of chemotherapy
  - Herceptin prevents damaged DNA from being repaired
Current methods

- Drive-net
  Connecting mutated genes (green genes on the left hand side) with deregulated genes (red genes on the right hand side) if they have “some interaction”.
  Finding driver genes by greedy set cover.

- Frequency-based
Method

Goal: *explain* deregulated gene by mutational events.

Input:

- Parameters: D,L,F
- An undirected gene-interaction graph $G = (V_m \cup V_d \cup V_r, E)$
- $V_m = \text{mutated}$
- $V_d = \text{Deregulated} \ (> F - \log \text{ fold change in expression})$
- $V_r = \text{regular}$
Method

Algorithm:
1) Finding phenotypes:
   A deregulated gene $v_d$ is considered explained by $v_m$ if there is a path $p = (v_m, v_2, \ldots, v_d)$ s.t:
   - $|p| \leq L$
   - $v_m \in V_m$
   - $\{v_2 \ldots v_d\} \in V_d$
   - $\forall i: \deg(v_i) < D$
   - $\forall i < d: (v_i, v_{i+1}) \in E$
• Can be solved by BFS in $O(|V_m|(V + E))$ time.
• *Phenotype* is a gene explained in at least T fraction of patients with the same cancer type (default 5%).
Method

2) Distinguish driver genes from back-seat:
Given a bipartite graph $G = (V_{dp} U V_{ph}, E)$
($E$ connects mutated genes with their alleged phenotypes as found in step (1))
Find the smallest subset of $V_{dp}$ covering $V_{ph}$
Method

- Can be solved by greedy Minimum Set-Cover $O(\log(n))$ approx. algorithm.

- Note: All driver genes keep their phenotypes from step (1), including the ones removed as part of the set cover algorithm.)
Method

3) Module-identification:
• A module is defined by a driver gene phenotypes explained by that gene.
• Modules that share a phenotype are merged together.
• Deregulated genes that do not belong to a path from a driver gene to a phenotype are trimmed from the module.
Method

- Module division after driver gene filtering
4) Calculate driver-gene impact = sum log folds of deregulated genes in the module.

5) **Overall impact** of gene $v_m$ is the it’s average impact across samples.

6) Output: driver gene list ordered by overall impact.

   • Possible issues: no impact normalization and permissive inclusion of genes in modules
Randomization testing

• Question: how to determine \{D,L,F\}?

• Assumption: with a suitable set of parameters, real data sets have many more associations than random data sets.

• Meaning for a given set \{d,l,t\}, the frequency with which a single gene is explained across samples is expected to be higher in the real data set.
Randomization testing

Randomization testing by the “permutation based” approach:

1) Generate random data sets by permuting gene labels for mutations and transcriptional data independently and with the same frequencies.

2) Given \{d,l,f\}, find the distribution with which each gene is explained for each random set.

3) Aggregating information across sets for each \{d,l,f\}, we compare it to the distribution for the real data given \{d,l,f\} using “Jensen-Shannon Divergence”.

4) Algorithm default = 100 Random sets.
Kullback-Leibler Divergence

• The Kullback-Leibler (KL) divergence of \( q(x) \) from \( p(x) \), denoted \( D_{KL}(q(x), ||p(x)) \), is a measure of the information lost when \( q(x) \) is used to approximate \( p(x) \).

• Typically \( p(x) \) represents the “true” distribution of data or an observations of the distribution. The measure \( q(x) \) typically represents an approximation of \( p(x) \).

\[
D_{KL}(Q||P) = \sum_{x \in X} P(x) \log\left(\frac{p(x)}{q(x)}\right) = E[\log\left(\frac{p(x)}{q(x)}\right)].
\]

• Theorem: \( D_{KL}(Q||P) \geq 0, D_{KL}(Q||P) = 0 \iff Q = P \)

• Problem: not smooth! (can not handle missing or 0 values)
Jensen-Shannon Divergence

• Solution: Jensen-Shannon Divergence:
  \[ JSD(Q||P) = \frac{1}{2} D_{KL}(Q||M) + \frac{1}{2} D_{KL}(P||M), M = \frac{1}{2} (P + Q) \]

• Smooth and Symmetric!

• Running time: \( O(n) \) for each sample, \( O(nmk) \) in total
  (\( n \) = distribution space size, \( m \) = num of samples and
  \( k \) = #random sets)

• We will look for \( \max_{\{d,l,f\}} JSD(Q||P) \)

http://web.engr.illinois.edu/~hanj/cs412/bk3/KL-divergence.pdf
http://www.math.ku.dk/~topsoe/ISIT2004JSD.pdf
Assessing phenotype genes significance

• In order to assess phenotype significance, 500 random data sets (permuting labels for genes only) were constructed to build empirical null distribution for the frequency with which a gene is explained.

• P-values were calculated for the observed frequencies.

• Correction for multiple hypothesis using FDR < 0.1 (on top of 5% frequency for phenotypes).
Proving FDR

• Recall: \( FDR = \frac{\# \text{ rejected true null hypothesis}}{\# \text{ rejected null hypothesis}} = Q \)

• Procedure: let \( H_1, H_2, \ldots, H_m \) be \( m \) hypothesis with corresponding \( p_1, p_2, \ldots, p_m \) in increasing order

• Find largest \( k \) for which \( p_k \leq \frac{k}{m} \alpha \) and reject \( H_1 \ldots H_k \)

• Theorem: for independent hypothesis and using the procedure above, \( FDR \leq \alpha \)
Proving FDR

- Lemma: For any $0 \leq m_0 \leq m$ independent p-values corresponding to true null hypotheses, and for any values that the $m_1 = m - m_0$ p-values corresponding to the false null hypotheses can take, the multiple-testing satisfies the inequality:

$$E(Q|P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) \leq \frac{m_0}{m} \alpha$$

Hence $E(Q) \leq \frac{m_0}{m} \alpha \leq \alpha$. \blacksquare
Clustering Driver-gene profiles

- To further investigate the prognostic effect of driver genes, the researchers clustered the driver gene profiles of the patients (0-1 vectors)

- Clustering by NMF- Non Negative Matrix Factorization
Non-Negative Matrix Factorization

\[
V (n \times m) \approx W (n \times k) H (k \times m)
\]

- Explain the observed data using a limited number of basis components, which when combined together approximate the original data as accurately as possible.
- Non-negative values.
- No need for orthogonally or independence.
- Performs well when clustering based on small set of features.

Objective: \( \min_{H,W \geq 0} D(V, WH) \)
Non-Negative Matrix Factorization

- NMF has an inherent clustering quality.
- $v_i = Wh_i$ hence $h_{ki}$ indicates the “affinity” of $v_i$ to cluster $k$.
- Hence $Cluster(v_i) = \max_k h_{ki}$.
- “Soft clustering” as opposed to k-means.
Gradient descent

- How to calculate W,H?
- One popular option – “Gradient Descent”
- General approach to minimize objective functions- “move in a direction that decreases the function”
- Simple case: for $f(x)$, the direction is $-f'(x)$
- For $x \in R^d$, the direction is $-\nabla f(x)$
- We will usually start with some random $x$, and iterate until $f(x)$ converges.
- We may need to make multiple repeats if the problem is not convex, and we should account the step size.
Solving NMF using Gradient Descent

- In our case: NMF objective: \( \min_{W,H \geq 0} \frac{1}{2} \| V - WH \|_F^2 \) (F- frobenius norm)

\[
f(W, H) = \frac{1}{2} \| V - WH \|_F
\]

\[
\nabla_H f(W, H) = W^T(WH - V)
\]

\[
\nabla_W f(W, H) = (WH - V)H^T
\]

- Gaujoux, A flexible R package for nonnegative matrix factorization
- Slim ESSID & Alexey OZEROV: A TUTORIAL ON NONNEGATIVE MATRIX FACTORISATION
- Wang, NON-NEGATIVE MATRIX FACTORIZATION BASED ON PROJECTED NONLINEAR CONJUGATE GRADIENT ALGORITHM
Gradient Descent

Algorithm:
1) Start with random $W, H$
2) Do until convergence:
   
   $2.1) W^{k+1} = \max\{0, W^k - \alpha \nabla_W f(W^k, H^k)\}$
   $H^{k+1} = \max\{0, H^k - \alpha \nabla_H f(W^k, H^k)\}$

   $\alpha = \text{"step size"}$

   Running time: $O(T(k(nk + mk) + mn))$

   $T = \#\text{iterations}$

   Note: small $\alpha$ - slow convergence, big $\alpha$ - might miss minimum.
Results

- Data sets: over 1000 samples from 5 cancer types from TCGA (GBM, Ovarian, Prostate, Bladder, Melanoma).

- 88 Ovarian and GBM cancer cell lines from CCLE (no controls).

- Top 10 driver genes in GBM as predicted by:
  - Frequency based vs. Oncoimpact
  - Drivenet vs. Oncoimpact
Concordance with CGC and “Pan Cancer Drivers” shows overall superior results of OncoIMPACT over frequency based and Drivenet.

- p-values are for enrichment in CGC and PCD
- Drivenet performs better on melanoma due to lack of controls
Results

- OncoIMPACT performs well with small sample size
- Stability: \(
\frac{\text{number of driver genes returned with } N \text{ samples}}{\text{number of driver genes returned with all samples}}\) (for instance 90% with ~20 samples in GBM)
- Recovery: %true positives
- Common: >5% mutational frequency
- Rare: <5% mutational frequency
Results

• Prediction on cell line datasets shows good concordance despite lack of controls.

• Overlap between predictions in samples and cell lines significant.
Validating results

1) Addition of *In Silico* random decoy mutations shows good FPR with up to 10%.

2) Driver gene predictions of OncoIMPACT on cell line dataset shows significant enrichment with ESP experiment driver genes. These results further improve with more stringent constraints on ESP experiment.
3) shRNA *in vitro* experiment:

Proliferation assay in a patient-derived melanoma cell line treated with siRNA targeting BRAF (well known melanoma driver) and TRIM24 (#671 in the list of frequently mutated gene in melanoma). *experiment data set = 160 melanoma samples and one cell line derived from melanoma patient.*
Validating results

7 genes with known oncogenic function in melanoma that were not predicted by OncoIMPACT were also tested with proliferation assay.
Out of the 7, only EZH2 show significant decrease in proliferation.
Co-drivers analysis

- One major (and frequently observed) conclusion is that driver genes tend to cluster close to one another in the network.
Most patients consist of multiple distinct modules (upper right) of which one is dominant (left).

- Dominant modules enriched with genes that play active roles in cancerous processes.
Clustering driver profiles

Glioblastoma

Ovarian Cancer

Survival Probability

Time (months)
Clustering driver profiles

• Comparison of mRNA- and driver- based stratification of patients (measured by Log Rank Test $p$-value for patient survival profiles across clusters).
Discussion/Future work

• The Authors didn’t mentioned any attempt to use weighted edges or to compare the tool to such methods.

• “Not all edges born equal”- for instance STRINGDB for human interactome: >87% of edges have a score of <0.5, ~40% of edges have <0.2 score.

• A similar approach can use the same algorithm with weighted edges and running Dijkstra instead of BFS (no major impact on running time).

• Expend the model to include modifications in order to include non-deregulated paths.
Thanks for listening!

GIVE THAT MAN

A COOKIE