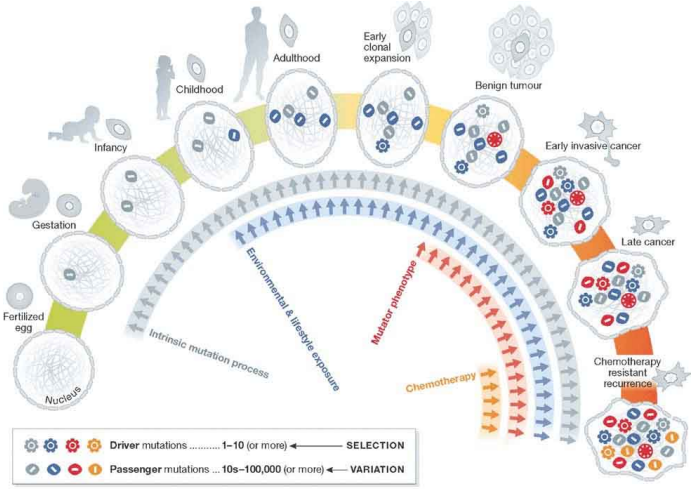
HIT'nDRIVE: Patient-Specific Multi-Driver Gene Prioritization for Precision Oncology

Raunak Shrestha

PhD Candidate, Bioinformatics, University of British Columbia
Dr. Colin Collins' Laboratory, Vancouver Prostate Centre
Dr. Cenk Sahinalp's Laboratory, Simon Fraser University / Indiana University

3rd November 2016

Background



Stratton MR. EMBO Molecular Medicine, 2011

- Cancer is mediated by somatic evolution of various alterations in genome

Background

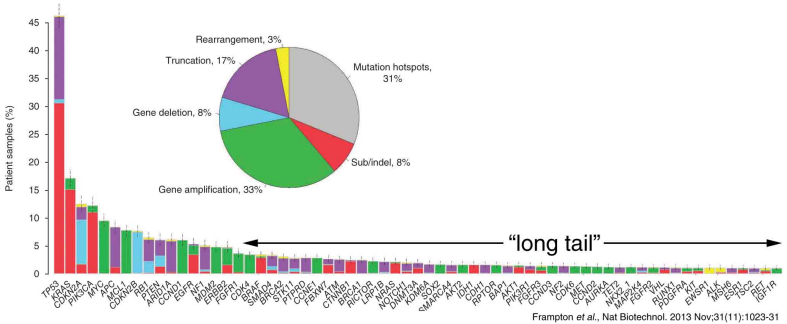
Driver vs Passenger

- Most of the alterations are neutral that provide no growth advantage to the tumor known as “passenger” alterations
- Very few alterations provide a net growth advantage and are positively selected for during tumorigenesis known as “driver” alterations

Challenge

- Driver alterations are diluted and are outnumbered by the passenger alterations
- This makes identification of driver alterations more complicated

Background



Frampton *et al.*, Nat Biotechnol. 2013 Nov;31(11):1023-31

- Most driver gene prediction methods rely on the recurrence frequency of alterations
- However, many drivers that affect only a small subset of cancer patients (long tail distribution)
- Infrequent driver genes may be functionally important and are likely to be missed by a frequency-based approach

HIT'nDRIVE

Shrestha *et al. RECOMB* 2014

- A combinatorial method that integrates genome and transcriptome data from tumor samples to prioritize genomic alterations as potential drivers using protein interaction network

HIT'nDRIVE: Multi-driver Gene Prioritization Based on Hitting Time

Raunak Shrestha[1,2,⋆], Ermin Hodzic[3,⋆], Jake Yeung[2,4,⋆], Kendric Wang[2], Thomas Sauerwald[5], Phuong Dao[6], Shawn Anderson[2], Himisha Beltran[7], Mark A. Rubin[7], Colin C. Collins[2,8], Gholamreza Haffari[9], and S. Cenk Sahinalp[3,10]

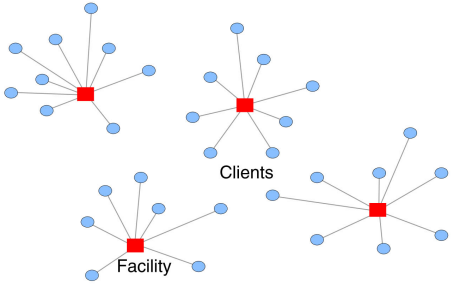
R. Sharan (Ed.): RECOMB 2014, LNBI 8394, pp. 293–306, 2014.

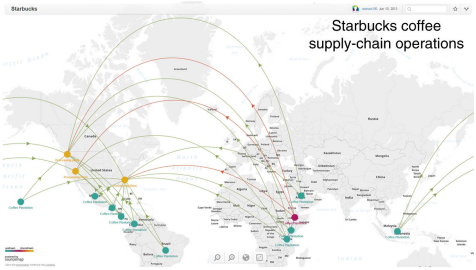
Our aim

- Identify the most parsimonious set of drivers that explain most of the observed gene expression alterations

Facility Location Problem (standard)

- Concerned with the **optimal placement of facilities** to **minimize associated costs**

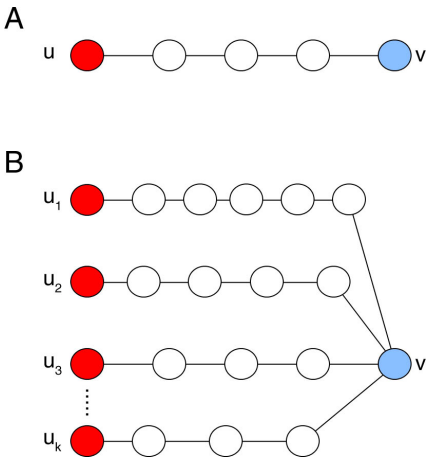




source: http://free.sourcemap.com/view/6039

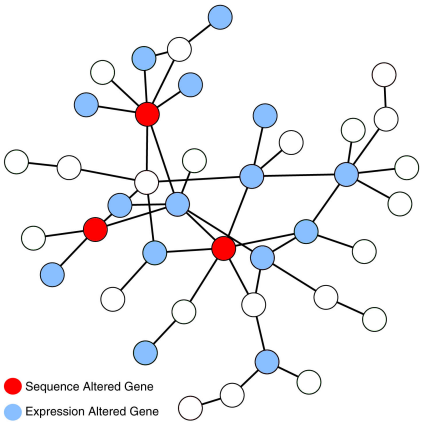
- Applications of (standard) facility location problem
  - Placement of Starbucks coffee outlets
  - Starbucks coffee supply-chain operations
  - Airline routing systems

Random Walk on Interaction Network



- In order to capture uncertainty of gene interactions we consider random walks in interaction networks, and use hitting times to measure distance
- Hitting time, $H_{u,v}$ is defined as **expected length of a random walk starting at u and visiting v for the first time**. If we let the random variable $\tau_{u,v}$ denote the number of hops in a random walk from u to v, then $H_{u,v} = E[\tau_{u,v}]$.
- Multi-hitting time $H_{U,v}$ is defined as **expected minimum length of a random walk starting at any of the nodes in set U and visiting v for the first time**. More specifically, $H_{U,v} = E[\min_{u \in U} \tau_{u,v}]$

Random Walk Facility Location (RWFL) Problem



- We define and model the problem as combinatorial optimization Random-Walk Facility Location problem

RWFL Problem

Let $\mathcal{X}$ be a set of potential driver genes and $\mathcal{Y}$ be a set of expression altered (outlier) genes. Then, for a user defined k, the solution to RWFL problem is:

$$\arg\min_{X \subseteq \mathcal{X}, |X|=k} \max_{y \in \mathcal{Y}} H_{X,y}$$

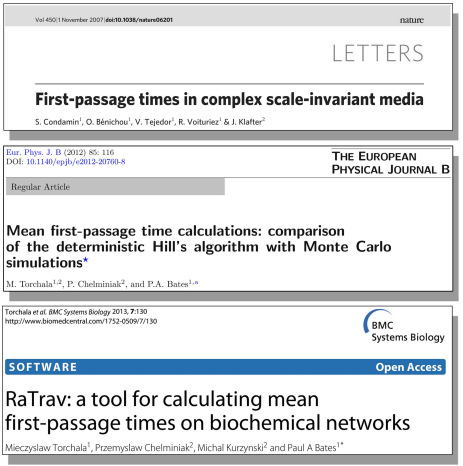
Estimating Multi-hitting Times

- RWFL problem is NP-hard to solve
- In order to overcome the difficulty of solving RWFL, we introduce the following estimate of the multi-hitting time:

$$H_{U,v} \approx \frac{1}{\sum_{i=1}^{k} \frac{1}{H_{u_i,v}}}$$

- It works well and allows us to simplify the problem, making it solvable in reasonable amount of time

Computing Single-Source Hitting Times



Vol 450|1 November 2007|doi:10.1038/nature06201 nature

LETTERS

First-passage times in complex scale-invariant media

S. Condamin[1], O. Bénichou[1], V. Tejedor[1], R. Voituriez[1] & J. Klafter[2]

Eur. Phys. J. B (2012) 85: 116
DOI: 10.1140/epjb/e2012-20760-8

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Mean first-passage time calculations: comparison of the deterministic Hill's algorithm with Monte Carlo simulations*

M. Torchala[1,2], P. Chelminiak[2], and P.A. Bates[1,a]

Torchala *et al. BMC Systems Biology* 2013, **7**:130
http://www.biomedcentral.com/1752-0509/7/130

BMC Systems Biology

SOFTWARE Open Access

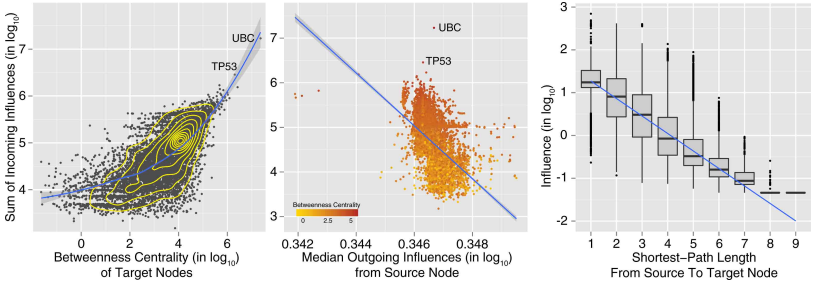
RaTrav: a tool for calculating mean first-passage times on biochemical networks

Mieczyslaw Torchala[1], Przemyslaw Chelminiak[2], Michal Kurzynski[2] and Paul A Bates[1*]

- 2007: Condamin *et al.* developed a method for accurate calculation of average hitting time in a complex network which depends on
  - Fractal Dimension: density of nodes
  - Random Walk Dimension: source-target distance in the network
- 2012: Torchala *et al.* extended Condamin *et al.* method
  - Hill's algorithm - uses transition probabilities between node
  - Efficient method to calculate average hitting time in a network
- 2013: Torchala *et al.* implements their work based on Hill's algorithm

Influence Matrix

- Influence value = inverse of average hitting time
- STRING ver.10 functional interaction network



- nodes occupying central positions in the network tend to receive more influence than the nodes in the periphery of the network ($R = 0.61$, $pvalue < 10^{-16}$)
- negative correlation between total incoming influence and the median outgoing influence of a node ($R = -0.54$, $pvalue < 10^{-16}$)

Weighted Multi-Set Cover Problem

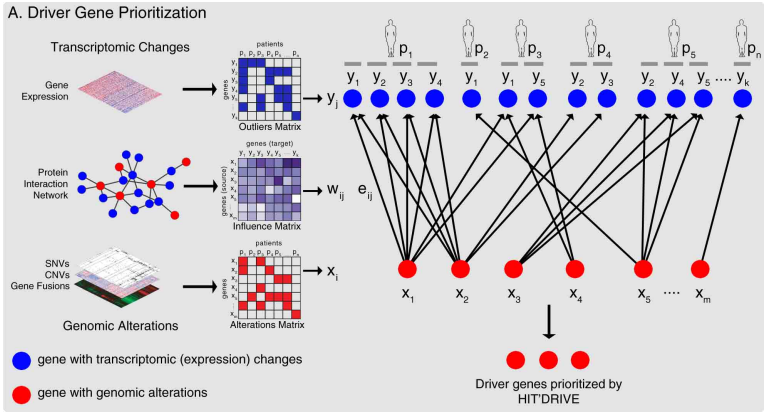
- Since RWFL is NP-hard, we reduce it to the Weighted Multi-Set Cover problem, which we can solve via ILP formulation

Weighted Multi-Set Cover (WMSC) Problem

WMSC asks to compute, the smallest set of drivers which "sufficiently" covers "most" of the patient specific expression altered genes:

$$\arg\min_{X \in \mathcal{X}} \min_{Y \subseteq \mathcal{Y}, |Y| \geq \alpha |\mathcal{Y}|} |X| \quad \text{such that} \quad \forall y \in Y : \sum_{x \in X} w_{x,y} \geq \gamma_y$$

Overview of HIT'nDRIVE Algorithm



- Now we can formulate WMSC as ILP

$$\min_{x_1,\ldots,x_{|\mathcal{X}|}} \sum_i x_i$$

s.t.

$$\forall i,j : x_i = e_{ij}$$

$$\forall j : \sum_i e_{ij} w_{ij} \geq y_j \gamma \lambda_j \sum_i w_{ij}$$

$$\sum_j y_j \geq \alpha |\mathcal{Y}|$$

$$\forall p : arg_{\beta\lambda_j}(y_j) = 1$$

$$x_i, e_{ij}, y_j \in \{0,1\}$$

Applications of HIT'nDRIVE

Shrestha et al., HITnDRIVE: Patient-Specific Multi-Driver Gene Prioritization for Precision Oncology (*In Submission*)

- Predicting cancer driver genes
- Predicting rare driver genes of cancer
- Predicting druggable driver gene targets of cancer
- Cancer sub-type classification using driver-seeded module
- Predicting driver-seeded module associated with patient's survival outcome
- Predicting mutually exclusive driver-seeded modules
- Predicting mechanisms of drug-response mechanisms

Acknowledgements

- Simon Fraser University
  - Dr. Cenk Sahinlap
  - Ermin Hodzic
- Vancouver Prostate Centre
  - Dr. Colin Collins
  - Kendric Wang
  - Jake Yeung
  - Shawn Anderson
- University of Cambridge
  - Dr. Thomas Sauerwald
- Monash University
  - Dr. Reza Haffari
- NCBI
  - Dr. Phuong Dao
- University of Padova
  - Dr. Fabio Vandin