MULTI-DRIVER GENE PRIORITIZATION BASED ON HITTING TIME

by

Ermin Hodzic
B.Sc., University of Sarajevo, Bosnia and Herzegovina, 2012

Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Summer 2014

APPROVAL

Name: Ermin Hodzic

Degree: Master of Science

Title of Thesis: MULTI-DRIVER GENE PRIORITIZATION BASED ON HITTING TIME

Examining Committee: Dr. Petra Berenbrink, Associate Professor
Chair

Dr. Cenk Sahinalp, Professor,
Senior Supervisor

Dr. Martin Ester, Professor,
Supervisor

Dr. Jian Pei, Professor,
Internal Examiner

Date Approved: August 22nd, 2014

Partial Copyright Licence

Abstract

A key challenge in cancer genomics is the identification and prioritization of genomic aberrations that potentially act as drivers of cancer. HIT'nDRIVE is a combinatorial method to identify aberrant genes that can collectively influence possibly distant "outlier" genes based on the "random-walk facility location" (RWFL) problem on an interaction network. RWFL uses "multi-hitting time", the expected minimum length of a random walk originating from any aberrant gene towards an outlier. HIT'nDRIVE aims to find the smallest set of aberrant genes from which one can reach outliers within desired multi-hitting time. It estimates multi-hitting time based on the independent hitting times and reduces the RWFL to a weighted multi-set cover problem, which it solves as an integer linear program (ILP). We apply HIT'nDRIVE to identify aberrant genes that potentially act as drivers in a cancer data set and make phenotype predictions using only the potential drivers, more accurately than alternative approaches.

keywords: drivers, cancer, multi-hitting time, interaction networks, multi-set cover

To my dear family.

Acknowledgments

I would like to thank most of all my senior supervisor, Dr. S. Cenk Sahinalp, for his guidance and for having me work with him.

And I would like to thank people who worked with me on the project which resulted in this thesis.

Contents

List of Figures

Chapter 1

Introduction

Cancer is known to be mediated by somatically acquired aberrations that accumulate in the human genome. Over the past decade, high-throughput sequencing efforts have revealed the importance of genomic aberrations in the progression of cancer [48]. During the time course of cancer evolution, tumor cells accumulate numerous genomic aberrations. However, not all aberrations are significant for the development of cancer and some of them do not contribute at all. We call such functionally inconsequential (for the progression of cancer) aberrations, "passenger" aberrations. Only a few "driver" aberrations are functionally relevant to progression of cancer and are expected to confer crucial growth advantage. The genes carrying these driver aberrations are known as "cancer genes". Identifying driver aberrations and their respective cancer genes is interesting because of their potential to be used as therapeutic targets in treatment of cancer patients. Identification of these driver aberrations and the specific genes that they alter poses a significant challenge as they are greatly outnumbered by passenger aberrations which do contribute further towards cancer heterogeneity [48, 25].

Finding cancer drivers is clearly a much harder problem than finding all genomic aberrations in cancer genome. Drivers not only seem to be outnumbered by passenger aberrations, but many cancer genes seem to be relevant to development of just a small fraction of tumors, or even cells in a tumor population. On other hand, some cancer genes, such as $TP53$ and $KRAS$, are frequently mutated in many types of cancer. This mutational heterogeneity further complicates identification of cancer drivers. The cancer is widely believed to be driven by multiple drivers, but the number of cancer drivers in a patient is believed to be just a few.

1.1 Related Work

While several methods for finding drivers of cancer have been described previously, most of them rely on the recurrence frequency of single nucleotide variants with respect to the background mutation rate in a population of tumors [24, 56]. These approaches are restricted to identifying only highly recurrent mutations as driver events. However, recent whole-genome studies have revealed that important genes may be recurrently mutated in only a small fraction of the tumor cohort under study, and can be subtype-specific [41, 9, 11]. Furthermore, personalized rare drivers are likely to arise during later stages of tumor evolution and be isolated to a small fraction of tumor cells [23, 17]. This makes mutation frequency a rather inadequate way to determine the importance of cancer genes for a particular patient, and makes it very hard to detect novel drivers specific to a patient.

Due to cancer heterogeneity, information about somatic mutations alone may not be enough to determine drivers. Methods that make use of additional information about phenotype have been designed. Perhaps the first computational method to consider large scale genomic variants as driver events is by Akavia *et al.* [1], which correlates genes with highly recurrent copy number alterations with variations in gene expression profiles within a Bayesian network. Copy number aberrations occur frequently in cancer due to genomic instability [1]. Similarly, Masica and Karchin [36] correlate gene mutation information with expression profile changes in other genes, using no prior knowledge of pathways or protein interactions. It is believed that driver mutations do not only target individual genes, but also groups of genes in cellular pathways. (Multi) Dendrix [32] assumes that on average there is just one driver aberration in a pathway in a patient, and aims to simultaneously identify multiple driver pathways, assuming mutual exclusivity of mutated genes belonging to the same pathway among patients, using either a Markov chain Monte Carlo algorithm or integer linear programming (ILP). Finally, MEMo by Ciriello *et al.* [13], identifies sets of proximally-located genes from interaction networks, forming a clique, which are also recurrently altered and exhibit patterns of mutual exclusivity across the patient population.

To the best of our knowledge, the first method to link copy number alterations to expression profile changes using an interaction network is by Kim *et al.* [30] which connects specific "causal" aberrant genes with potential targets in a protein interaction network. Similarly, method PARADIGM [53] computes gene-specific inferences using factor graphs to integrate various genomic data to infer pathways altered in a patient.

A more recent tool, HotNet by Vandin *et al.* [51], was the first to use a network diffusion approach to compute a pairwise influence measure between the genes in the (gene interaction) network and identify subnetworks enriched for mutations. TieDIE [42] also uses the diffusion model to identify a collection of pathways and subnetworks that associate a fixed set of driver genes to expression profile changes in other genes. Briefly, the network diffusion approach aims to measure the influence of one node over another by calculating the stationary proportion of a "flow" originating from the starting node, that ends up in the destination node. Since it is based on the stationary distribution, the inferences that can be made by the diffusion model are time independent. In that sense, the diffusion approach is very similar to Rooted PageRank, the stationary probability of a random walk originating at a source node, being at a given destination node.

A final method, DriverNet by Bashashati *et al* [5], also aims to correlate single nucleotide alterations with target genes' expression profile changes, but only among direct interaction partners. The novel feature of DriverNet is that it aims to find the "minimum" number of potential drivers that can "cover" targets.

1.2 Our Contributions

In this thesis we present a novel integrative method that considers potential driver events at the genomic level, i.e. single nucleotide mutations, structural or copy number changes, linking them to observed abnormal phenotype events using an interaction network to measure "influence" of possible drivers over the events.

We present HIT'nDRIVE, an algorithm that aims to identify "the most parsimonious" set of patient-specific driver genes which have sufficient "influence" over a large proportion of outlier (differentially expressed) genes. HIT'nDRIVE formulates this as a "random-walk facility location" problem (RWFL), a combinatorial optimization problem, which, to the best of our knowledge, has not been explored earlier. RWFL differs from the standard facility location problem (which is NP-hard) by its use of "multi-source hitting time" (or multi-hitting time) as an alternative distance measure between a set of aberrant genes (potential drivers) and an outlier gene. Multi-hitting time generalizes the notion of hitting time [34]: we define it as the expected minimum number of hops in which one of the random walks, which are started simultaneously from all aberrant genes from the set, reaches the outlier for the first time in the human gene or protein interaction network. In other words, our

distance measure is the expected amount of time needed for the potential driver gene set to "find" the outlier if all genes in the set "search" for the outlier at the same time by running random walks. This corresponds to the driver set influencing the changes in the outlier via random propagation of alteration events throughout the network. RWFL problem thus asks to find the smallest (the most parsimonious) set of aberrant genes from which one can reach (at least a given fraction of) all outliers within a user defined multi-hitting time. We believe that applications of RWFL problem (and the use of multi-hitting time for measuring distance among sets of nodes in networks) may extend beyond its application to driver gene identification - to influence analysis in social networks, disease networks, etc.

Since RWFL problem is NP-hard, we estimate the multi-hitting time based on the independent pair-wise hitting times of the drivers towards an outlier, which provides an upper bound on the multi-hitting time. Our experiments, in which we perform brute-force calculation of multi-hitting times on a very small sample of random genes, show that this estimate works well for the human protein interaction network.

More importantly, our estimate enables us to reduce the RWFL problem to a weighted multi-set cover problem (WMSC), for which we give an ILP formulation. Assuming our estimate is correct, the solution to WMSC problem is also a solution to RWFL problem. For the specific problem instances we consider, our ILP formulation is solvable exactly by CPLEX in less than two days on a standard PC.

Note that hitting time as a measure of influence of one potential driver on an outlier gene is quite different from the diffusion-based measures or the Rooted PageRank: hitting time essentially measures the expected distance/time between a source node and a destination node in a random walk. We argue that hitting time is a better measure to capture the influence of one (driver) node over another as it is (i) parameter free (diffusion model introduces at least one additional parameter - the proportion of incoming flow "consumed" at a node in each time step), (ii) it is time dependent (while the diffusion model and PageRank measures the stationary behavior) and (iii) it is more robust (w.r.t. small perturbations in the network; see [26]).

We also show that, by a simple Monte Carlo method in which we perform random walk simulations, the hitting time in networks with n nodes that have constant average degree and small diameter (as per the human protein interaction network) can be estimated in $\tilde{O}(n^2)$ time. For computing the hitting time in general networks, alternative methods [49] require to perform a complete matrix inversion, which takes $O(n^{2+c})$ time for some $c > 0.37$.

We have applied HIT'nDRIVE to identify genes subject to somatic mutation and copy number changes that potentially act as drivers in glioblastoma cancer. We then used the identified potential drivers to perform phenotype prediction on the cancer data set, solely based on gene expression profiles of small subnetworks "seeded" by the drivers. For that we extended the OptDis method [15] by focusing only on driver-seeded subnetworks and achieved a higher accuracy than the alternative approaches.

Chapter 2

Methods

HIT'nDRIVE naturally integrates genome and transcriptome data from a number of tumor samples for identifying and prioritizing aberrated genes as potential drivers. It "links" aberrations at the genomic level to gene expression profile alterations through a gene or protein interaction network. For that, it aims to find the *smallest* set of aberrated genes that can "explain" most of the observed gene expression alterations in the cohort. In other words, HIT'nDRIVE identifies the minimum number of potential drivers which can "cause" a user-defined proportion of the downstream expression effects observed.

HIT'nDRIVE uses a particular "influence" value of a potential driver gene on other (possibly distant) genes based on the (gene or protein) interaction network in use. In order to capture the uncertainty of interaction of genes with their neighbours, it considers a random walk process which propagates the effect of sequence alteration in one gene to the remainder of the genes through the network. The influence is defined to be the inverse of *hitting-time*, which is defined as the expected length (number of hops) of a random walk which starts at a given potential driver gene and "hits" a given target gene the first time in a (protein or gene) interaction network. More specifically, given any two nodes $u, v \in V$ of an undirected, connected graph $G = (V, E)$, let the random variable $\tau_{u,v}$ denote the number of hops of a random walk started from u that visited v for the first time. The hitting-time $H_{u,v}$, thus is defined as $H_{u,v} = E[\tau_{u,v}]$ [33]. Therefore, the smaller the hitting-time from a potential driver to an outlier is, the more "influence" is it considered to have over it - the more likely it was to influence the change in phenotype.

In order to capture synthetic lethality like scenarios, HIT'nDRIVE also considers multiple aberrated genes as potential drivers. For that purpose we measure the collective influence

of a set of potential driver genes over an outlier. We define the influence value of a set of potential driver genes on a target as the inverse of multi(source)-hitting time, i.e., the expected value of the smallest number of hops in any of the random walk processes, simultaneously started at each one of the potential drivers, and "hitting" at a given outlier for the first time. More specifically, let $U \subseteq V$ be a subset of nodes of G and $v \in (V - U)$ be a single node. We thus define the multi(source)-hitting time $H_{U,v}$ as $H_{U,v} = E[\min_{u \in U} \tau_{u,v}]$.

2.1 Problem Definition

HIT'nDRIVE formulates the process of potential driver gene discovery in terms of the "random-walk facility location" (RWFL) problem, which, for a single patient, can be described as follows.

Problem 1 (RWFL)**.** *Let G be an interation network with set of nodes V. Let $\mathcal{X} \subseteq V$ be a set of potential driver genes and $\mathcal{Y} \subseteq V$ be a set of expression altered (outlier) genes. Then, for a user defined k, HIT'nDRIVE aims to return k potential driver genes as solution to the following optimization problem:*

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

where $H_{X,y}$ denotes the multi-hitting time from the gene set X to the gene y in the interaction network G.

RWFL problem resembles the standard (minimax) "facility location" problem in which one seeks to determine a subset of nodes as facilities in a graph such that the maximum distance from any node in the graph to its closest facility is minimized. RWFL differs from standard facility location by its use of $H_{X,y}$ as a distance measure between a collection of nodes to any other node, which aims to capture the uncertainty in molecular interactions during the propagation of one or more signals, by random walks starting from one or more origins (reminiscent of the underlying Brownian motion).

Since the standard facility location is an NP-hard problem, RWFL problem is NP hard as well. As shown in the next section, we overcome this difficulty by introducing a good estimate on the multi-hitting time that helps us to reduce RWFL problem (the quality of the solution depends on the quality of our estimate) to the weighted multi-set cover problem (WMSC), which we solve through an ILP formulation in Section 2.4. (Although the use of

set-cover for representing the most parsimonious solution in a bioinformatics context is not new [27], to the best of our knowledge this is the first use of the multi-set cover formulation for maximum parsimony.) In this formulation, we use a slightly different objective: given a user defined upper bound on the maximum multi-hitting time allowed, we now aim to minimize the number of potential drivers that can "cover" (a user defined proportion of) the outlier genes. For more than one patient, we minimize the number of drivers that can "cover" (a user defined proportion of) patient-specific outliers such that each such outlier is covered by potential drivers that are aberrant in that patient. The best way to pick these "user defined constants" for a new user might be to run the algorithm multiple times and determine which combination of input parameters gives the best solution, as it is difficult for a human to estimate distance measured in hitting times.

2.2 Estimating Hitting Time on a Protein-Protein Interaction (PPI) Network

As mentioned before, HIT'nDRIVE estimates the multi-hitting time $H(U, v)$ between a set of nodes U and a single node v as a function of independent pair-wise hitting times $H(u, v)$ for all $u \in U$ - as will be shown later. However, even computing $H(u, v)$ is not a trivial task in a general graph $G = (V, E)$ as it requires a solution to a system of $|V|$ linear equations with $|V|$ variables. Below we show how to efficiently calculate $H(u, v)$ for all $u, v \in V$ for a graph $G = (V, E)$ with constant average degree and small diameter, as per the available human protein interaction network (or any small world network), using a simple Monte Carlo method. We would like to stress out that our method of calculating hitting times is highly parallelizable as all random walks performed are independent. The method is also more space-efficient, as it does not require one to store whole matrix. Keeping track of hitting times only from driver candidates is enough.

Once again, let the random variable $\tau_{u,v}$ denote the number of hops of a random walk started from u that visited v for the first time, and let $H_{max} = \max_{u,v}\{H_{u,v}\}$. Our aim is to estimate $H_{u,v}$ empirically by performing independent random walk experiments and taking the average of the observed lengths. More formally, for any given number of iterations $m > 1$ and pair $u, v \in V$, let $X_1, X_2, ..., X_m$ be a sequence of independent random variables which have the same distribution as $\tau_{u,v}$ for every $1 \leq i \leq m$. Then the *empirical hitting time* is defined as $\tilde{H}_{u,v} = \frac{1}{m} \cdot \sum_{i=1}^{m} X_i$. The following theorem shows how fast $\tilde{H}_{u,v}$ converges to

the actual hitting-time $H_{u,v}$.

Theorem 1. *Assume that G is a graph such that the maximum hitting time satisfies $H_{max} \leq Cn$ for some constant $C > 0$ and let $u, v \in G$ be an arbitrary pair of nodes. Then for any $\varepsilon \in [\frac{1}{n^4}, 1]$, after $m = (128C)^2(1/\varepsilon)^2(\log_2 n)^3$ iterations, the returned estimate $\tilde{H}_{u,v}$ satisfies*

$$\Pr\left[|\tilde{H}_{u,v} - H_{u,v}| \leq \varepsilon n\right] \geq 1 - n^{-3}.$$

Moreover, with probability at least $1 - n^{-7}$, the total number of random walk hops made is at most $m \cdot 32Cn \log_2 n = O((1/\varepsilon)^2 n \log^4 n)$.

Proof. By Markov's inequality, we have for every $1 \leq i \leq m$, $\Pr[X_i \geq 2H_{max}] \leq \frac{1}{2}$. Dividing the random walk into k consecutive sections of length $2H_{max}$ yields for any integer $k \geq 1$,

$$\Pr[X_i \geq k \cdot 2H_{max}] \leq \left(\frac{1}{2}\right)^k.$$

Let us define $\mathcal{E}$ as the event which occurs if for every $1 \leq i \leq m$, $X_i \leq 32 \log_2 n \cdot H_{max}$. By the union bound,

$$\Pr[\mathcal{E}] \geq 1 - m \cdot \left(\frac{1}{2}\right)^{16 \log_2 n} = 1 - m \cdot n^{-16} \geq 1 - n^{-7},$$

where the last inequality is due to the definition of m and the lower bound on ε. Observe that if the event $\mathcal{E}$ occurs, then the total number of random walk steps made is at most $m \cdot 32 \log_2 n \cdot H_{max} \leq m \cdot 32Cn \log_2 n$, which yields the second statement of the theorem.

We now prove the first statement of the theorem. Conditioning the expectation of X_i yields

$$E[X_i] = \Pr[\mathcal{E}] \cdot E[X_i \mid \mathcal{E}] + \Pr[\neg\mathcal{E}] \cdot E[X_i \mid \neg\mathcal{E}].$$

By the memoryless property of the random walk,

$$E[X_i \mid \neg\mathcal{E}] \leq 32Cn \log_2 n + H_{max}.$$

Consequently,

$$E[X_i] \leq 1 \cdot E[X_i \mid \mathcal{E}] + n^{-7} \cdot (32Cn \log_2 n + Cn) \leq E[X_i \mid \mathcal{E}] + \frac{\varepsilon}{2} \cdot n.$$

By definition of $\mathcal{E}$, $E[X_i] \geq E[X_i \mid \mathcal{E}]$, and combining the previous two inequalities yields

$$|E[X_i] - E[X_i \mid \mathcal{E}]| \leq \frac{\varepsilon}{2} \cdot n. \tag{2.1}$$

Note that in the probability space conditional on the event $\mathcal{E}$, the random variables X_1, X_2, ..., X_m are mutually independent, identically distributed random variables with expectation $E[X_1 \mid \mathcal{E}]$ each. Furthermore, each random variable takes values in $\{1, 2, ..., 32Cn\log_2 n\}$. Hence Hoeffding's inequality gives for any $\lambda > 0$,

$$\Pr\left[\left|\sum_{i=1}^{m} X_i - m \cdot E[X_1 \mid \mathcal{E}]\right| \geq \lambda \,\middle|\, \mathcal{E}\right] \leq 2 \cdot \exp\left(-\frac{2\lambda^2}{m \cdot (32Cn\log_2 n)^2}\right).$$

Choosing $\lambda = 64C\sqrt{m} \cdot n \cdot (\log_2 n)^{1.5}$ yields

$$\Pr\left[\left|\sum_{i=1}^{m} X_i - m \cdot E[X_1 \mid \mathcal{E}]\right| \geq 64C\sqrt{m} \cdot n \cdot (\log_2 n)^{1.5} \,\middle|\, \mathcal{E}\right] \leq 2n^{-4}.$$

With our lower bound on $\Pr[\mathcal{E}]$, we conclude that

$$\begin{aligned}
&\Pr\left[\left|\sum_{i=1}^{m} X_i - m \cdot E[X_1 \mid \mathcal{E}]\right| \leq 64C\sqrt{m} \cdot n \cdot (\log_2 n)^{1.5}\right] \\
&\geq \Pr[\mathcal{E}] \cdot \Pr\left[\left|\sum_{i=1}^{m} X_i - m \cdot E[X_1 \mid \mathcal{E}]\right| \leq 64C\sqrt{m} \cdot n \cdot (\log_2 n)^{1.5} \,\middle|\, \mathcal{E}\right] \\
&\geq (1 - n^{-7}) \cdot (1 - 2n^{-4}) \geq 1 - n^{-3}.
\end{aligned}$$

If the above event occurs, then our returned estimate $\tilde{H}_{u,v}$ satisfies

$$\left|\tilde{H}_{u,v} - E[X_1 \mid \mathcal{E}]\right| < \frac{64C(\log_2 n)^{1.5}}{\sqrt{m}} \cdot n = \frac{\varepsilon}{2} \cdot n,$$

where the last equality follows from the definition of m. Combining this with equation (2.1) yields

$$\left|\tilde{H}_{u,v} - E[X_1]\right| \leq \left|\tilde{H}_{u,v} - E[X_1 \mid \mathcal{E}]\right| + |E[X_1 \mid \mathcal{E}] - E[X_1]| = 2 \cdot \frac{\varepsilon}{2} \cdot n = \varepsilon \cdot n,$$

which completes the proof of the first statement as $X_1 = H_{u,v}$. □

To obtain the empirical estimates of all n^2 hitting times $H_{u,v}$ efficiently, observe that taking a single random walk starting from u until all other nodes are visited gives an estimate for all n hitting times $H_{u,v}$ with $v \in V$. Since for fixed $v \in V$, all m estimates for $H_{u,v}$ (coming from m iterations) are independent, we conclude by the first statement of Theorem 1 and the union bound that with probability at least $1 - n^{-2}$, for fixed $u \in V$ all n estimates $\tilde{H}_{u,v}$ approximate $H_{u,v}$ up to an additive error of ϵn. Similarly, the total number of random

walk hops to obtain all these n approximations is $O((1/\epsilon)^2 n \log^4 n)$ with probability at least $1 - n^{-6}$. Finally, we do the above procedure for all n possible starting vertices $u \in V$, so that with probability at least $1 - n^{-1}$, we have an ϵn-additive approximation for each of the n^2 hitting times, and the total number of random walk hops is $O((1/\epsilon)^2 n^2 \log^4 n)$ with probability at least $1 - n^{-5}$.

2.3 Estimating Multi-Source Hitting Time via Single-Source Hitting Times

Given $U = \{u_1, u_2, \ldots, u_k\}$, we now show how to estimate multi-hitting time $H_{U,v}$ towards a node v by a function of independent pairwise hitting times $H_{u_i,v}$ for all $u_i \in U$, in order to overcome the difficulty of solving the RWFL problem. We use the following estimate:

$$H_{U,v} \approx \frac{1}{\sum_{i=1}^{k} \frac{1}{H_{u_i,v}}} \tag{2.2}$$

Let the conductance of graph G be defined as $\Phi(G) = \min_{\emptyset \subsetneq S \subsetneq V} \frac{|E(S,V\setminus S)|}{\min\{(S),(V\setminus S)\}}$. Many real-world networks, including preferential attachment graphs, are known to have large conductance [38]. For such graphs, our next theorem provides mathematical evidence for the accuracy of our estimate in (2.2).

Theorem 2. *Let $G = (V, E)$ be any graph with constant conductance $\Phi > 0$. Then there is an integer $C = C(\Phi) > 0$ such that, given an integer k, a set of nodes $U = \{u_1, u_2, \ldots, u_k\}$ and node $v \in V$ satisfying $\frac{1}{k \cdot \frac{\deg(v)}{2|E|}} \geq \log^{1.5} n$, the following inequality holds:*

$$H_{U,v} \leq C \cdot \frac{1}{\sum_{i=1}^{k} \frac{\deg(v)}{2|E|}}.$$

In particular, for any pair of nodes $u, v \in V$ with $\deg(v) \leq \frac{2|E|}{\log^{1.5} n}$, we have $H_{u,v} = O(\frac{|E|}{\deg(v)})$.

For the proof of Theorem 2, it will be convenient to consider a lazy version of the random walk which stays at the current node in each step with probability $1/2$. Note that any hitting time (single-source or multi-source) of the lazy verion of the random walk is always an upper bound on the corresponding hitting time of the standard random walk.

Lemma 1. *Let $G = (V, E)$ be a graph with constant conductance $\Phi > 0$. For any pair of nodes $u, v \in V$ and number of steps t with $\omega(\log n) \leq t \leq \frac{2|E|}{\deg(v)}$, let $\mathcal{A}_{u,v,t}$ be the event that a random walk starting from u visits v within t steps. Then*

$$\Pr\left[\mathcal{A}_{u,v,t}\right] \geq \frac{\Phi^2}{280} \cdot t \cdot \frac{\deg(v)}{2|E|}.$$

Proof. We first record the following useful inequality (cf. [33]). Let $P^s_{x,y}$ be the probability that a random walk starting at x visits node y in step s. Then,

$$\left| P^s_{x,y} - \frac{\deg(y)}{2|E|} \right| \leq \sqrt{\frac{\pi(y)}{\pi(x)}} \cdot \lambda^t_{\max},$$

where $\pi(w) = \frac{\deg(w)}{2|E|}$ for any $w \in V$, $\lambda_{\max} = \max\{\lambda_2, |\lambda_n|\}$ with $1 = \lambda_1 \geq \cdots \geq \lambda_n > -1$ being the eigenvalues of the transition matrix P. Since the random walk has loop probability $1/2$, $\lambda_n \geq 0$ and thus $\lambda_{\max} = \lambda_2$. Furthermore, by Cheeger's inequality, $\lambda_2 \leq 1 - \frac{\Phi^2}{8}$. Hence

$$\left| P^s_{x,y} - \frac{\deg(y)}{2|E|} \right| \leq \sqrt{\frac{\pi(y)}{\pi(x)}} \cdot \left(1 - \frac{\Phi^2}{8}\right)^t,$$

which implies for every s with $t/2 \leq s \leq t$, as $t = \omega(\log n)$,

$$\left| P^s_{u,v} - \frac{\deg(v)}{2|E|} \right| \leq n^{-4}.$$

Let X be the random variable counting the number of visits to v within the time-interval $[t/2, t]$. Then, from the above,

$$\frac{t}{2} \cdot \frac{\deg(v)}{2|E|} \leq X \leq 2t \cdot \frac{\deg(v)}{2|E|}.$$

To apply the second moment method, we will now analyze the variance of X, denoted by $V[X]$. Note that $X = \sum_{s=t/2}^{t} X_s$, where $X_s = 1$ if the random walk visits u in step s and

$X_s = 0$ otherwise. Then,

$$\begin{aligned}
V[X] &\leq \sum_{s=t/2}^{t} E[X_s] + 2 \sum_{t/2 \leq s < s' \leq t} \Pr[X_s = 1 \wedge X_{s'} = 1] - \Pr[X_s = 1] \cdot \Pr[X_{s'} = 1] \\
&= \sum_{s=t/2}^{t} E[X_s] + 2 \sum_{t/2 \leq s < s' \leq t} \Pr[X_s = 1] \cdot (\Pr[X_{s'} = 1 \mid X_s = 1] - \Pr[X_{s'} = 1]) \\
&\leq E[X] + 2 \sum_{t/2 \leq s < s' \leq t} \left(\frac{\deg(v)}{2|E|} + n^{-4} \right) \cdot \left(\left(\frac{\deg(v)}{2|E|} + (1 - \frac{\Phi^2}{8})^{s'-s} \right) - \left(\frac{\deg(v)}{2|E|} - n^{-4} \right) \right) \\
&\leq E[X] + 2 \sum_{t/2 \leq s \leq t} \sum_{1 \leq i \leq t/2} \left(\frac{\deg(v)}{2|E|} + n^{-4} \right) \cdot \left((1 - \frac{\Phi^2}{8})^{i} + n^{-4} \right) \\
&\leq E[X] + 2 \sum_{t/2 \leq s \leq t} \left(\frac{\deg(v)}{2|E|} + n^{-4} \right) \cdot \left(\frac{8}{\Phi^2} + t/2 \cdot n^{-4} \right) \\
&\leq E[X] \cdot \left(2 + \frac{32}{\Phi^2} \right) + O(n^{-2}) \leq \frac{35}{\Phi^2} \cdot E[X].
\end{aligned}$$

By the Paley-Zygmund inequality, for any $0 < \delta < 1$,

$$\Pr[X \geq \delta \cdot E[X]] \geq (1-\delta)^2 \cdot \frac{E[X]^2}{V[X] + E[X]^2} \geq (1-\delta)^2 \cdot \frac{1}{\frac{35}{\Phi^2} \cdot \frac{1}{E[X]} + 1} \geq (1-\delta)^2 \cdot \frac{\Phi^2}{2 \cdot 35} \cdot E[X],$$

where the last inequality follows from $E[X] \leq 2$ which holds thanks to our upper bound on t. Choosing $\delta = \frac{1}{2}$ implies, as X is an integer-valued random variable,

$$\Pr[A_{u,v,t}] = \Pr[X \geq 1] \geq \Pr\left[X \geq \frac{1}{2} \cdot E[X] \right] \geq \frac{\Phi^2}{8 \cdot 35} \cdot E[X],$$

and due to the lower bound on $E[X]$ derived earlier, the proof is finished. □

With the lemma at hand, we are now able to complete the proof of Theorem 2.

Proof. For any integer $\alpha \geq 1$, define $\tau = \tau(\alpha) := \alpha \cdot \frac{280}{\Phi^2} \cdot \frac{1}{\sum_{i=1}^{k} \frac{\deg(u)}{2|E|}}$. For any $1 \leq i \leq k$, let $\mathcal{E}_i$ be the event that the random walk starting from v_i does *not* visit u within τ steps. By partitioning the τ steps into consecutive sections of length $\log^{1.5} n$ and applying Lemma 1 to every section, we conclude that

$$\Pr[\mathcal{E}_i] \leq \left(1 - \frac{\Phi^2}{280} \cdot \log^{1.5} n \cdot \frac{\deg(u)}{2|E|} \right)^{\tau / \log^{1.5} n} \leq \exp\left(-\tau \cdot \frac{\Phi^2}{280} \cdot \frac{\deg(u)}{2|E|} \right).$$

As all k random walks are independent, it follows that

$$\Pr\left[\bigwedge_{i=1}^{k} \mathcal{E}_i \right] = \prod_{i=1}^{k} \Pr[\mathcal{E}_i] \leq \exp\left(-\tau \cdot \sum_{i=1}^{k} \frac{\Phi^2}{280} \cdot \frac{\deg(u)}{2|E|} \right) = \exp(-\alpha) \leq 2^{-\alpha}.$$

Hence the expected multi-source hitting time can be estimated as follows,

$$H_{\{v_1,\ldots,v_k\},u} \leq \frac{280}{\Phi^2} \cdot \frac{1}{\sum_{i=1}^{k} \frac{\deg(u)}{2|E|}} \cdot \sum_{\alpha=1}^{\infty} \alpha \cdot 2^{-\alpha} \leq \frac{560}{\Phi^2} \cdot \frac{1}{\sum_{i=1}^{k} \frac{\deg(u)}{2|E|}}$$

□

Note that the bound in Theorem 2 differs from our estimate in equation (2.2) in that $\frac{1}{H_{u_i,v}}$ is replaced by $\frac{\deg(v)}{2|E|}$. However, for graphs with constant conductance, we have $H_{u,v} \leq H_{\pi,v} + O(\log n)$, where $H_{\pi,v}$ is the hitting time for a random walk starting according to the stationary distribution π, given by $\pi(w) = \frac{\deg(w)}{2|E|}$ for every $w \in V$. Hence $\frac{2|E|}{\deg(v)} = H_{v,v} \leq H_{\pi,v} + O(\log n)$. Since $H_{\pi,v} = \sum_{u \in U} \pi(u) \cdot H_{u,v}$, it follows that, given any fixed node v, it holds for "most nodes" u that $H_{u,v}$ is not much smaller than $\frac{2|E|}{\deg(v)} - O(\log n)$.

2.4 Reformulation of RWFL as a Weighted Multi-Set Cover Problem

As mentioned before, since RWFL is NP-hard we reduce it to the weighted set cover problem via our estimate of the multi-hitting time. We solve the new problem via an ILP formulation. This formulation also generalizes RWFL to allow patient-specific drivers and outlier genes. Consider a bipartite graph $G_{bip}(\mathcal{X}, \mathcal{Y}, \mathcal{E})$ where $\mathcal{X}$ is the set of aberrant genes, $\mathcal{Y}$ is the set of patient-specific expression altered genes, and $\mathcal{E}$ is the set of edges among the two partitions. If gene g_i is mutated in a patient p, we set edges between g_i and all of the expression altered genes in the same patient (g_j, p) where the edges are weighted by the inverse pairwise-hitting times $w_{i,j} := H^{-1}_{g_i,g_j}$; see the Figure 2.1 for more details.

We now slightly reformulate RWFL problem by introducing limit on the maximum allowed multi-hitting time:

$$\arg\min_{X \subseteq \mathcal{X}} |X| \qquad \text{such that} \qquad \max_{y \in \mathcal{Y}} H_{X,y} \leq \Delta, \tag{2.3}$$

where Δ is the maximum allowed multi-hitting time from the drivers to any expression altered gene. Based on this formulation, we define a minimum weighted multi-set cover (WMSC) problem on G_{bip}, whose solution provides an exact solution to RWFL problem, provided our estimate of the multi-hitting times is accurate.

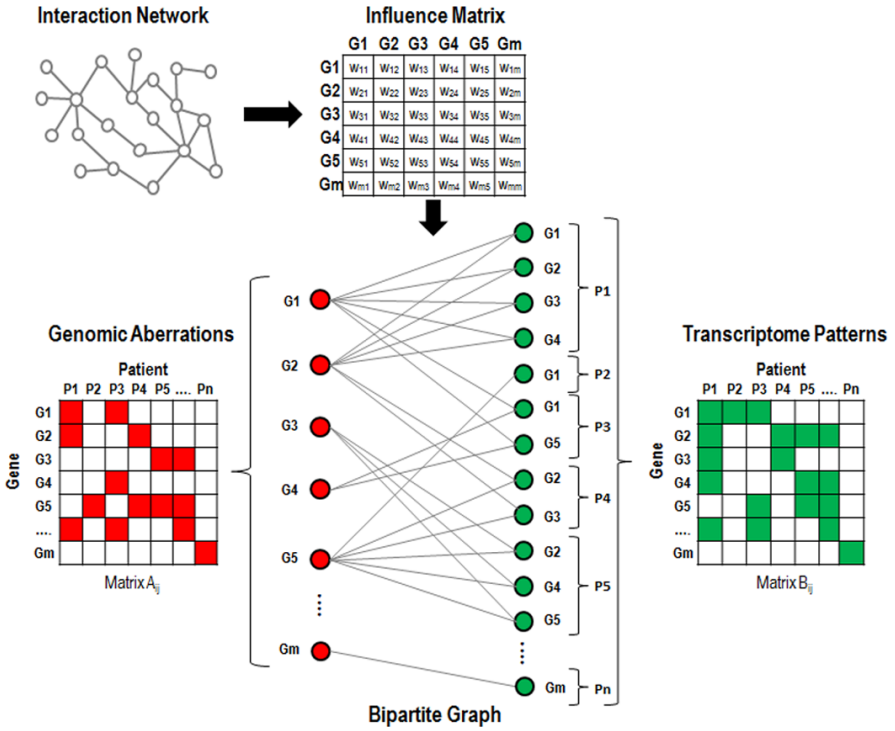

Figure 2.1: **Schematic overview of construction of bipartite graph in HIT'nDRIVE.** The influence matrix derived from the interaction network contains the *inverse* hitting times between every pair of genes. A and B are gene-patient matrices showing the genomic abberations and expression alteration events, respectively. The red color in A indicates the aberration status of a gene in a patient. Similarly, the green color in B indicate expression altered genes in a patient. The edges in the bipartite graph are weighted by the inverse hitting times within the PPI network.

Problem 2 (WMSC)**.** *Given G_{bip}, WMSC asks to compute as the potential driver gene set, the smallest set which "sufficiently" covers "most" of the patient-specific expression altered genes:*

$$\arg\min_{X\subseteq\mathcal{X},Y\subseteq\mathcal{Y},|Y|\geq\alpha|\mathcal{Y}|}|X| \quad \textit{such that} \quad \forall y\in Y: \sum_{x\in X} w_{x,y}\geq\gamma_y \tag{2.4}$$

where $0<\alpha\leq 1$ represents the fraction of patient-specific expression altered genes that we believe are causally linked to the potential drivers, and γ_y represents for each outlier how "influenced" by chosen driver set we want it to be (or equivalently, how distant, measured by multi-hitting time, from the chosen driver set we tolerate it to be).

The right-hand-side of the constraints in (2.3) and (2.4) are related by $H^{-1}_{X,y} \approx \sum_{x \in X} H^{-1}_{x,y}$, as mentioned in Section 2.3. The introduction of γ_y makes it possible to control the minimum amount of "influence" required for *individual* expression alteration events (each patient potentially indicates a unique expression alteration event for each gene).

2.5 An ILP Formulation for WMSC

$\min_{x_1,..,x_{|\mathcal{X}|}} \sum_i x_i$
s.t.
(1) $\forall i, j : x_i = e_{ij}$
(2) $\forall j : \sum_i e_{ij} w_{ij} \geq y_j \gamma \sum_i w_{ij}$
(3) $\sum_j y_j \geq \alpha |\mathcal{Y}|$
(4) $x_i, e_{ij}, y_j \in \{0, 1\}$

Figure 2.2: ILP formulation.

We formulate WMSC as an ILP and solve it using an off-the-shelf ILP solver. The ILP formulation for our combinatorial optimization problem is as Figure 2.2 where there is a binary variable $x_i,\ y_j,\ e_{ij}$, respectively, for each potential driver, expression alteration event, and edge in the bipartite graph.

- The first constraint ensures that a selected driver contributes to the "coverage" (by its influence) of each of the expression alteration events it is connected to - in each patient.
- The second constraint ensures that selected (patient-specific) driver genes cover at least a (γ) fraction of the sum of all incoming edge weights to each expression alteration event. This constraint corresponds to setting a lower bound on the joint influence (i.e. setting an upper bound on the multi-hitting time), as estimated by out method, of selected (patient specific) drivers on an expression alteration event.
- The third constraint ensures that the selected driver genes collectively cover at least an α fraction of the set of expression alteration events.
- The fourth constraint simply asks all the variables to be binary so as to resemble "picking" corresponding elements in G_{bip}.

2.6 Evaluation Framework

Evaluating computational methods for predicting cancer drivers, given genomic and transcriptomic data, is challenging in the absence of the ground truth (i.e. follow-up biological experiments). We refer to previous studies [5] that observe the overlap between predicted driver genes and known cancer genes compiled in public resources such as the Cancer Gene Census (CGC) database [22] or the Catalogue of Somatic Mutations in Cancer (COSMIC) database [21] and we provide those numbers as well.

However, since we do not want to restrict our evaluation of the predicted drivers to overlap with known cancer genes, we mainly focused on testing whether our predictions provide insight into the cancer phenotype and improve classification accuracy on an independent cancer dataset. The classifiers we evaluate are based on network "modules", a set of functionally related genes (e.g. in a signaling pathway), which are connected in an interaction network and include at least one potential driver. They then use module features, such as the average expression of genes in the module, for phenotype classification. Using such module features, we hope that the classifier in use does not *overfit* on rare drivers and is able to *generalize* the signal coming from rare drivers to new patients.

For classification purposes we primarily use OptDis [15] for *de novo* identification of gene modules in the interaction network which include (i.e. are seeded by) at least one predicted driver gene. Briefly, OptDis performs supervised dimensionality reduction on the set of connected subnetworks (modules). It projects the high dimensional space of all connected subnetworks to a user-specified lower dimensional space of subnetworks such that, in the new space, the samples belonging to the same (different) class are closer (respectively, more distant) to one another with respect to a normalized distance measure (typically L_1). For each predicted driver gene, OptDis starts with a subnetwork of size 1, containing only the gene, and expands it into a connected subnetwork of a limited size such that its "discriminative score" (a linear combination of the average in-class distance and out-class distance [15]) is maximized.

Since the human PPI network has a small diameter, there is significant overlap between many modules seeded by potential driver genes. In order to limit the number of overlapping modules (and achieve further dimensionality reduction) we first compute the top 10 modules seeded by each driver gene that have the best individual discriminative scores. The modules seeded by all potential drivers are then collectively sorted based on their discriminative

score. Among these modules, we greedily pick a subset in a way that the i^{th} module is added to our result subset R if its maximum pairwise gene overlap with any module already in R is no more than a user-defined threshold.

Chapter 3

Results

3.1 Datasets

We use a publicly available cancer dataset representing matched genomic aberrations (somatic mutation, copy-number aberration) and transcriptomic patterns (gene-expression data) of 156 Glioblastoma Multiforme (GBM) samples [41] from The Cancer Genome Atlas (TCGA). We make use of a global network of protein-protein interactions (PPI) from the Human Protein Reference Database (HPRD) version April 2010 [44] to derive the influence values based on the hitting time. We use the same PPI Network for module identification using our modification to OptDis. We ran HIT'nDRIVE with different combinations of values for the parameters α and γ and Figure 3.1-A shows number of drivers picked for those values.

3.2 Genomic Drivers for Glioblastoma Multiforme (GBM)

3.2.1 Evaluation Based on CGC and COSMIC Databases.

To assess whether the genes identified by HIT'nDRIVE are essential players in cancer, we first analyzed the concordance of the predicted drivers with the genes annotated in CGC and COSMIC database. Gene sets resulting with the parameters $\gamma = 0.7$ and $\alpha = \{0.1, 0.2, ..., 0.9\}$ were analyzed (Figure 3.1-B).

The remainder of results are obtained for parameter values $\gamma = 0.7$ and $\alpha = 0.9$ this results in 107 driver genes covering the majority (22933) of outlier genes in 156 patients.

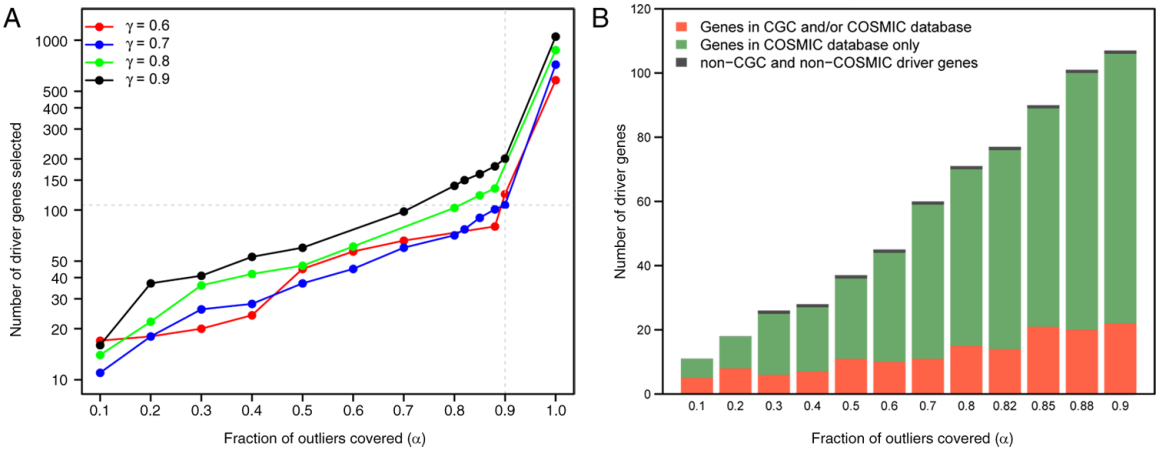

Figure 3.1: **Behavior of HIT'nDRIVE as a function of α and γ.** (A) The number of selected drivers and covered outliers as α increases for various values of γ. (B) Concordance of GBM driver genes with that of COSMIC and Cancer Gene Census database for $\gamma = 0.7$.

3.2.2 Phenotype Classification Using Dysregulated Modules Seeded with the Predicted Drivers.

We evaluated the driver genes identified by HIT'nDRIVE using phenotype classification (as described in Section 2.6 and results are shown in Figure 3.2). Briefly, drivers identified from the TCGA dataset were used as seeds for discovering discriminative subnetwork modules. The module expression profiles were used to classify normal vs. glioblastoma samples through repeated cross-validation on the validation dataset. First, HIT'nDRIVE using hitting time based influence values, was compared against DriverNet, which greedily identifies driver genes using direct gene interactions from the HPRD network. Across the appreciable range of discriminative modules discovered by OptDis, HIT'nDRIVE demonstrates better accuracy in classifying the cancer phenotype, with a maximum accuracy of 97.05% and a mean accuracy of 94.52% (Figure 3.2). Next, comparing the HIT'nDRIVE deduced drivers against a comparable number of genes with the highest node-degrees in the PPI network reveals a clear advantage to HITnDRIVE. This trend was observed when genes were used as individual classification features (blue vs. orange plots) as well as when they were used

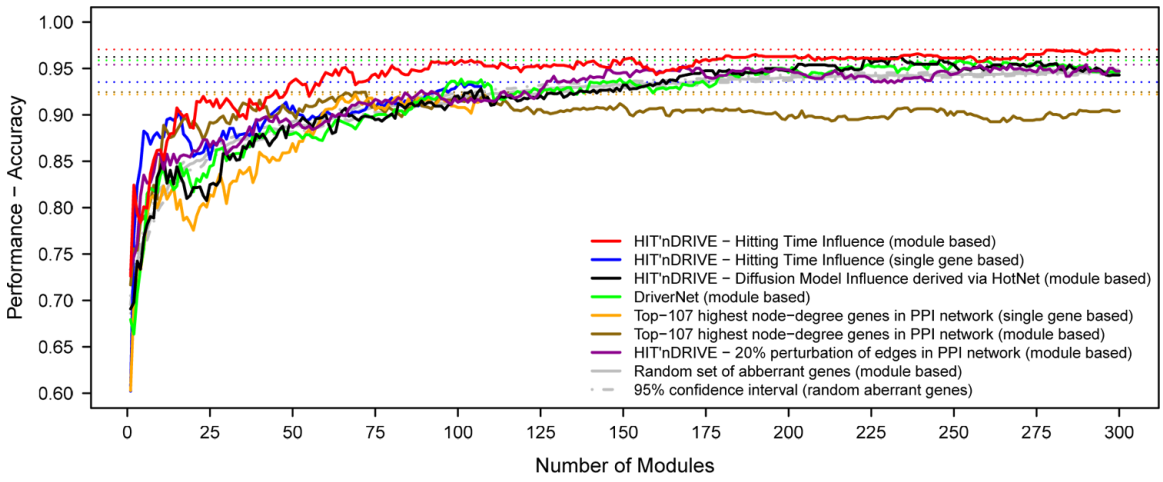

Figure 3.2: **Phenotype classification using the identified drivers obtained by various methods.** The dysregulated sets of modules seeded by the 107 chosen drivers are used to predict phenotype in the validation dataset using using k-nearest neighbour classifier with k=1. We used the HPRD-PPI Network for module identification using our modification to OptDis.

as seeds for module-based features (red vs. brown plots). Comparing the classification accuracy of HITnDRIVE deduced drivers against 107 genes randomly selected from the entire list of aberrant genes (red vs. grey plots) provides additional support for the relevance of drivers selected by HITnDRIVE. This is also confirmed by comparing the performance of hitting-time based influence values against those derived from the diffusion model [51] (red vs. black plots) both employed by HITnDRIVE.

3.2.3 Evaluating our multi-hitting time estimate.

For the purpose of evaluating our multi-hitting time estimate in the human PPI network, we ran a small test. We picked the following 10 genes at random: ATP7A, BMP15, CNPY2, FHL1, PDZD4, PIK3CG, RAB3D, TRIM3, TSPY1, ZRSR2. On this set, we computed the exact solution to the RWFL problem for 3 "facilities" using a brute-force approach: CNPY2, RAB3D and TRIM3. Then we applied Hit'nDrive where only the above 10 genes were kept on the left-hand side of the bipartite graph, using parameter value of $\gamma = 0.7$. After taking α to be as high as 0.99, the solution obtained included the three genes offered by the exact solution, i.e. CNPY2, RAB3D, TRIM3, plus one more gene, TSPY1. This

experiment, together with the discussion following Theorem 2, suggests that our estimate of multi-hitting time works well in practice.

3.2.4 Sensitivity of HIT'nDRIVE to Small Perturbations of the PPI Network.

We perturbed the PPI network by swapping endpoints of 20% edges at random and recalculated pairwise hitting times. We observed that almost all changes in the hitting times are less than 10% relative to the original values, most of them being between 1% and 5%. However, impact on accuracy of classification using HIT'nDRIVE output can be noticed in Figure 3.2.

3.2.5 Prediction of Frequent and Rare Drivers.

The 107 driver genes nominated by HIT'nDRIVE are aberrated at varying frequencies in the tumor population (Figure 3.3-A). CHEK2 and EGFR are the two most frequently aberrated drivers (at 46.8% and 42.3% respectively), followed by CDKN2A (31.4%), MTAP (30.1%) and CDKN2B (29.5%). Some of these frequent drivers harbour different types of genomic aberrations in different patients. For example, EGFR shows somatic mutation and high copy-number gain in 14.2% and 32.7% of the patients, respectively. Similarly, PTEN harbours somatic mutation in 12.8% and homozygous deletion in 3.9% of the patients. Amplification in EGFR, PDGFRA, mutations in CHEK2, TP53, PTEN, RB1, and deletions in CDKN2A have been previously associated with GBM [41, 8, 54]. HIT'nDRIVE also identified infrequent drivers, which we defined as genes that are genomically aberrant in at most 2% of the cases. Out of 26 (16.66%) rare driver genes identified, four genes (MYST4, FLI1, BMPR1A and BRCA2) were implicated in the CGC database. Despite being aberrant in a small fraction of patients, the rare drivers are specifically associated with cancer development, DNA repair, cell growth and migration, cell death and survival. Some rare drivers like MAG and BMPR1A have also been closely linked with GBM progression [37, 43].

3.2.6 Prediction of Low-degree and High-degree Drivers.

The drivers predicted by HIT'nDRIVE include a number of well-known high-degree "hubs" such as TP53, EGFR, RB1 and BRCA1, which occupy the central position (with high degree

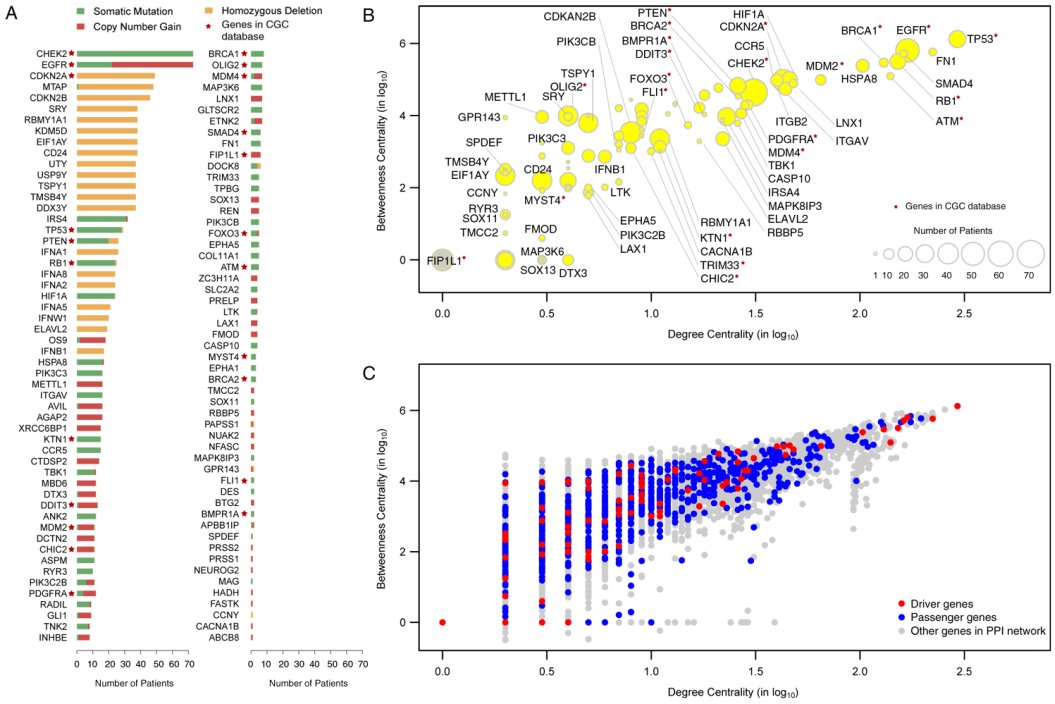

Figure 3.3: **Characteristics of driver genes of GBM predicted by HIT'nDRIVE.** (A) Recurrence frequency of the aberration in the driver genes predicted by HIT'nDRIVE. (B) The centrality of the predicted drivers in the PPI network.The size of the circles is proportional to the recurrence frequency of the genomic aberration of the gene. (C) Centrality of the "driver" and "passenger" genes is colored by red and blue dots respectively; all other nodes in the PPI network apart from the driver and passenger genes are represented as grey dots.

and high betweenness, i.e. the proportion of shortest paths between all pairs of nodes that go through that node, and high degree - computed by the igraph [14] R package.) in the PPI network (Figure 3.3-B). If these genes are perturbed, they dysregulate several other genes and the associated signaling pathways. Moreover, HIT'nDRIVE also identified low-degree genes (such as FIP1L1, SOX11 and RYR3) that reside in the periphery of the PPI network. Some of these low-degree genes are only aberrant in a small fraction of patients. Since driver genes and passenger genes display similar network characteristics (Figure 3.3-C), and identified driver genes have both low and high degrees in the network, HIT'nDRIVE likely selects drivers irrespective of known network biases.

Chapter 4

Conclusion

We have presented HIT'nDRIVE, a combinatorial method to capture the collective effects of driver gene aberrations on possibly distant "outlier" genes based on what we call the "random-walk facility location" (RWFL) problem. We introduced the notion of "multi-source hitting time" and presented efficient and accurate methods to estimate it based on single-source hitting time in large-scale networks. We applied HIT'nDRIVE to identify genes subject to somatic mutation and copy number aberrations in GBM. Our results showed that the predicted driver genes identified by HIT'nDRIVE are well-supported in databases of important cancer genes. Furthermore, these drivers were able to perform phenotype predictions more accurately than those identified by the alternative approaches, even when applied to a different dataset. Importantly, the discovery of these drivers was not biased by the frequency of aberrations among patients and/or the degree of a gene in the PPI network, not their central position in the network. It shows that our method is able to discover novel drivers as well as personalized drivers.

Since our framework is very general in the sense of its input, our approach can easily integrate various aberration types such as single nucleotide changes, copy number changes, structural variations, and splice variations. Furthermore, it can be straightforwardly extended to incorporate epigenome and/or gene-fusions data. As gene networks increase in density and volume of interaction, HIT'nDRIVE will be able to capture such improvements naturally. Finally our method is well suited to identify patient-specific driver-aberrations which can potentially be used as therapeutic targets.

We believe that the applications of multi-hitting time that we introduced can be extended beyond its use for driver gene discovery and can be used for analysis in (social) networks as

it represents a general measure of distance of a node from a whole set of other nodes in the network, assuming random walk model, which reflects the network topology.

Good phenotype prediction power of the modules identified in the last step of our framework suggests that our method is also able to discover driver pathways. The way we identify driver modules, by extending a small subnetwork around driver nodes, agrees with the assumption (made in some of the relate work) that drivers pathways typically contain a single mutated gene per patient, on average. We are conducting further work on that matter, hoping to further extend and improve our method.

Bibliography

[1] Uri David Akavia, Oren Litvin, Jessica Kim, Felix Sanchez-Garcia, Dylan Kotliar, et al. An integrated approach to uncover drivers of cancer. *Cell*, 143(6):1005–17, December 2010. 2

[2] Ludmil B. Alexandrov, Serena Nik-Zainal, David C. Wedge, Samuel a. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Bø rresen Dale, Sandrine Boyault, Birgit Burkhardt, Adam P. Butler, Carlos Caldas, Helen R. Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörd, John a. Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinsk, Natalie Jäger, David T. W. Jones, David Jones, Stian Knappskog, Marcel Kool, Sunil R. Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C. Munshi, Hiromi Nakamura, Paul a. Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V. Pearson, Xose S. Puente, Keiran Raine, Manasa Ramakrishna, Andrea L. Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N. Schumacher, Paul N. Span, Jon W. Teague, Yasushi Totoki, Andrew N. J. Tutt, Rafael Valdés-Mas, Marit M. van Buuren, Laura van t Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R. Yates, Jessica Zucman-Rossi, P. Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M. Grimmond, Reiner Siebert, Elías Campo, Tatsuhiro Shibata, Stefan M. Pfister, Peter J. Campbell, and Michael R. Stratton. Signatures of mutational processes in human cancer. *Nature*, August 2013.

[3] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106, 2010.

[4] Derek W. Barnett, Erik K. Garrison, Aaron R. Quinlan, Michael P. Strmberg, and Gabor T. Marth. Bamtools: a c++ api and toolkit for analyzing and managing bam files. *Bioinformatics*, 27(12):1691–1692, 2011.

[5] Ali Bashashati, Gholamreza Haffari, Jiarui Ding, Gavin Ha, Kenneth Lui, et al. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome biology*, 13(12):R124, December 2012. 3, 17

[6] Nicole C Berchtold, David H Cribbs, Paul D Coleman, Joseph Rogers, Elizabeth Head, et al. Gene expression changes in the course of normal brain aging are sexually dimorphic. *Proceedings of the National Academy of Sciences of the United States of America*, 105(40):15605–10, October 2008.

[7] Marija Buljan, Guilhem Chalancon, Sebastian Eustermann, Gunter P Wagner, Monika Fuxreiter, et al. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Molecular cell*, 46(6):871–883, 2012.

[8] Cancer Genome Atlas Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–8, October 2008. 22

[9] Cancer Genome Atlas Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–15, June 2011. 2

[10] Cancer Genome Atlas Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–15, June 2011.

[11] Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–7, July 2012. 2

[12] Han-Yu Chuang, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker. Network-based classification of breast cancer metastasis. *Molecular systems biology*, 3(140):140, January 2007.

[13] Giovanni Ciriello, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome research*, 22(2):398–406, February 2012. 2

[14] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. 23

[15] P. Dao, K. Wang, C. Collins, M. Ester, A. Lapuk, and S. C. Sahinalp. Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics*, 27(13), Jul 2011. 5, 17

[16] Charles J David and James L Manley. Alternative pre-mrna splicing regulation in cancer: pathways and programs unhinged. *Genes & development*, 24(21):2343–2364, 2010.

[17] Li Ding, Timothy J Ley, David E Larson, Christopher a Miller, Daniel C Koboldt, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–10, January 2012. 2

[18] e.

[19] e.

[20] Uriel Feige. A tight lower bound on the cover time for random walks on graphs. *Random Struct. Algorithms*, 6(4):433–438, 1995.

[21] Simon a Forbes, Nidhi Bindal, Sally Bamford, Charlotte Cole, Chai Yin Kok, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic acids research*, 39(Database issue):D945–50, January 2011. 17

[22] P Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, et al. A census of human cancer genes. *Nature reviews. Cancer*, 4(3):177–83, March 2004. 17

[23] Mel Greaves and Carlo C Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–13, January 2012. 2

[24] Chris Greenman, Richard Wooster, P Andrew Futreal, Michael R Stratton, and Douglas F Easton. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics*, 173(4):2187–98, August 2006. 2

[25] Christopher Greenman, Philip Stephens, Raffaella Smith, Gillian L Dalgliesh, Christopher Hunter, et al. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–8, March 2007. 1

[26] John Hopcroft and Daniel Sheldon. Manipulation-resistant reputations using hitting time. In *Algorithms and Models for the Web-Graph*, pages 68–81. Springer, 2007. 4

[27] Fereydoun Hormozdiari, Can Alkan, Evan E. Eichler, and S. Cenk Sahinalp. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome research*, 19(7):1270–1278, July 2009. 8

[28] E. Karakoc, A. Cherkasov, and S. C. Sahinalp. Distance based algorithms for small biomolecule classification and structural similarity search. *Bioinformatics*, 22(14):e243–251, Jul 2006.

[29] Yarden Katz, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods*, 7(12):1009–15, December 2010.

[30] Yoo-Ah Kim, Stefan Wuchty, and Teresa M Przytycka. Identifying causal genes and dysregulated pathways in complex diseases. *PLoS computational biology*, 7(3):e1001095, March 2011. 2

[31] Anna V Lapuk, Chunxiao Wu, Alex Wyatt, er W, Andrew McPherson, et al. From sequence to molecular pathology, and a mechanism driving the neuroendocrine phenotype in prostate cancer. *The Journal of pathology*, 227(3):286–97, July 2012.

[32] Mark D M Leiserson, Dima Blokh, Roded Sharan, and Benjamin J Raphael. Simultaneous identification of multiple driver pathways in cancer. *PLoS computational biology*, 9(5):e1003054, May 2013. 2

[33] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2008. 6, 12

[34] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, May 2007. 3

[35] Elizabeth a Maher, Cameron Brennan, Patrick Y Wen, Laura Durso, Keith L Ligon, et al. Marked genomic differences characterize primary and secondary glioblastoma subtypes and identify two distinct molecular and clinical secondary glioblastoma entities. *Cancer research*, 66(23):11502–13, December 2006.

[36] David L Masica and Rachel Karchin. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer research*, 71(13):4550–61, July 2011. 2

[37] L McKerracher, S David, D L Jackson, V Kottis, R J Dunn, et al. Identification of myelin-associated glycoprotein as a major myelin-derived inhibitor of neurite growth. *Neuron*, 13(4):805–11, October 1994. 22

[38] Milena Mihail, Christos H. Papadimitriou, and Amin Saberi. On certain connectivity properties of the internet topology. *J. Comput. Syst. Sci.*, 72(2):239–251, 2006. 11

[39] Anastasia Murat, Eugenia Migliavacca, Thierry Gorlia, Wanyu L Lambiv, Tal Shay, et al. Self-Renewal Signature and High Epidermal Growth Factor Receptor Expression Associated With Resistance to Concomitant Chemoradiotherapy in Glioblastoma. 26(18), 2013.

[40] Hiroko Ohgaki, Pierre Dessen, Benjamin Jourde, Sonja Horstmann, Tomofumi Nishikawa, Pier-Luigi Di Patre, Christoph Burkhard, Danielle Schüler, Nicole M Probst-Hensch, Paulo César Maiorka, Nathalie Baeza, Paola Pisani, Yasuhiro Yonekawa, M Gazi Yasargil, Urs M Lütolf, and Paul Kleihues. Genetic pathways to glioblastoma: a population-based study. *Cancer research*, 64(19):6892–9, October 2004.

[41] D Williams Parsons, Sian Jones, Xiaosong Zhang, Jimmy Cheng-Ho Lin, Rebecca J Leary, Philipp Angenendt, et al. An integrated genomic analysis of human glioblastoma multiforme. *Science (New York, N.Y.)*, 321(5897):1807–12, September 2008. 2, 19, 22

[42] Evan O Paull, Daniel E Carlin, Mario Niepel, Peter K Sorger, David Haussler, et al. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics (Oxford, England)*, pages 1–8, September 2013. 3

[43] S G M Piccirillo, B a Reynolds, N Zanetti, G Lamorte, E Binda, et al. Bone morphogenetic proteins inhibit the tumorigenic potential of human brain tumour-initiating cells. *Nature*, 444(7120):761–5, December 2006. 22

[44] T S Keshava Prasad, Kumaran Kandasamy, and Akhilesh Pandey. Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Methods in molecular biology (Clifton, N.J.)*, 577:67–79, January 2009. 19

[45] Benjamin Purow and David Schiff. Advances in the genetics of glioblastoma: are we reaching critical mass? *Nature reviews. Neurology*, 5(8):419–26, August 2009.

[46] Satu-Leena Sallinen, Tarja Ikonen, Hannu Haapasalo, and Johanna Schleutker. CHEK2 mutations in primary glioblastomas. *Journal of neuro-oncology*, 74(1):93–5, August 2005.

[47] Matthias Simon, Michael Ludwig, Rolf Fimmers, Ralph Mahlberg, Angelika Müller-Erkwoh, Gertraud Köster, and Johannes Schramm. Variant of the CHEK2 gene as a prognostic marker in glioblastoma multiforme. *Neurosurgery*, 59(5):1078–85; discussion 1085, November 2006.

[48] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–24, April 2009. 1

[49] Prasad Tetali. Design of on-line algorithms using hitting times. *SIAM J. Comput.*, 28(4):1232–1246, 1999. 4

[50] Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[51] Fabio Vandin, Eli Upfal, and Benjamin J Raphael. Algorithms for detecting significantly mutated pathways in cancer. *Journal of computational biology : a journal of computational molecular cell biology*, 18(3):507–22, March 2011. 3, 21

[52] Fabio Vandin, Eli Upfal, and Benjamin J Raphael. De novo discovery of mutated driver pathways in cancer. *Genome research*, 22(2):375–85, February 2012.

[53] Charles J Vaske, Stephen C Benz, J Zachary Sanborn, Dent Earl, Christopher Szeto, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics (Oxford, England)*, 26(12):i237–45, June 2010. 2

[54] Roel G W Verhaak, Katherine a Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer cell*, 17(1):98–110, January 2010. 22

[55] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis a Diaz, et al. Cancer genome landscapes. *Science (New York, N.Y.)*, 339(6127):1546–58, March 2013.

[56] Ahrim Youn and Richard Simon. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics (Oxford, England)*, 27(2):175–81, January 2011. 2