FUNCTIONAL CHARACTERISTICS OF CANCER DRIVER GENES
IN COLORECTAL CANCER

by

GURKAN BEBEK

Submitted in partial fulfillment of the requirements for the degree of Master of
Science

Clinical Research Scholars Program

CASE WESTERN RESERVE UNIVERSITY

May 2017
CASE WESTERN RESERVE UNIVERSITY

SCHOOL OF GRADUATE STUDIES

We hereby approve the thesis/dissertation of

Gurkan Bebek

candidate for the degree of Master of Science*.

Committee Chair
Masaru Miyagi

Committee Member
Thomas LaFramboise

Committee Member
Jennifer Yu

Date of Defense
April 5, 2017

* We also certify that written approval has been obtained for
any proprietary material contained therein.
Dedicated to my family.
Contents

1 Introduction .......... 1
  1.1 Colorectal Cancer .......... 1
  1.2 Driver Genes in Cancer .......... 2
  1.3 Contributions .......... 3

2 Methods .......... 7
  2.1 Cancer Driver Gene Networks .......... 7
  2.2 Quantifying Network Similarity .......... 8
    2.2.1 Constructing the Alignment Graph .......... 8
    2.2.2 Alignment Scoring .......... 11
    2.2.3 Scoring Nodes of the Alignment Graph .......... 11
    2.2.4 Scoring Sequence Similarity .......... 13
    2.2.5 Scoring Functional Similarity .......... 13
    2.2.6 Overall Similarity Between Two Interactions .......... 15
    2.2.7 Scoring the Alignment Graph and Normalization of Scores .......... 15
    2.2.8 Finding the Optimal Alignment .......... 16
  2.3 Gene Ontology-Based Classification of Genes .......... 17
  2.4 Sequence (Protein Families) Based Classification of Genes .......... 18
  2.5 Sequence and Gene Ontology Based Classification of Genes .......... 18
  2.6 Somatic mutation profiles of tumor samples and statistical evaluation .......... 19
3 Results and Discussion

3.1 Cancer driver genes and their synergistic activities .......................... 20
3.2 Finding similarities among synergistic networks ............................... 22
3.3 Clustering Cancer Driver Genes .................................................. 22
3.4 Testing synergistic activities by comparing somatic mutation profiles
   of tumor samples ............................................................................. 26
3.5 Validation of results with cell line epistasis experiments .................... 27

4 Conclusion ...................................................................................... 30

A Appendix ......................................................................................... 32

A.1 Gene Ontology Based Classification of Cancer Driver Genes ............. 32
A.2 Sequence Based Classification of Cancer Driver Genes ...................... 32

References ......................................................................................... 35
### List of Figures

1. Overview of the study design ........................................... 4
2. Alignment graph nodes ............................................. 10
3. Algorithm for within species network alignment .................. 12
4. Example alignment of two CAN-gene networks .................... 21
5. Histogram of normalized scores ................................... 23
6. Classification of CAN-gene networks and observed somatic mutations in various tumors ........................................ 24
7. Hierarchical clustering of CAN-genes ................................ 25
8. Validation Experiments .................................................. 28
9. Multidimensional Scaling of knockdown experiments in triplicate . . 29
10. Classification of CAN-genes based on their gene ontology annotations ........................................ 33
11. Classification of CAN-genes based on their Pfam families ........ 34
Functional Characteristics of Cancer Driver Genes
in Colorectal Cancer

Abstract

by

GURKAN BEBEK

Colorectal Cancer is the third most commonly diagnosed cancer and third leading cause of cancer death in both men and women. Recent studies have led to the discovery of cancer driver genes whose mutations are mostly observed in low frequencies compared to the total number of tumors analyzed, suggesting that mutations in distinct subsets of driver genes are sufficient for tumorigenesis. We hypothesize that comparing the networks by which the driver genes contribute to tumorigenesis can generate more meaningful functional similarities. We develop an algorithm for within-species network similarity quantification. We compare subnetworks corresponding to colorectal cancer driver genes using this tool to group driver genes. We validate our approach with cell line experiments and show we can predict driver gene similarities based on functional similarities. The relationships between driver genes whose mutations lead to similar phenotypic outcomes are best understood through comparison of pathways that are dysregulated.
1 Introduction

1.1 Colorectal Cancer

Colorectal cancer (CRC) is a major worldwide health problem owing to its high prevalence and mortality rates. CRC is the third most commonly diagnosed cancer and third leading cause of cancer death in both men and women. It is estimated that about 134,490 people were diagnosed with CRC in the USA, and about 49,190 people will die of the disease. Colon and rectum cancer represents 8.0% of all new cancer cases in the U.S \[1\]. Incidence of CRC is higher in men than women. The risk factors for CRC increases with age. The median age at diagnosis is about 70 years in developed countries. If diagnosed early CRC is also one of the most curable types of cancer with 5-year survival rates as high as 90%. Most CRCs could be prevented with current clinical practices such as screening tests \[? ? ? \].

Recent advances in the diagnosis and treatment of CRC improved the overall outcome of patients through new screening methods, biomarker and genomic analysis, personalized therapies and chemotherapy. However, many patients with advanced and metastatic tumors will still die from the disease. Novel diagnosis and treatment options are still needed for these patients.

There are various risk factors associated with colon cancer. These include the personal history of colorectal cancer or polyps, family history of colon cancer, inherited syndromes that increase colon cancer risk, diabetes, obesity, smoking, alcohol use, diet etc. While inherited gene mutations make up only a small percentage of colon
cancers, they can increase an individual’s risk of cancer significantly.

Most colorectal cancer cases are sporadic. They develop slowly over several years through the adenoma–carcinoma sequence [? ? ]. The major treatment options for CRC are surgery, neoadjuvant radiotherapy (for patients with rectal cancer), and adjuvant chemotherapy (for patients with stage III/IV and high-risk stage II colon cancer). While 5-year relative survival rate for stage I disease is %90, stage IV patients survival rate is slightly greater than 13% [? ].

1.2 Driver Genes in Cancer

In recent years, projects utilizing next generation sequencing technologies have resulted in an excess of genomic data. Studies that sequenced cancer exomes allowed us to understand complex diseases such as colorectal and breast cancer [? ], ovarian cancer [? ] and glioblastoma multiforme (GBM) [? ] in greater depth. These studies confidently recovered genes that are frequently mutated in different cancers. Further analyses revealed so-called candidate cancer genes (CAN-genes), which are shown to drive carcinogenesis (hence named “driver” genes), as opposed to the genes that are also mutated in carcinogenic samples, but likely have no impact on tumor growth (passenger genes). Recent research efforts have introduced methods for identifying such sets of genes for non-small cell lung carcinoma, glioblastoma multiforme, and ovarian cancer using computational methods [? ? ? ? ].

One of the key challenges associated with the introduction of these CAN-genes is to understand the mechanistic link between the mutations in these genes and the car-
cinogenic phenotype. Some researchers utilized network approaches to characterize the synergistic relationships among such genes. A variety of approaches were presented for the identification of protein subnetworks (small groups of proteins that are linked to each other through protein-protein interactions) that would describe pathways that are likely to be operative when particular sets of mutations are observed [? ? ? ]. However, these methods have not yet adequately explained the observation that, although CAN-genes are key drivers of carcinogenesis, they are mostly observed in very low abundance in the patient groups. For instance, the most frequently mutated GBM driver gene, TP53, is mutated in at most 27% of GBM patients [? ]. In colorectal cancer patients, mutations of well-known driver genes, other than the highly mutated APC gene [? ], are also found in lower frequencies than expected [? ]. However, via downstream alterations manifested by several of these driver mutations, cancers that share similar characteristics develop. Therefore, it is quite likely that cancer manifests itself via synergistic activities of various combinations of driver gene mutations.

1.3 Contributions

In this work, we present a comparative network analysis approach to further study cancer driver genes in terms of their functional similarity, with the objective of gaining insights into the relationships between these driver genes. The proposed computational framework for comparative analysis of cancer driver genes is illustrated in Figure ?? . Overall, the proposed framework provides a computational method for
Figure 1: Overview of the study design.

A. The cancer driver genes are identified mostly by exome sequencing [? ? ].

B. The driver genes are imputed into the signaling pathway search framework [? ]. Using this framework, networks connecting each colorectal cancer (CRC) driver gene with $APC$ are generated.

C. All pairs of the top cancer driver gene networks are aligned to each other, utilizing protein family information and gene ontologies. Using this measure of similarity, hierarchical clustering is performed to classify driver genes networks. The results are validated computationally and experimentally.
studying the functional characteristics of CAN-genes by approaching them from the standpoint of network similarity. Our hypothesis is that network similarities will better describe the synergistic similarities of CAN-genes than traditional ontology-based comparison of these genes. This analysis is carried out within a specific cancer type.

To test this hypothesis, we must first determine the networks by which each of the driver genes contribute to tumorigenesis. We obtain these networks by integrating annotations and tissue-specific microarray data \[?\]. Next, we produce organism-specific alignments of these networks by optimizing a novel measure of edge similarity based on sequence information and gene ontology annotations. We then use cluster analysis to test our hypothesis. Our results show that driver genes that are part of more similar signaling subnetworks are grouped together, while naive, non-network-based measures produce less meaningful results. We also validate our approach with an epistasis analysis of a candidate pair.

An important contribution of our approach is that, instead of directly quantifying the similarity between the driver genes themselves, we quantify the similarity between the pathways (or more generally PPI subnetworks) associated with each driver gene. In order to quantify subnetwork similarity, we develop a novel network-alignment algorithm for the specific purpose of aligning PPI subnetworks that originate from the same organism, i.e. within-species network alignment. Although there exists a rich literature on comparing PPI networks of different species \[? \ldots ?\] and in principle these approaches would be directly applicable to the comparison of our driver gene subnetworks, there are several considerations that prohibit direct application of these methods to these subnetworks. In particular, we identify the
following as the key characteristics of a network alignment algorithm that will be used to compare driver gene subnetworks:

1. The alignment should be global in the sense we should be able to compute a single “similarity score” between the two subnetworks being compared.

2. There should not be any ambiguous mapping (i.e. every mapped gene from one network should be mapped to only one gene from the other network).

3. Two identical genes in different subnetworks (which is not possible in cross-species network alignment) should have the maximum score, as they clearly have an identical function.

Various solutions to the network alignment problem have been proposed for cross-species alignment, but none of them feature all of the above characteristics. For instance, IsoRank [?] is a global network alignment tool that finds matches of a protein in one PPI network in another PPI network if the former’s neighbors are good matches for the latter’s neighbors. While, IsoRank satisfies (1) and (2), it does not satisfy (3), as it uses the PageRank algorithm where the score of a protein alignment is related to the scores of its neighbors. Therefore, the alignment of two identical genes may not be the maximum score, because their neighbors may not be identical. All other cross-species alignment algorithms [?] focus on local alignments, and hence do not satisfy the first requirement.
2 Methods

2.1 Cancer Driver Gene Networks

In this study, we utilize an integrative -omics methodology, where we first generate CAN-gene networks by which the driver genes contribute to tumorigenesis. It has been shown that mutation of $APC$ is the first hit in colorectal cancer, followed by the mutation of a set of 22 cancer driver genes [? ]. To generate the proposed tumorigenesis pathways for each of the driver genes, we therefore generate protein-protein interaction networks that connect each of the driver genes with $APC$. To do so, we first obtain a tissue-specific microarray gene expression data collected in replicates from the $Apc^{1638N+/−}$ mouse model (available through the Gene Expression Omnibus data series GSE19338). Using the Blossom algorithm [? ] we search publicly available PPI databases [? ? ] and score these by integrating the tissue-specific gene expression data, gene ontologies [? ], and known pathways [? ? ? ? ]. $APC$ and 22 other CAN-genes are linked via networks using this network discovery method. While we have utilized these networks to demonstrate utility of our approach, any other set of networks demonstrating synergistic links between driver genes could be utilized in this pipeline. The Blossom networks and the approach have been validated within [? ] and [? ] and various synergy studies of pairwise driver genes in other studies.
2.2 Quantifying Network Similarity

While quantifying the functional similarity between two PPI subnetworks, our goal is to quantify 1) the functional similarity between proteins that are part of their subnetworks and 2) topological consistency of the interactions in the network. In other words, for two subnetworks to be considered functionally similar, both the proteins and the context in which they interact with each other should be similar

In order to incorporate both topology and protein similarity, we formulate the network similarity problem as a network alignment problem, where proteins in a network are matched to proteins in another network. The resulting score is the sum of the scores of the individual alignments of the proteins (See 2.3 Alignment Scoring), penalized by topological inconsistency. A network alignment algorithm optimizes this score, to produce the highest quality alignment.

We represent the two subnetworks as graphs $G_A = (V_A, E_A)$, $G_B = (V_B, E_B)$, where the vertices $v \in V$ correspond to proteins/genes, the edges $e \in E$ correspond to interactions. We define an alignment as a map $A : E_0 \to E_B$, $E_0 \subseteq E_A, |E_A| \leq |E_B|$. $A$ is constrained to be one-to-one: $\forall e_1, e_2 \in E_0, A(e_1) = A(e_2) \implies e_1 = e_2$.

We define an alignment score $W(A) = \sum_{i=0}^{\left|E_0\right|} w(e_i)$, where $w(e_i)$ is the score of alignment of interaction $e_i$ with $A(e_i)$. The best alignment is $A$ that maximizes $W(A)$.
2.2.1 Constructing the Alignment Graph

In order to find the best alignment between $G_A$ and $G_B$, a computationally feasible algorithm for optimizing the alignment score must be developed. When aligning PPI networks across species, Guo et al. [?] construct an alignment graph, where each node represents a pair of possibly alignable interactions. The edges of the alignment graph are based on network connectivities and these edges are weighted according to the score of mapping the two pairs of edges that are mapped to each other and the topological inconsistency created by this mapping. They then use a greedy heuristic to find non-redundant subgraphs on this graph with high total edge weight. To align PPI networks within the same species, we construct a similar graph, in which the edge similarities are quantified based on Pfam and Gene Ontology annotation terms (see next subsection). We then find the optimal alignment using a search algorithm where the graph is pruned greedily to form a topologically correct alignment of the two networks and the maximal alignment is returned at the end.

The alignment graph $\Psi(G_A, G_B) = (U_\Psi, F_\Psi)$ is generated as follows: every vertex $u_i \in U_\Psi$ corresponds to a pair of interactions $(e_A, e_B)$ where $e_A \in E_A$ and $e_B \in E_B$, such that $U_\Psi = E_A \times E_B$. In essence, the nodes of the alignment graph correspond to every possible alignment of the interactions of the two networks (In Figure ?? these are the nodes displayed on the right hand side). A pair of nodes $(u_i, u_j)$ where $u_i, u_j \in U_\Psi$ is connected by an edge $f_{ij} \in F_\Psi$ in the alignment graphs if the alignment follows one of the following three types of connectivities:

1. The two pairs of aligned edges share a protein in both graphs (Extension)
2. The two pairs of aligned edges share a protein in one graph, but in the other graph they are separated by one hop (Insertion/Deletion)

3. The two pairs of aligned edges are separated by one hop in both networks (Jump)

Figure 2: Alignment graph nodes. The nodes that are generated while building the alignment graph are depicted. (A) an extension, where two edges follow each other in the two networks. (B) while one network has two connected edges, the other has an additional edge in the other network (insertion/deletion). (C) the two pairs of edges are separated with one interaction on both networks (a jump).
After the network is thusly constructed, the edges are weighted with a combination of the similarity score for each aligned interaction and a penalty for insertions/deletions and jumps, biasing the graph towards fewer topological inconsistencies. We next describe the scoring of nodes and edges in detail.

### 2.2.2 Alignment Scoring

In the alignment graph, the weight of an edge consists of two components: (i) the weights of nodes that are connected by the edge and (ii) the relative topology of the corresponding interactions in the two subnetworks. Since each node in the alignment graph represents a possible mapping of two interactions, each from one of the subnetworks being compared, the weight of a node reflects the functional similarity between these interactions. We now discuss how we quantify the functional similarity between two interactions. Subsequently, we discuss how topological inconsistencies in the mapping are penalized.

### 2.2.3 Scoring Nodes of the Alignment Graph

An important factor that will result in proper alignments is to develop a biologically accurate quantitative measure of the similarity between interactions in the two subnetworks. Here, we define such a measure based on sequence similarities identified as in the protein domain family database Pfam [? ] and through Gene Ontology annotations [? ]. However, instead of aligning proteins to proteins (node-node alignments), we align interactions (edge-edge alignments). A similar approach was proposed earlier for cross-species alignments of networks [? ]. This approach allows us to focus more
Figure 3: Algorithm for within species network alignment.

1. Construct a node for every combination of interactions of networks A and B. Note that alignment nodes are not transitive, but in this figure we only consider one direction of alignment. Weight the nodes with the “alignment weight,” which is proportional to the number of shared Pfam families and shared gene ontology terms between corresponding proteins in an aligned interaction.

2. Connect the nodes by edges corresponding to extensions, insertions/deletions, and jumps. Note that only extensions are shown in this figure, and also, reverse alignments are not included (hence the lone node).

3-5. Find the “modified” minimum spanning tree, corresponding to an alignment, eliminating nodes with each step. Repeat, until an alignment has been generated from each node, and 6. pick the alignment with the minimum weight.

on functional associations across networks (Figure ??).
2.2.4 Scoring Sequence Similarity.

Pfam provides a public database of protein domains and their associated proteins as discovered by sequence similarity analysis. We incorporate Pfam into our algorithm under the conjecture that alignment between two edges is more plausible if the corresponding proteins in the alignment share a high number of Pfam domains. Hence, utilization of Pfam allows us to incorporate sequence similarity into our alignment score, highlighting the functional viewpoint of these sequences.

Given two interactions \( e_A = (v_{A,i}, v_{A,j}), e_A \in E_A \) and \( e_B = (v_{B,k}, v_{B,l}), e_B \in E_B \), we compute the sequence-based similarity between these two interactions as follows:

\[
\text{s}_{pfam}(e_A, e_B) = \frac{|\text{PF}(v_{A,i}) \cap \text{PF}(v_{B,k})|}{|\text{PF}(v_{A,i}) \cup \text{PF}(v_{B,k})|} \cdot \frac{|\text{PF}(v_{A,j}) \cap \text{PF}(v_{B,l})|}{|\text{PF}(v_{A,j}) \cup \text{PF}(v_{B,l})|}
\]

Where \( \text{PF}(v) \) represents the set of Pfam domains corresponding to the given protein \( v \).

2.2.5 Scoring Functional Similarity.

The functional relationship between two interacting proteins can be assessed using biological process (BP) Gene Ontology annotations, where operations or sets of molecular events are described. Molecular function annotations from Gene Ontology describe elemental activities of gene products at the molecular level, and cellular component annotations lists parts of a cell the gene product is active. Hence, in this context, only biological processes can possibly describe the underlying processes and can help capture the nature of interactions. For example, if an interaction consists of a protein
with the GO term kinase activity (GO: 0016301) and another protein with the GO
term protein kinase binding (GO: 0019901), when aligning two interactions, it follows
that both interactions will have a similar relationship. Therefore, given an interac-
tion alignment, corresponding proteins should have similar GO terms, indicating that
they have the same role in their respective interactions. Indeed, in recent years, many
studies have demonstrated that interactions in signaling and regulatory networks can
be abstracted as interactions between different functional categories, which are used
recurrently in various biological processes [? ? ]. We compute the GO-based similar-
ity between two proteins utilizing a semantic measure where GO terms of each gene
pair are assessed by their specificity and depth on the GO tree [? ? ]. This measure
quantifies the information content of the shared functional annotations of the two
proteins being compared and is defined as:

\[ IC(x, y) = -\log_2 \frac{|G_{\Lambda(S_i, S_j)}|}{|G_r|} \]

Here, \( G_{\Lambda(S_i, S_j)} \) is the set of all genes that are described by the minimum common
ancestor set of the genes \( S_i \) and \( S_j \), and \( G_r \) is the set of all genes in GO. If the two
proteins have highly specific common ancestors, then they are scored higher than if
their common ancestors describe many other genes.

Using this measure, we quantify the functional similarity between two interactions
\( e_A = (v_{A_i}, v_{A_j}) \), \( e_A \in E_A \) and \( e_B = (v_{B_k}, v_{B_l}) \) as follows:

\[ s_{GO}(e_A, e_B) = IC(v_{A_i}, v_{A_j}) \times IC(v_{B_k}, v_{B_l}) \]
2.2.6 Overall Similarity Between Two Interactions.

The overall similarity score, and hence weight of alignment, is defined as:

\[ w(e_A, e_B) = s_{pfam}(e_A, e_B) + s_{GO}(e_A, e_B) \]

2.2.7 Scoring the Alignment Graph and Normalization of Scores

The alignment score is essentially the cost of the tree, \( t_{AB} \), generated as a result of aligning networks \( G_A(E_A, V_A) \) and \( G_B(E_B, V_B) \). This cost, \( \text{cost}(t_{AB}) \), is simply calculated by summing up the weight of nodes on the tree.

\[ \text{cost}(t_{AB}) = \sum_{(e_A, e_B) \in t_{AB}} w(e_A, e_B) \]

The networks that are considered for alignment have varying number of protein-protein interactions, and the network alignments essentially find similar interactions across two networks. To assess the network alignments, after a score is calculated, a normalized alignment score is calculated. The normalization is done with respect to the maximum possible alignment that is possible. The normalized score, \( \hat{s}_{AB} \), is:

\[ \hat{s}_{AB} = \frac{k(\min(|E_A|, |E_B|) - 1)}{\text{cost}(t_{AB})} \]

Where \( k \) is the constant value that represents the maximum similarity score an edge can be assigned. In case of a perfect alignment, e.g. self-alignment of a network, \( \hat{s}_{AB} = 1 \), as the tree will return one less edge than the size of the smallest network (if there is a cycle in the graph). In case of no similarity, the score \( t_{AB} \) will be infinite, which will correspond to a normalized score \( \hat{s}_{AB} = 0 \). Based on the normalized scores,
unsupervised hierarchical clustering of the networks was done in R using the hclust function.

2.2.8 Finding the Optimal Alignment

Successful traversal of the alignment graph can discover an optimal alignment. However, due to the three types of connectivity that are possible, the number of edges in the graph can become very large and the traversal can be cumbersome. We here propose a novel method to approximate the global maximum of the alignment on the graph. In an alignment graph, $\Psi$, a single edge in $G_A$ is mapped to many possible edges in $G_B$, therefore a traditional minimum spanning tree would not make sense, as it would not correspond to a one-to-one alignment of an edge of $G_A$ with an edge in $G_B$. Therefore, when a node corresponding to the alignment $(e_A, e_B)$ is added to the network, all other nodes containing possible alignments of $e_A$ or $e_B$ are removed from the graph. This dramatically reduces the computational time of the algorithm, and it removes the redundancy inherent of the alignment graph. The alignment graph is calculated starting from every node, and the modified network with the least weight is picked as the optimum alignment. In essence, the traversal of the alignment graph generates a tree that is representative of the alignment. A minimum spanning tree would not work, because it would inevitably pick nodes that violate the one-to-one constraint on $A$.

Starting at a random node, the algorithm proceeds as follows:
Initialize:

\[ E_{\text{traversed}} = \emptyset \]

\[ V_{\text{traversed}} = \{ x \}, \text{ where } x = (e_{A_x}, e_{B_x}) \text{ is a random node from } A. \]

\[ V_{\text{available}} = V_\Psi \setminus \{(e_{A_x}, *), (*, e_{A_x}), (e_{B_x}, *), (*, e_{B_x})\} \]

Repeat until \( V_{\text{available}} = \emptyset \)

Choose \( (x, y) \),

where \( y = (e_{A_y}, e_{B_y}) \in V_{\text{available}} \) with minimum weight in \( \Psi \),

\[ x \in V_{\text{traversed}} \text{ and } y \notin V_{\text{traversed}}. \]

\[ V_{\text{traversed}} = V_{\text{traversed}} \cup y \]

\[ E_{\text{traversed}} = E_{\text{traversed}} \cup (x, y) \]

\[ V_{\text{available}} = V_{\text{available}} \setminus \{(e_{A_y}, *), (*, e_{A_y}), (e_{B_y}, *), (*, e_{B_y})\} \]

The algorithm is repeated for every node in the alignment graph. The traversal with the minimum weight is the one that corresponds to the best alignment of the graphs.

### 2.3 Gene Ontology-Based Classification of Genes

We have utilized the same semantic ontology similarity measure to assess the functional similarities of the genes [? ]. The similarity scores of driver genes based on gene ontologies, \( s'_{GO} \), are utilized to generate an unsupervised clustering of the genes in R using the hclust function, which performs hierarchical clustering.

\[ s'_{GO}(v_i, v_j) = IC(v_i, v_j) \]
2.4 Sequence (Protein Families) Based Classification of Genes

We have utilized a measure that is similar to the alignment scoring to assess the sequence-based similarities of the cancer driver genes. Shared protein family domains of genes obtained from PFAM [?] were used to generate a scoring matrix, where given two genes \( v_i, v_j \), we compute the sequence-based similarity between these two genes as follows:

\[
s'_{pfam}(v_i, v_j) = \frac{|PF(v_i) \cap PF(v_j)|}{|PF(v_i) \cup PF(v_j)|}
\]

where \( PF(v) \) represents the set of Pfam domains corresponding to the given gene \( v \). Unsupervised clustering of the genes was generated using hierarchical clustering.

2.5 Sequence and Gene Ontology Based Classification of Genes

To generate a comparable measure at the gene level, we have also utilized a weight function that is similar to the weight function used in the alignment algorithm. However, since only driver genes are considered, and not networks, a modified scoring function \( w' \) is used.

\[
w'(v_i, v_j) = s'_{pfam}(v_i, v_j) + s'_{GO}(v_i, v_j)
\]
2.6 Somatic mutation profiles of tumor samples and statistical evaluation

We have acquired mutation data for the driver genes that correspond to the generated networks. The CAN-gene identification dataset and The Cancer Genome Atlas publicly available somatic mutation data for have been acquired for the driver genes that correspond to the generated networks. The CAN-gene identification dataset [?] and The Cancer Genome Atlas publicly available somatic mutations for colon and rectum adenocarcinomas [?] were merged. In this mutational dataset, the 94 samples mutation profile for the driver genes of interest are represented in a binary fashion, i.e. a positive value is assigned to a gene if there is at least one non-synonymous mutation on that gene. The Mantel test is utilized to test any significant correlation between the network alignments and mutation profiles of patient samples. The normalized network score is converted to distances and the mutation profiles across patients are used for each gene to calculate a distance. Mantel test as it is implemented in R package ade4 is used with ten thousand permutations to generate a p-value showing the likelihood of correlation.
3 Results and Discussion

In this work, we investigate frequently mutated genes that are part of the neoplastic process, also known as cancer driver genes. Although there are various genes that have been identified as key driver genes, most of these genes are mutated in lower frequencies than expected in sequenced tumors. We hypothesize that functionality of cancer driver genes can be better assessed by looking at their functional role within the tumorigenesis process. In this work, we have started from driver genes identified for colorectal cancer, tested our hypothesis through a novel computational approach, where networks connecting these driver genes are used to assess these genes (Figure ??).

3.1 Cancer driver genes and their synergistic activities

In recent years various cancer genomes and exomes were sequenced with the new technologies that have become available. These studies have resulted in identification of key driver genes for various cancers such as colorectal and breast cancer [? ], ovarian cancer [? ] and glioblastoma multiforme [? ]. In this study, we focus on colorectal cancer driver genes as identified in [? ].

Colorectal tumorigenesis starts mostly with a mutation in $APC$ gene, and tumor progression results from mutations in the other genes as observed previously [? ? ]. These sequences of molecular and genetic events lead to transformation of adenomatous polyps to malignancies. In addition to mutations, various genetic and epigenetic
Figure 4: Example alignment of two CAN-gene networks (SMAD3 (left) vs. SMAD4 (right) depicted). CAN-genes are colored grey, and dashed (red) edges show aligned nodes (hence edges) that returned the best score. This alignment has a normalized score of 0.82.

Events such as abnormal DNA methylation can silence tumor suppressor genes or activate oncogenes. In this study, we focus on somatic mutations that would compromise the genetic balance and ultimately lead to tumorigenesis. Utilizing an earlier network discovery approach, we gather 22 CAN-gene signaling pathway networks that represent likely synergistic relationships between these CAN-genes and APC gene.

These networks reflect significant signaling pathway events, reflecting the content of the gene expression data used. In other words, utilization of a different gene expression data might change the set of driver gene networks we gather. Moreover, we are able to extend the number of discovered networks via utilizing estimated interactions as previously suggested [? ?] to overcome with missing links in the PPI network. However, we choose to use only known interactions collected from public PPI databases [? ?] and these 22 networks are utilized for further evaluation.
3.2 Finding similarities among synergistic networks

We employ a novel network alignment tool catered towards aligning networks that originate from the same organism. The alignments resulted in normalized similarity scores among the 22 networks representing CAN-genes in colorectal cancer (e.g. $SMAD3$-$SMAD4$ alignment as in Figure ???). Figure ?? shows the histogram generated based on pairwise alignments of the driver gene networks. The normalized similarity score ranges between 0 and 1, and the distribution of the scores shows that while most of the networks are not similar to each other, there are smaller groups of networks that show high similarity, i.e. have high normalized alignment score.

Next, hierarchical clustering is conducted with these similarity measures, and a dendogram showing close relationships of these networks are shown (Figure ??). As expected, networks that connect genes that belong to the same family such as $SMAD2$, $SMAD3$ and $SMAD4$, or $EPHA3$ and $EPHB6$ cluster closely. $SMAD2$, $SMAD3$, and $SMAD4$ networks also show increased similarity to other driver genes based on their highly similar interacting partners. For instance, $HIST1HB1$ and $SMAD2$ tumor suppressor genes were previously classified as chromatin remodeling/transcription genes, and showed close relationship in our unsupervised classification [? ? ].

3.3 Clustering Cancer Driver Genes

In order to depict the utility of the network comparison approach, we also employed a similar clustering of the driver genes by utilizing their gene ontology terms and/or Pfam domains. While sequence only and ontology only comparisons generated very
Figure 5: Histogram of normalized scores are plotted. All networks are compared pairwise, and the frequencies are calculated. While most networks are not similar to each other, there is a limited number that is highly similar, i.e. have a normalized score greater than 0.8.

Different results than the network approach (See Appendix), we have combined the two measures to create a better comparison of gene-based, and network-based approaches. The clustering of the cancer driver genes shown in Figure ?? and ?? clearly show that network-based approach can capture a different clustering of these genes.

For instance, TGF$\beta$ signaling associated SMADs are clustered with P2X$_7$ receptor (P2RX7) in the leftmost cluster in Figure ??, are not clustered together in the gene-based approach. These genes are shown to have a relation where activation of P2X$_7$ receptors enhanced TGF-β signaling which is upstream of SMADs [? ]. Another
Figure 6: Classification of CAN-gene networks and observed somatic mutations in various tumors. The hierarchical clustering of CAN-gene networks are shown at top. The heatmap below depicts the mutations observed in 94 patient samples from two independent studies [? ? ] Each column corresponds to a gene. Mutational profiles of samples for these driver genes are given in rows.
Figure 7: Hierarchical clustering of CAN-genes using a distance measure that incorporates semantic similarity [? ] and sequence information extracted from Pfam [? ] is shown.

An example of the differences is cytoskeleton remodeling genes LMO7 and SYNE1 [? ]. While these were within the same subtree in our network-based clustering, the gene only approaches placed these two genes in different subtrees. Interestingly, if only sequence-based information is used, these genes would appear within close proximity (Figure A??). However, the gene ontology only clustering has placed them in different subtrees (Figure A??). Note that gene ontology includes expert annotation of gene/protein functions and is not dependent on sequence information. This clearly shows that while our network-based approach can capture functional properties of CAN-genes, sequence and/or ontology-based approaches are more limited in this assessment.
3.4 Testing synergistic activities by comparing somatic mutation profiles of tumor samples

In order to test our hypothesis, we have collected somatic mutations identified from colorectal cancer samples collected from humans [? ? ]. To improve the rigor of our study, we have combined data from two studies that followed different approaches. Overall, the 94 colorectal cancer tumor samples are tabulated for their mutation status at these 22 genes (Figure ??). We utilize Mantel test [? ] as a statistical test. This method looks for significant correlation between two distance matrices. The CAN-gene network similarities acquired from the network alignments and mutation profiles of patient samples are used for this purpose.

We calculate distances between driver gene mutation profiles based on their commonality in the tumor samples. For instance, observing a mutation in the same tumor would increase the similarity of these genes, or decrease the distance between them in this context. We gather such distance measures based on all 94 samples. The same distance measure used for hierarchical clustering put to work for CAN-gene networks.

As a result, the distance matrices (not shown) had no correlation ($p \geq 0.4$). This result rejects that the two datasets, one from the alignment of the networks and one from the somatic mutations have no correlation. If there was a correlation, this would mean that by utilizing the mutation profiles only, we could generate a similar clustering of the genes without using driver gene networks. This result rejects this claim and hence suggests that the mutations are mostly mutually exclusive for functionally similar CAN-genes.
3.5 Validation of results with cell line epistasis experiments

We validated our approach via a cell line experiment where we have picked two genes that are closely linked within the clustering of CAN-gene networks and performed an epistasis analysis. Epistasis refers to genetic interactions that the mutation of one gene masks the phenotypic effects of a mutation at another locus. Our hypothesis is that given the CAN-gene clustering if two networks are closely similar, the suppression of one gene followed by the second gene should not create a big shift in the phenotype. This is justified by the functional similarity of the two genes, and it would be redundant to lose the same function twice.

We have performed siRNA knockdown experiments in a patient-derived cell (donated by Markowitz Lab, CWRU). The cell line did not have a mutation in \textit{EPHA3} or \textit{EPHB6} — the two genes we have selected. \textit{EPHA3} and \textit{EPHB6} were knockdown individually and in sequence (both ways) to observe the phenotypical change (Figure ??). The knockdown samples in triplicate are then analyzed with Illumina HumanHT12 microarray chip (data pending on Gene Expression Omnibus). The results were quantile normalized and gene level expression values were further analyzed. All data was utilized to generate a multidimensional scaling plot to observe changes in the cell lines (Figure ??).

Based on our results, we observe that the single knockdown experiments have generated similar phenotypic changes. The second knockdown did not have additive or synergistic effect. Hence the two similar CAN-genes identified by the network analysis validates our approach that CAN-genes with similar functions are not necessarily
Figure 8: The experimental design for validations is shown. The two genes are knockdown one at a time, and also in tandem. All cell lines are hybridized and analyzed with mRNA gene expression chips.

mutations are mutually exclusive for functionally similar CAN-genes identified by our framework.
Figure 9: **Multidimensional Scaling of knockdown experiments in triplicate.**

The multidimensional scaling plot as a result of microarray gene expression profiles of samples are shown. WT shows Wild type CRC Cells, A3 shows knockdown of *EPHA3* gene (siEPHA3), B6 shows knockdown of *EPHB6* gene (siEPHB6) and AB shows knockdown of both genes (siEPHA3 first and siEPHB6 later or siEPHB6 first and siEPHA3 later). The phenotypes of single knockdowns shift cells drastically in a similar direction. However the additional knockdown of the second gene after the first one does not create incremental burden in the phenotype, hence siEPHA3-siEPHB6 knockdowns are still close to siEPHA3 only or siEPHB6 only knockdowns of samples.
4 Conclusion

In this study, we introduce a network alignment based approach to analyze signaling networks in carcinogenesis, and we demonstrate this method using colorectal cancer driver gene signaling networks. While recent advancement in sequencing technologies allow us to identify these changes, little is known as to how these alterations come together to manifest themselves as malignancies. By utilizing a novel network alignment tool we were able to study how functional similarity between driver gene signaling networks can be used to predict functional similarities among driver genes in carcinogenesis.

By repeating the clustering experiment with only functional annotations and sequence information, we found that the synergistic networks represent a better way of elucidating these relationships, and it is likely that similar driver genes are grouped together. In other words, given a group of similar driver genes, these work in similar synergistic activities with the \textit{APC} gene in the tumorigenic progress.

By testing against an independent set of mutation data, we found that given a specific tumor, it is unlikely that similar driver genes are mutated together. In other words, given a group of similar driver genes, a single mutation in this group is sufficient to continue the tumorigenic progress.

We further validated our computational results with cell line experiments, where we demonstrated that the clustering of CAN-Genes can predict the functional role of these genes in cancer phenotypes. This work represents a novel approach to studying
cancer driver gene networks: analyzing the networks in terms of their functional similarity, using graph alignment algorithms.

In future work, this framework for analysis of signaling pathway networks can be applied to analyze larger numbers of driver genes, resulting in a more complete timeline of likely events of tumorigenesis. Here, we have concluded a small study for colorectal cancer; this can be applied to other cancers that benefit from recent large-scale cancer genome projects. These methods will be important for understanding the relationships between cancer signaling networks, resulting in improved diagnosis and characterization of tumors and finding appropriate and efficient regimens for treatments.
A Appendix

A.1 Gene Ontology Based Classification of Cancer Driver Genes

The clustering of the cancer driver genes shown in Figure ?? and ?? show very clear differences. That network-based approach can capture a different clustering of these genes. For instance, \textit{EPHA3} and \textit{EPHB6} which have a close functional bimodality are put together in the context of CRC networks, whereas the gene ontologies does not show the same similarity for these two genes. Another example is \textit{SMAD2}, \textit{SMAD3} and \textit{SMAD4} genes where in the latter clustering these are not observed to be in the same subcluster. Moreover, the hierarchical tree also shows a more level clustering rather than grouping the genes. This result suggests that the networks captured functional similarities of CAN-genes where better than the annotation only based approach.

A.2 Sequence Based Classification of Cancer Driver Genes

We have also utilized Pfam domains [?] as a measure of similarity at the sequence level, and generated a scoring matrix that compared the driver genes that we have studied. The subsequent clustering is shown in Figure ??, where most genes are grouped together, since most genes had similar distances to others. While some protein families are captured quite well, the dendogram does not present much information about the majority of the driver genes.
Figure 10: Classification of CAN-genes based on their gene ontology annotations using hierarchal clustering are shown. The functional similarities of genes are assessed by the semantic similarity of the ontology terms that were reported in Gene Ontology database. The red lines show expected groupings of the terms that were placed in different subtrees in the unsupervised method.
Figure 11: Classification of CAN-genes based on their Pfam families using hierarchal clustering are shown. The sequence similarities of genes are assessed by the shared terms in the Pfam database.
References


[7] Laura D Wood, D Williams Parsons, Siân Jones, Jimmy Lin, Tobias Sjöblom,


