Universiteit Leiden
The Netherlands

Bioinformatics

Investigating the Hallmarks of Cancer
Using Protein-Protein Interaction Networks

Laurens Engwegen

Supervisor:
Dr. K.J. Wolstencroft

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

16/08/2020

Abstract

The hallmarks of cancer comprise ten capabilities acquired by cancer cells. This generalisation of the functionalities of cancer cells has enabled classification of cancer genes, regarding their role in the disease, by means of hallmark annotations. In this thesis, a protein-protein interaction network of cancer gene products was constructed. Since a part of these gene products had a hallmark annotation, the hallmarks could be investigated on a network level. Using network topology and the Gene Ontology, connections between hallmarks were discovered. In addition, annotated cancer gene products were used to predict hallmark annotations for not yet annotated cancer gene products on basis of network features and semantic similarity between those gene products.

Contents

1 Introduction

Cancer is an extremely complex disease and is still one of the main causes of death [1]. Over the past decades a lot of cancer research has been done, which has led to a great amount of discoveries. *TP53* and *PTEN* have been identified as guardians of the genome and key players in the disease, among others [2][3]. Nonetheless, the complexity of cancer is mainly caused by the fact that there are a lot of other genes involved, besides the key players. Furthermore, an enormous amount of different forms of cancer exist, dependent on a lot of different factors.

The use of novel technologies have caused progression in cancer research, prevention and treatment. The most recent example is the analysis of 2,658 whole-cancer genomes in the Pan-Cancer Analysis of Whole Genomes [4]. Due to the complexity of the disease and the advances in research, many genes are now linked to the disease. For a large part of the discovered genes, however, there is a lack of functional characterisation. Effective detection and treatment of cancer is obviously dependent on the identification of the functions of the cancer genes. For example, Zhan *et al.* showed the possible impact of using CRISPR/Cas9 in cancer reasearch and give an overview of the first trials in which this technique is used as therapy against cancer [5].

As a result of the extensive research to cancer cells and their differences with healthy cells, a tremendous amount of information has been obtained. Subsequently, the discovered information about cancer is widely spread across the literature. In addition with the fact that the complexity of the disease is overwhelming, this makes cancer research challenging. Several efforts have been done to generalise and simplify the disease and to centralise the available information. Already in 2000, the complexities of the disease were made more understandable by combining them into a small number of general characteristics: the Hallmarks of Cancer [6].

1.1 The Hallmarks of Cancer

The multi-step process of cancer development is hard to understand on a molecular level. However, several characteristics that normal cells need to acquire to become cancer cells are known, such as their ability to grow uncontrollably and to spread across the body. Hanahan and Weinberg have used existing knowledge to identify such characteristics and initially proposed six hallmarks of cancer [6]. As more research was done and more information about the behaviour of cancer cells on a molecular level had come to light, they added four more hallmarks in 2011 [7]. The set of the ten hallmarks of cancer generalised the disease and is widely used in research. The hallmarks are visualised in Figure 1 and are briefly described. (The hallmarks have been enumerated in this thesis as shown below. This enumeration will be used to refer to the hallmarks.)

(1) Sustaining proliferative signaling

Cancer cells are able to sustain chronic proliferation. This is achieved by deregulating the signals that promote or suppress growth. These signals are often growth factors that bind to cell surface receptors that contain tyrosine kinase domains. One of the most important signaling pathways is the mitogen-activated protein kinase (MAPK) pathway [8], which has been associated with different types of cancer. As a consequence of the dergulation of pathways like this, cancer cells are able to

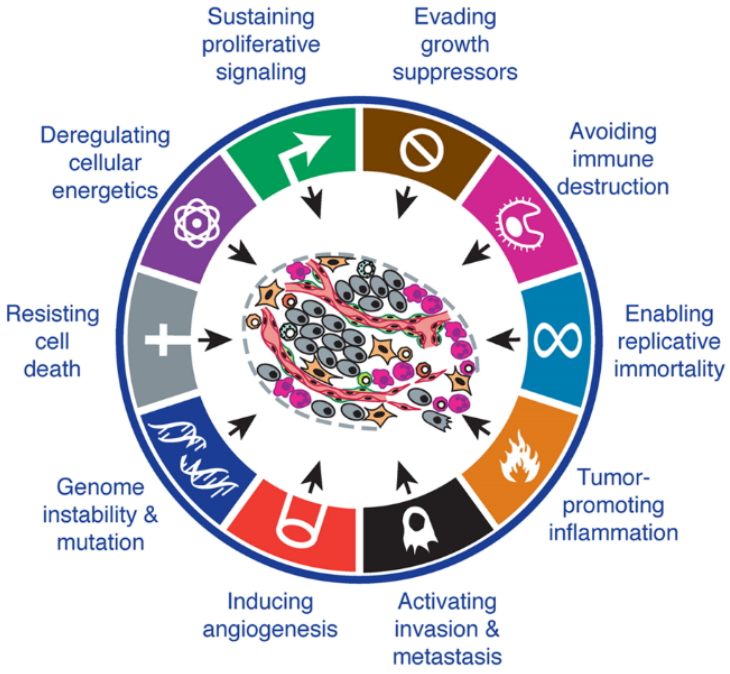

Figure 1: The ten hallmarks of cancer, a generalisation of the disease on basis of the acquired capabilities of cancer cells. (From: *Hallmarks of Cancer: The Next Generation*, https://doi.org/10.1016/j.cell.2011.02.013.)

grow and differentiate faster than normal cells.

(2) Evading growth suppressors

Besides the signals that promote growth, cells can also receive signals that suppress growth. Cancer cells have acquired the capability of becoming insensitive to those signals. As for the first hallmark, this allows for rapid growth and differentiation of cancer cells.
The first two hallmarks seem to be related in terms of the ability of cells to communicate in order to create a micro-environment in which cells help each other to grow cancerous tissue.

(3) Avoiding immune destruction

A healthy immune system is able to detect abnormal cells and destroy them, often by apoptosis. Cancer cells have, however, acquired the capability of evading this destruction by the immune system.

(4) Enabling replicative immortality

Another barrier that cancer cells have to overcome is the fact that cells normally have a limited life cycle. This is caused by the erosion of telomeres on the chromosomes during the lifetime of a cell. Cancer cells show a significantly high expression of telomerase, the DNA polymerase that is able to

repair telomeres, which allows the cells to live longer and, thus, enable them to replicate more than normal cells.

(5) Tumour promoting inflammation

This is one of the two enabling hallmarks. Inflammation of cancer tissue is known to contribute to the development of tumours by supplying molecules to tumours, including growth factors, survival factors and angiogenic factors among others. As a side note, the immune system is able to cause inflammation of cancer cells as a response to the disease and thereby indirectly contribute to the development of larger cancer tissue.

(6) Activating invasion & metastasis

Arguably one of the most well-known and threatening capabilities of cancer is its ability to spread through the body. The development of multiple malignant tumours at a distance from its origin is known as metastasis. The invasion of organs in combination with metastasis is known to make tumours malignant. This process is very complex and consumes a lot of energy.

(7) Inducing angiogenesis

In order to acquire and maintain most of the hallmarks of cancer, the cancer cells require a lot of oxygen and nutrients. Furthermore, the cells produce a lot of waste that has to be evacuated. Cancer cells, thus, need to create and/or expand the vasculature system from which oxygen and nutrients can be taken up and to which waste can be evacuated.

(8) Genome instability & mutations

This is the second enabling hallmark. Especially because of the lack of DNA damage repair, genome instability and mutations arise in cancer cells. This is an enabling hallmark because a higher rate of mutations results in a higher chance to acquire other hallmarks.

(9) Resisting cell death

Abnormal processes such as uncontrolled growth and DNA damage can be detected by cells. This will trigger signaling pathways that cause the cell to die. However, cancer cells have the ability to deregulate these signals to survive.

(10) Deregulating cellular energetics

Many of the hallmarks require a lot of energy from the cell, especially accelerated growth. Furthermore, cancer cells seem to be inefficient in their energy metabolism. This causes the need of deregulating cellular energetics. In this way, this hallmark might be looked at as yet another enabling hallmark, although it wasn't introduced as such.

As mentioned, the hallmarks of cancer are used in research. When the function of a cancer gene is known, this gene could be annotated with one or more hallmarks. There are still a lot of genes that have been shown to play a role in the disease, but their specific function and the hallmark to

which this function contributes are still unknown. The hallmarks of cancer give the opportunity to investigate cancer on a broader level and answer questions about the mutli-step process of cancer development.

1.2 COSMIC's Cancer Gene Census

Besides the efforts of generalising and simplifying the mechanics of cancer cells to better understand the disease, there are also initiatives to combine available information about cancer into databases. The Catalogue of Somatic Mutations in Cancer (COSMIC) [9] is such a source of information about cancer. It provides several databases with information related to human cancers. One of those databases is the Cancer Gene Census (CGC) [10], an ongoing effort to catalogue the genes of which mutations have been implicated in cancer. This database is manually curated by experts and includes a hallmark annotation for a part of the genes.
When functioning normally, genes in this database either have promoting or suppressing functions on hallmarks of cancer. When a mutation of a promoting gene causes overexpression or *gain-of-function*, this will contribute to the emergence of the corresponding hallmark, where underexpression or *loss-of-function* of a suppressing gene will also cause a contribution of this gene to the hallmark with which its function corresponds. In this way, the genes in the CGC are annotated to promote or suppress one or more of the hallmarks. In this thesis, the promoting and suppressing annotations will be taken together when investigating the hallmarks, since it will not be investigated whether not annotated cancer genes have a suppressing or promoting effect on the hallmarks. Rather, the genes and their hallmark annotations in the CGC will be used to investigate the hallmarks of cancer by means of the relations between those cancer genes.

1.3 Protein-Protein Interaction Network

Proteins perform a lot of cellular functions, where each protein has its own role. However, isolated proteins don't function [11]. Their interactions with other proteins and molecules initiate biological processes. In this way, the function of a protein can be expressed as its interactions with other proteins and molecules. As a result of advances in research regarding high-throughput interaction detection, such as the yeast two-hybrid system [12] among others, an enormous amount of protein-protein interactions has been discovered. Protein-protein interaction (PPI) data has enabled the possibility to create networks of proteins to use for the investigation of cellular behaviour.

In a PPI network nodes represent proteins and an edge between two nodes represents a possible interaction between the two connected proteins. In this way a PPI network that includes many proteins that are related to biological functions, pathways or diseases could be constructed. Using a PPI network, cell function could be investigated on a large scale in which a lot of information is included, in contrast to traditional single-gene research.

Network topology shows properties of the network and can be used to identify important nodes or relevant sub-structures of a network. The degree k of a node is defined as the amount of edges that are connected to that node. When a node has a high degree, this indicates that the node plays an important role in the network. There are other features regarding the topology of a network that can give an indication about the properties of the network and its nodes. The closeness centrality

of a node is based on its average distance to all other nodes in the network. When a protein has a high closeness in the network, it is "close" to all other proteins in the network, i.e. the path lengths (amount of edges that lay between two nodes) to all other nodes is small. The betweenness centrality of a node is based on the number of times a node occurs on a shortest path connecting two other nodes. Thereby, this gives an indication about the importance of this protein in the network. Namely, when its betweenness is high, the protein connects a lot of other nodes in the network and, thus, could be considered important.
Biological networks are known to be scale-free [13], i.e. their degree distribution approximates a power law. This means that there are many nodes with a small degree and a few nodes with a very high degree. The latter, thus, represent proteins that connect many other proteins and could thereby be considered the most important proteins in the network that is studied. Furthermore, these proteins are often central in the network, resulting in a relatively high closeness and betweenness of these proteins.

PPI networs have shown to be useful for protein function prediction [14][15]. Furthermore, many complex diseases are caused by complex interactions between genes (and/or gene products), which implies that PPI networks could be of use in studying these diseases on a (multi)cellular level. In cancer research, PPI networks have shown to be useful for analysis of the disease [16] and for revealing properties of cancer-related proteins [17].
Complex networks often show a hierarchical topology, which could be used to identify sub-structures [18]. Highly interconnected nodes in a network could be identified as clusters (or modules). The proteins in a cluster are relatively densely connected and refer to a group of proteins that are physically or functionally linked. In addition, cliques often form the core of a cluster. Cliques are maximally connected groups nodes, where, thus, every node is linked to all other nodes in the clique. As a result, this indicates that proteins in a clique are very related to each other.

1.4 Research Question

The introduction of the hallmarks of cancer in combination with the use of a PPI network has created the opportunity to do cancer research on a (multi)cellular level. In this thesis, the aim is to investigate the interconnections between the cancer gene products in the CGC to discover how hallmarks of cancer interact with each other and to use the interconnections to predict hallmark annotation for gene products that have not yet been annotated by COSMIC. The research questions are formulated as follows:

1. How do the hallmarks of cancer interact with each other over the network? Do hallmarks occur in overlapping groups or are the hallmark annotations distributed across the network?

2. Can we use network features to predict hallmark annotations for cancer gene products that have not yet been annotated?

First of all, it is pointed out how the network of cancer genes is constructed in Section 2. In this Section, the network analysis that is performed and the bioinformatics tools that are used will also be explained. In Section 3 the results are described after which they will be linked to the goals of this thesis in Section 4.

2 Methods

2.1 Cytoscape and STRING Database

The rapidly accumulating amount of PPI data is available in different databases. In general, this data is not only obtained through lab experiments, but also by creating models to predict interactions. To integrate and visualise this data with networks, different software could be used. Cytoscape is a software environment used by bioinformaticians to visualise and analyse biomolecular interaction networks [19]. In Cytoscape various plug-ins could be used for the analysis of networks, allowing an efficient way to investigate biological networks. One of the popular plug-ins for Cytoscape is the StringApp [20]. This plug-in provides the user with an easy way to import PPI data from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database [21]. The STRING database collects and integrates all the publicly available physical and functional protein-protein interaction information and extends this information using computational predictions. The database currently covers over 24 million proteins and 3 billion interactions from more than 5000 different organisms. The interactions are obtained from five main sources: (1) genomic context predictions, (2) high-throughput lab experiments, (3) (conserved) co-expression, (4) automated textmining and (5) previous knowledge in databases. The confidence of the predicted interactions often differ from the known interactions. Hereby, confidence scores are given for the interactions. These confidence scores are derived from measuring the quality of the predictions on a reference dataset (consisting of known interactions).

2.2 Creating the Protein-Protein Interaction Network

Cytoscape (version 3.8.0) was used in combination with the StringApp (version 1.5.1) to create a PPI network of the products of the genes in the Cancer Gene Census. The StringApp takes a list of gene product identifiers as input and outputs all known and predicted interactions that can occur between the gene products in the specified organism (Homo Sapiens in this case). A confidence score cut-off of 0.4 was chosen for the interactions to include only the interactions with a medium or higher confidence. In this way the obtained network consists of enough edges to apply analysis, while the confidence of the interactions/edges is maintained.

For some genes multiple possible gene product matches were found in the STRING database. Those genes and their possible matches are shown in Table 1. The genomic locations of the genes were compared with that of their possible matches to determine whether one of the matches is an alias for the input gene identifier. In this way it was determined that this was certainly the case for the underlined terms in Table 1, which were thereby imported. For the TRA gene, none of the possibilities had a matching genomic location.

The list of 722 genes (723 in the census minus the TRA gene) was input to the STRING database using StringApp, after which the gene products were imported into Cytoscape together with the interactions between those gene products. In this way a PPI network of the genes in the Cancer Gene Census was generated and visualized in Cytoscape. The obtained PPI network consisted of only 716 gene products instead of 722.
As mentioned, STRING takes a list of genes as input and converts this list into gene products. It is

Table 1: Genes with multiple possible matches in the STRING database. The underlined terms show the correct matches that were chosen to include in the PPIN.

Genes	Possible matches
TRA	CENPU, AIRE, HEY1, CPEB1
U2AF1	U2AFBP, U2AF1L4, U2AF35
KNL1	CASC5, CENPK
SHTN1	KIAA1598, KIF20B
IGL	IGLL5, FAIM2
CHD2	ENSG00000173575, DSCAM, ENSG00000279765

possible that a gene identifier differs from the identifier of the corresponding gene product in the STRING database. It is also possible that the gene product of a gene is unknown or that there is no information about interacting partners for a gene product in the STRING database. The latter must be the case for 6 genes in the census, since the network consists of 716 gene products instead of 722. To find out for which genes a gene product was missing, the list of genes from the Cancer Gene Census was compared to the list of gene products obtained from STRING. Several gene identifiers in the Cancer Gene Census were not present in the list of gene product identifiers and vice versa. This meant that there were not only missing gene products, but also differences in identifiers between genes and corresponding gene products. For example, the gene LHFPL6 has a different identifier for its gene product: LHFP. Using the information (genomic location among others) about those genes and gene products available on the National Center for Biotechnology Information (NCBI), the genes in the census were mapped to products obtained from the STRING database. The genes for which no matching gene product was found were: *HMGN2P46*, *IGK*, *MRTFA*, *TENT5C*, *TRB* and *TRD*. These genes were input manually on the STRING database website to ensure there was no matching protein found by STRING, which was the case. The genes that had a different identifier for its gene product are shown in Table 2.
The resulting PPI network thus consisted of 716 gene products (nodes) from the genes in the CGC. The network statistics will be discussed in Section 3.1.

2.3 Hallmark Annotations

Hallmark annotations were available for roughly one third of the genes in the Cancer Gene Census. In the complete download of the census it was only specified whether genes had a hallmark annotation. Information about which hallmarks were annotated to which genes was not included in the download. However, this information was available on the website and, thus, was extracted manually. The hallmark annotations were written to a file to be able to directly import this information into the network in Cytoscape. The annotations can be found in the supplementary materials.

In the previous section it was pointed out that some genes in the census had an identifier that was different from their corresponding protein identifier in Cytoscape. None of those genes had a hallmark annotation, which means that all annotations could be imported directly from the manually created table without having to change identifiers.

Table 2: Genes of which the corresponding gene product in the STRING database had a different identifier.

Gene identifier in census	Product identifier in STRING
AFDN	MLLT4
CHD2	ENSP00000377747
IGH	IGHV4-38-2
IGL	IGLL5
KNL1	CASC5
LHFPL6	LHFP
MALAT1	ENSP00000485396
NSD2	WHSC1
NSD	WHSC1L1
SHTN1	KIAA1598
U2AF1	U2AFBP
WDCP	C2orf44

In the Cancer Gene Census 293 genes were said to have a hallmark annotation. However, when extracting the annotations manually from the website, it was found that 26 of those genes were not annotated. Thus, in total there were 267 annotated genes in the CGC.

2.4 Network Analysis

The obtained network with hallmark annotations was analysed using different methods in order to discover connections between hallmarks and to predict new hallmark annotations. First, clusters of densely connected regions in the network were identified, after which the common functions of the gene products in these clusters were determined using Gene Ontology enrichment analysis. Furthermore, cliques of gene products in the clusters were identified in order to predict hallmark annotations for gene products that show high similarity to annotated gene products. The methods that were used are explained in detail in this subsection.

2.4.1 Network Clustering

The network as a whole shows that the hallmarks of cancer are interconnected and spread across the network. The topology of the network was analysed in order to discover biological information about the cancer genes and the hallmarks of cancer. First, the network was explored using centrality analysis to identify highly connected gene products. These 'hubs' are very important players in cancer, since they influence or are influenced by the most other gene products in the network. NetworkAnalyzer [22], a popular Cytoscape plug-in for centrality analysis, was used to obtain metrics for each node such as degree, average shortest path, closeness and betweenness. On basis of these metrics, hubs were identified.

In biology, the most densely connected regions of a network implicate groups of gene products that are involved in the same biological processes or pathways, since these clusters may likely represent protein complexes [23]. This is useful for discovering connections between the hallmarks of cancer, one of the goals of this thesis. For each identified cluster of relatively closely related gene products, the biological processes in which those gene products are involved could be determined. In this way, it could be discovered whether gene products in a cluster contribute to similar processes and whether these processes could be linked to hallmarks of cancer.

The next step was thus to identify clusters in the network. Many algorithms and tools for discovering clusters in PPI networks have been developed [24]. The approaches of clustering algorithms could be divided into two groups: bottom-up and top-down approaches. The bottom-up approach obtains clusters one at a time by selecting a seed node to start a cluster with and expanding the cluster until certain criteria, often regarding the cluster density, are met. The top-down approach identifies different clusters simultaneously. The network as a whole is divided into different clusters by eliminating edges. Using the bottom-up approach, often a lot of nodes with few interactions are discarded as an effect of the constraints that have to be satisfied for adding nodes to a cluster, while the accuracy of the clusters (predicted true complexes) may be relatively high [25]. Top-down approaches generally result in a smaller amount of nodes that are disgarded, with a lower accuracy of the clusters as a consequence.

The clustering algorithm Molecular Complex Detection (MCODE) was used to identify clusters in the PPIN [26]. MCODE is a bottom-up clustering method that allows identification of both overlapping and non-overlapping clusters and is one of the most popular and accurate algorithms for clustering [27]. A bottom-up method was chosen because the accuracy of the clusters was considered more important than the size.

MCODE first weights every node in the network. For each node a weight is assigned on basis of the local network density and the highest k-core of the node neighbourhood. The neighbourhood of node v is defined as all nodes directly connected to v. A k-core is a graph in which each node has a minimum degree (number of edges) of k. The highest k-core is therefore the most densely connected subgraph, where the highest k-core of node v's neighbourhood is, thus, the set of nodes directly connected to v with minimum degree k, where k is maximal. In the algorithm the term core-clustering coefficient is introduced, defined as the density of the highest k-core of the neighbourhood of a node. Finally, the weight that is given to each node is the product of the core-clustering coefficient the node and the highest k-core level of its neighbourhood. The minimum amount of edges for a node to be scored, the degree cut-off, was set to 2 to prevent singly connected nodes from getting a (high) weight.

The next step of the algorithm is the prediction of complexes. The highest weighted node from the previous step will be taken as the seed node for a cluster. Then recursively moving outwards from nodes in the cluster to its neighbours, nodes that have a weight above a given threshold are added to the cluster. This threshold is the node score cut-off, a parameter that has to be set, which is the maximum percentage that a node's weight can differ from the weight of the seed node in order to be added to the cluster. This parameter was set to 0.1 (10%) to promote smaller and denser clusters. The maximum search distance from the seed node for nodes to be added to the cluster was set to 100. In this way, the focus of adding nodes to the cluster is more focussed on the density

of the cluster, rather than the distance between the node and the seed.
The creation of a cluster is completed when no more nodes can be added on basis of the threshold. Then a new cluster will be created in the same way, starting again with the highest weighted node as seed. As a result, the most densely connected regions in the network are identified as clusters.

The final step is post-processing of the identified clusters using set parameters. First, clusters that do not contain at least a k-core are filtered out, with k set to 2. Then, the "fluff" parameter allows for expansion of the clusters by checking the neighbourhood of each node in the cluster. Nodes that are added to a cluster in this way are not checked as seen and can thus be added to multiple clusters, possibly resulting in overlapping clusters. Since clusters don't have to be expanded and since there is no need for overlapping clusters, this parameter was set to false. Finally, the "haircut" option removes singly connected nodes from the clusters. The aim is to obtain densely interconnected clusters, so the 'haircut' option was turned on.

The Cytoscape plug-in MCODE (version 1.6.1) was used to identify clusters in the PPI network. The clusters of which the average of the weights of the nodes assigned by MCODE are at least 10 are selected, to investigate only the clusters that are most densely connected. This resulted in six clusters, covering 320 proteins in total, to be selected for further analysis.

For the selected clusters, statistics were calculated and the amount of annotations per hallmark were counted to compare with the whole cluster. This might give an indication of which hallmarks were overrepresented among the annotated gene products in each cluster.

2.4.2 Gene Ontology Enrichment Analysis

Connections between hallmarks of cancer can be discovered using the identified clusters. The hallmarks of cancer are essentially very general and broad biological processes, where each gene involved in a hallmark plays its own specific role to contribute to this biological process. Thus, specific biological processes, regulated by specific genes, together constitute a more general biological process: a hallmark of cancer. In this way, biological processes in which different genes are involved can be grouped together, resulting in more general processes that link to one or more specific hallmarks. When a set of genes related to each other, as the sets of genes in the clusters, show involvement in biological processes that can be linked to multiple hallmarks, it can be said that an overlap a connection between those hallmarks exists.

As a result of novel sequencing techniques, over the last decades an enormous amount of genes has been discovered. For a part of those genes the functionalities have also been identified. The Gene Ontology (GO) project is an ongoing effort for the integration of this biological information [28]. With this project the GO database was created, in which GO terms, brief descriptions regarding the functionality of genes, are provided in a structural manner. The GO database is organized in a hierarchical way using a tree structure, where nodes are GO terms and the directed edges are relationships between GO terms. There are three domains, which can be seen as the roots of three trees: biological process, molecular function and cellular component. The roots are, thus, the most general GO terms and the leaves are more specific GO terms. The relationship between two nodes is defined as either 'is-a' or 'part-of', e.g. the GO term "growth" *is a* "biological process" (the

relationship is always from child to parent). Later, more relationships such as "regulates", "occurs in" and "capable of" were added. Furthermore, new GO terms (nodes) are added regularly, as new knowledge comes to light.
An example of a GO term with respect to its parent terms is shown in Figure 2, where one of the 29 annotated GO biological processes of the *POLB* gene, a polymerase that's involved in DNA damage repair, is visualised.

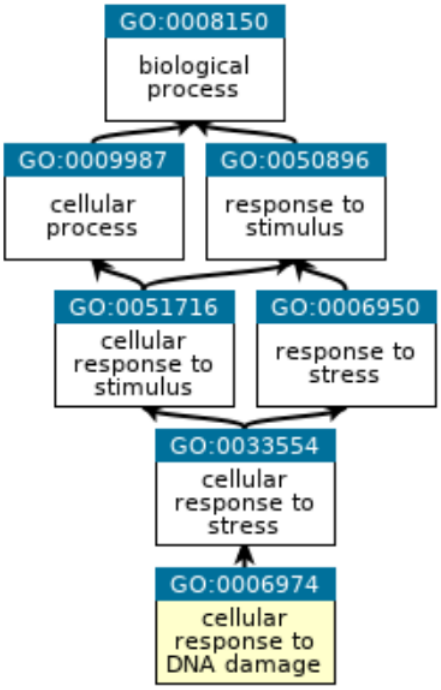

Figure 2: Visualisation of seven GO biological process terms, including the root, and their (directed) relationships. (From: https://www.ebi.ac.uk/QuickGO/term/GO:0006974)

With around four thousand cellular component terms, eleven thousand molecular function terms and almost thirty thousand biological process terms, the Gene Ontology currently consists of 44262 terms. The annotation database includes over 8 million annotations for more than 1.5 million gene products across 4643 different species. The GO database is commonly used by bioinformaticians and biologists to obtain and analyse the list of processes in which specified genes are involved. The specified genes are searched for in an annotation database, often the Gene Ontology Annotation database [29], to obtain the list of GO terms annotated to those genes. In this way, biological information about specific genes could be retrieved. This is called GO enrichment analysis and was performed in this thesis to identify the biological processes in which the genes in each cluster are involved. For each cluster, the gene product IDs were extracted and input to the analysis tool from the PANTHER classification system [30]. This system is up to date with the latest GO annotations. By choosing biological process as GO aspect and Homo Sapiens as species, the GO terms of the clusters were obtained. The Fisher's Exact test type was used and the p-values and false discovery rates were calculated by PANTHER.
The obtained lists of GO terms for the different clusters were very long, despite the choice of a

strict cut-off p-value of 0.001.

2.4.3 Summarising and Visualising GO Terms

As a result of the hierarchical structure of the GO database, the large amount of GO terms in it and the fact that each gene has multiple GO annotations, GO enrichment analysis often leads to a very long list of GO terms ranging from very specific to very general terms. Moreover, the list often includes a lot of redundant terms, e.g. in Figure 2 the term "Response to stress" fully encompasses its child term "Cellular response to stress". In a list of enriched GO terms, it's often the case that such a parent term is significantly enriched only due to the fact that it includes all the genes enriched with the child term.

To discover which biological processes are regulated by the clusters and to link these processes to hallmarks of cancer, the obtained lists of GO terms had to be simplified. This could be done by comparing the GO terms and removing redundancies. While comparing protein sequences or structures can be done directly due their systematic representation, this is not as straightforward for biological processes. Nonetheless, as a result of the directed acyclic graphs in the GO database in which terms are represented by means of their relation to other terms, the similarity between two GO terms or sets of GO terms could be determined. Functional or semantic similarity is defined as a function that returns a value indicating how closely related two GO terms are and could be used to remove redundancies and summarise a list of GO terms.

Research has pointed out the different approaches to calculate semantic similarity scores [31]. These approaches can be divided into two types: edge-based and node-based. In general, edge-based approaches are based on counting the amount of edges between two GO terms that have to be compared. In this way, the distance between two terms is used to calculate the semantic similarity of the terms. Node-based approaches focus on comparing the properties of two terms, their parents and/or their children. A recent evaluation of different semantic similarity measures has shown that node-based methods perform best [32], of which the most popular ones will be briefly explained.

Information content (IC) is a concept that forms the basis of the node-based approaches for calculating semantic similarity. IC is a numerical value that's based on the specificity and informativeness of a term. The IC of term t can be defined as the negative log of the probability of t,

$$IC(t) = -\log\left(p(t)\right) \tag{1}$$

where the probability of t corresponds to the frequency of occurrence of t in the GO annotation database. When term t is annotated to a lot of genes, its frequency of occurrence in the database is high and, thus, its information content is low.

Resnik's method [33], which is the oldest and yet the most popular method, defines the semantic similarity between two terms t_1 and t_2 by the IC of their most informative common ancestor (MICA), the common ancestor with the highest IC:

$$Sim_{Resnik}(t_1, t_2) = IC(t_{MICA}) \tag{2}$$

When the MICA is a very general term, it occurs more often, hence, the IC is lower and the compared terms are not considered to be very similar. In this way, the information shared by two terms is taken into account. However, the distance between two terms in the GO graph is not considered: it doesn't matter how far away the MICA is from the terms. The measures introduced by Lin [34] and Jiang and Conrath [35] try to take this into account by including the IC of the terms that are compared:

$$Sim_{Lin}(t_1, t_2) = \frac{2 \times IC(t_{MICA})}{IC(t_1) + IC(t_2)} \tag{3}$$

$$Sim_{Jiang\&Conrath}(t_1, t_2) = 1 - IC(t_1) + IC(t_2) - 2 \times IC(t_{MICA}) \tag{4}$$

These methods don't take the level of detail of the MICA of the two terms into account. Schlicker's method [36] tries to include the level of specificity of the MICA by adding a weight regarding its probability (frequency of occurence) to Lin's method (3):

$$Sim_{Schlicker}(t_1, t_2) = SemSim_{Lin}(t_1, t_2) \times (1 - p(t_{MICA})) \tag{5}$$

A few tools that summarise long lists of GO terms by removing redundancies on basis of the semantic similarity of GO terms are available. REVIGO was chosen to summarise the list of GO terms for each cluster. REVIGO comes with the option of choosing one of the four discussed semantic similarity measures to use for removing redundancies. Since Schlicker's method not only takes the information shared by two terms into account, but also the distance between those terms and the level of detail of its MICA, this method seems the best choice for summarising the long lists of GO terms annotated to the gene products in each cluster with REVIGO.

REVIGO first uses the semantic similarity measure to group the provided GO terms into clusters of very similar terms. This procedure of clustering is similar to the neighbour joining approach for hierarchical clustering [37]. For each of these clusters, one or more representatives have to be chosen. In this way, the long list of GO terms are reduced by removing redundant terms.

First of all, the semantic similarity between all pairs of GO terms is calculated. Then, one of the two most similar GO terms is removed from the list. Which of the two terms will be deleted is determined by several ordered criteria. First it is checked whether one of the terms is very uninformative, i.e. has a frequency > 0.5 in the specified GO annotation database and, thus, a low IC. If this is not the case, then the term that has a less significant p-value is deleted. However, when the p-values are fairly close, the parent term will be deleted if a parent-child relationship exists between the terms. Lastly, if there is no such relationship between the terms, the choice is made *at random*. This process of deleting one of the two most similar GO terms is done recursively, until those terms' similarity is beneath the user-specified cut-off value, which was set to 0.5 ('medium'). A flowchart that visualises this procedure is shown in Figure 3. In this way, the GO terms annotated to the gene products of each cluster are summarised by removing redundancies.

Visualisation of the remaining non-redundant GO terms is key for interpretation. Which biological processes are overrepresented in each cluster? Can these processes be linked to hallmarks of cancer? The answers to those questions must be determined as objective as possible in order to say

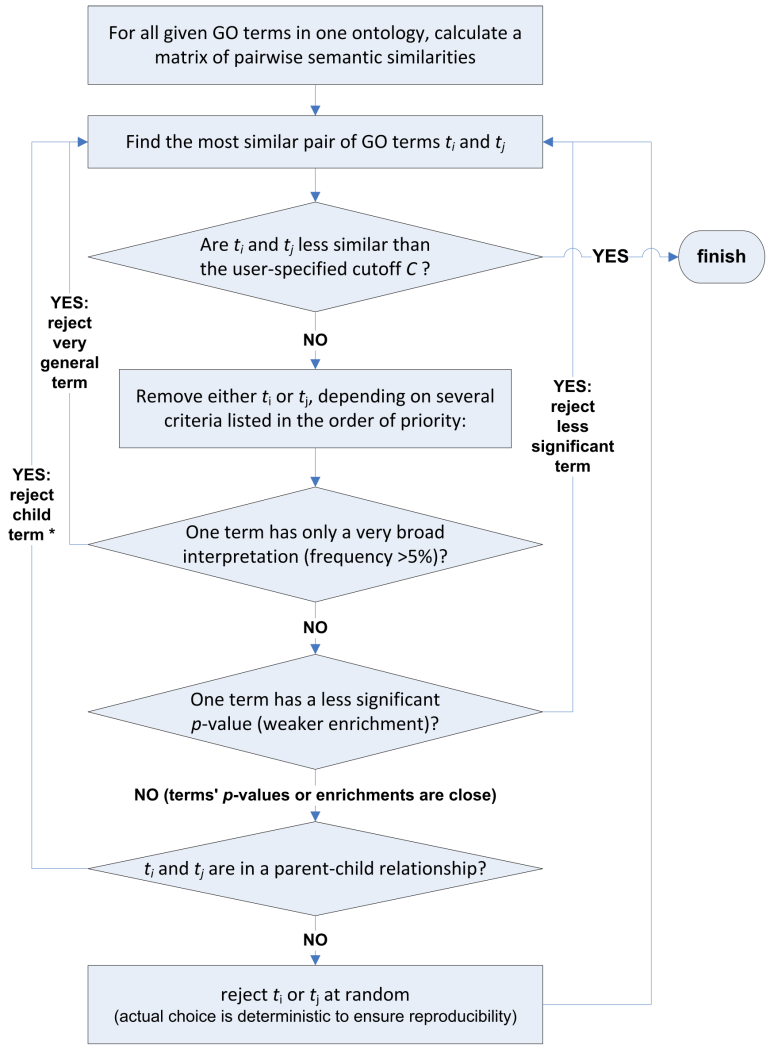

Figure 3: Flowchart visualising the procedure used by REVIGO to summarise a list of redundant GO terms into representative terms. (From: *REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms*, https://doi.org/10.1371/journal.pone.0021800)

something about connections between hallmarks. This is hard, since the biological processes in the summarised list of terms are often more specific than the hallmarks itself and there are still quite a lot of GO terms left. REVIGO offers several visualisation options, of which the treemap seems best to overcome the mentioned difficulties. The remaining GO terms are here joined into several high-level groups, labelled with keywords that are overrepresented in the GO terms of those groups. Examining the group labels and its (more specific) GO terms, could result in links to hallmarks. However, the treemap has some downsides. Most importantly, GO terms in the groups and the proportion of biological processes in each group is visualised using different sizes for the groups (instead of numerical values), which is not very intuitive or exact.

Several other tools are available for visualising lists of GO terms. CirGO [38] is recently developed software for visualising non-redundant hierarchically structured GO terms that doesn't contain the mentioned downsides of REVIGO's visualisation tools. The treemap from REVIGO could be exported as csv file, which could be used as input for the CirGO visualisation tool. CirGO creates an informative circular plot with a two-layer hierarchy, where the inner circle represents the high-level groups as labelled by REVIGO and the outer circle contains the representative GO terms. Furthermore, the proportion of each group is provided as numerical information. CirGO was used to visualise and interpret the summarised lists of GO terms obtained from each cluster, in order to discover overlaps between hallmarks of cancer on basis of overrepresented biological processes that could be linked to hallmarks.
For each cluster, the amount of annotations per hallmark was also counted and compared with the whole network to possibly get an indication of which hallmarks are overrepresented in which clusters.

2.4.4 Discovering Cliques in the Clusters

The second goal of this thesis was to predict hallmark of cancer annotations using existing annotations. In order to do this, gene products without a hallmark annotation that are very functionally similar to annotated gene products had to be identified. Clusters contain gene products that are likely to form molecular complexes and to be involved in the same functionalities. However, to obtain groups of gene products for which this is even more likely, cliques had to be discovered from the clusters. Cliques are, namely, fully connected subgraphs, and, thus, consist of gene products that could interact with every other gene product in the same clique. Thereby, a clique contains gene products that are very likely to form a complex and fulfill the same task(s).

MClique is a Cytoscape plug-in for identifying maximal cliques in a network. This plug-in was used to discover all maximal cliques in the different clusters. This resulted in long lists with possible maximal cliques to further examine. A script was written in Python (with the code available in the supplementary materials) to identify interesting cliques on basis of the existing hallmark annotations. For each cluster different cliques for which at least one hallmark was annotated to all annotated gene products were selected.

2.4.5 Semantic Similarity between Gene Products in Cliques

The gene products in the selected cliques are very likely to form a complex that regulates similar processes. Moreover, a high semantic similarity between the gene products in a clique makes this evidence even stronger. Therefore, the semantic similarity between all pairs of gene products in each clique was calculated. Since there is no limit to which semantic similarity measures could be used (as was the case with using REVIGO), there is a wide range of measures that could be used. A recent analysis of commonly used semantic similarity measures has shown that the node-based measures perform best, as mentioned in section 2.4.3 . In this analysis, the difference in accuracy between the node-based methods was not significance. Wang's method [39], however, seems to overcome the limitations of classic node-based methods (from Resnik and Lin). One of these limitations is that the more often a term is annotated, the lower its IC is, which may lead to different IC scores when using different GO annotation databases. Wang also overcomes the problem of the fact that two GO terms that occur higher in the tree-structured graph are more general terms and, thus, are less similar than two GO terms with the same distance that are lower in the graph, where other node-based measures don't take this into account. Wang does this by first defining the "Knowledge" of term t as:

$$K(t) = 1/IC(t) \tag{6}$$

where the IC of t is defined as in Equation (1). K is further normalized to the "Semantic Weight" of term t as follows:

$$SW(t) = \frac{1}{1 + e^{-K(t)}} \tag{7}$$

The Semantic Weight is calculated for the ancestors of the terms that are to be compared (including the terms itself), which are summed up to constitute the "Semantic Value" SV of term a that has to be compared with another term, thus:

$$SV(a) = \sum_{t \in T_a} SW(t) \tag{8}$$

In this way, the *knowledge* we have about all ancestor terms of a could be taken into account. The semantic similarity between two terms a and b is calculated by dividing two times the sum of the semantic weights of all of their overlapping ancestor terms by the sum of the semantic values of term a and b, i.e. the semantic similarity is based on their aggregate information content:

$$Sim_{Wang}(a, b) = \frac{\sum_{t \in T_a \cap T_b} 2 \times SW(t)}{SV(a) + SV(b)} \tag{9}$$

Wang's method was chosen to calculate the semantic similarity between all pairs of gene products in each cluster. The package GOSemSim [40] for R, released under the GNU General Public License within Bioconductor [41], was used to compute these values. Since gene products are very often annotated with more than one GO term, these different terms have to be combined in order to calculate the semantic similarity of two gene products (which can essentially be seen as two sets of GO terms in this case). GOSemSim provides four methods for this: *maximum*, *average*, *rcmax* and *best-match averages* (BMA). The max and average methods have been proven to be outperformed by the BMA approach [42], mainly because the BMA method is independent of the amount of terms to be combined. For two genes g_1 and g_2 the sets of annotated GO terms could be noted

as $GO_1 = \{go_{11}, go_{12}, ..., go_{1m}\}$ and $GO_2 = \{go_{21}, go_{22}, ..., go_{2n}\}$, respectively. The BMA is then defined as the average of the maximum similarities for all terms:

$$Sim_{BMA}(g_1, g_2) = \frac{\sum_{i=1}^{m} \max_{1 \leq j \leq n} Sim(go_{1i}, go_{2j}) + \sum_{j=1}^{n} \max_{1 \leq i \leq m} Sim(go_{1i}, go_{2j})}{m + n} \tag{10}$$

The rcmax method uses the maximum of the average maximum similarity of each term in GO_1 and the average maximum similarity of each term in GO_2:

$$Sim_{rcmax}(g_1, g_2) = \max\left(\frac{\sum_{i=1}^{m} \max_{1 \leq j \leq n} Sim(go_{1i}, go_{2j})}{m}, \frac{\sum_{j=1}^{n} \max_{1 \leq i \leq m} Sim(go_{1i}, go_{2j})}{n}\right) \tag{11}$$

Since the BMA method takes all terms into account, where rcmax doesn't, and this method has been a popular choice in research, BMA was chosen to combine the semantic similarities of GO terms into a single value. For each selected clique, the semantic similarity between all pairs of gene products has been calculated.

Considering the fact that the gene products in a clique could very likely form a complex, in combination with high semantic similarity between those gene products, annotations for unannotated gene products were predicted with the hallmarks that were annotated to all (annotated) gene products in the clique. Often a threshold of 0.5 is chosen for two genes to be significantly similar. However, there is no real method for determining this threshold. Here we chose to use 0.650 as a threshold for the semantic similarity of the gene products to be called significantly similar, since the goal here was to accurately predict hallmark annotations. Nonetheless, ten randomly selected subgraphs of size 15 from the network with all gene products were compared to the cliques, on basis of the average semantic similarity between their pairs of gene products. In this way, it could be checked whether each clique was more likely to regulate the same biological processes and, thus, was more likely to actually form a complex than the random samples. Finally, only for the gene products that had a significantly high semantic similarity ($Sim > 0.650$) with one of the annotated gene products, the hallmark(s) that was/were annotated to all annotated gene products in the clique was predicted to be annotated, but only if the average pairwise semantic similarity of the whole clique was higher than that of the random samples.

3 Results

After the genes in the Cancer Gene Census were imported into Cytoscape, the PPI network was constructed using the StringApp. In this section, first the network statistics will be discussed, followed by the cluster statistics and results from the GO enrichment analysis, and finally information about the identified cliques and the results of the analysis of the semantic similarities between its gene products will be pointed out.

3.1 Network Statistics

The whole network of cancer gene products consisted of 716 nodes (gene products) and 14867 edges (interactions). The visualisation of the PPI network in Cytoscape is shown in Figure 4, where blue

nodes indicate not annotated gene products, red nodes indicate annotated gene products (where darker red equals more annotations) and the size of a node correlates with its degree. The PPI network shows a lot of interconnections, however, there are several (mainly not annotated gene products) with a low degree. As can be seen, there were 10 disconnected nodes in the network. These nodes could not be used for the analysis and were therefore not used to calculate the network statistics.

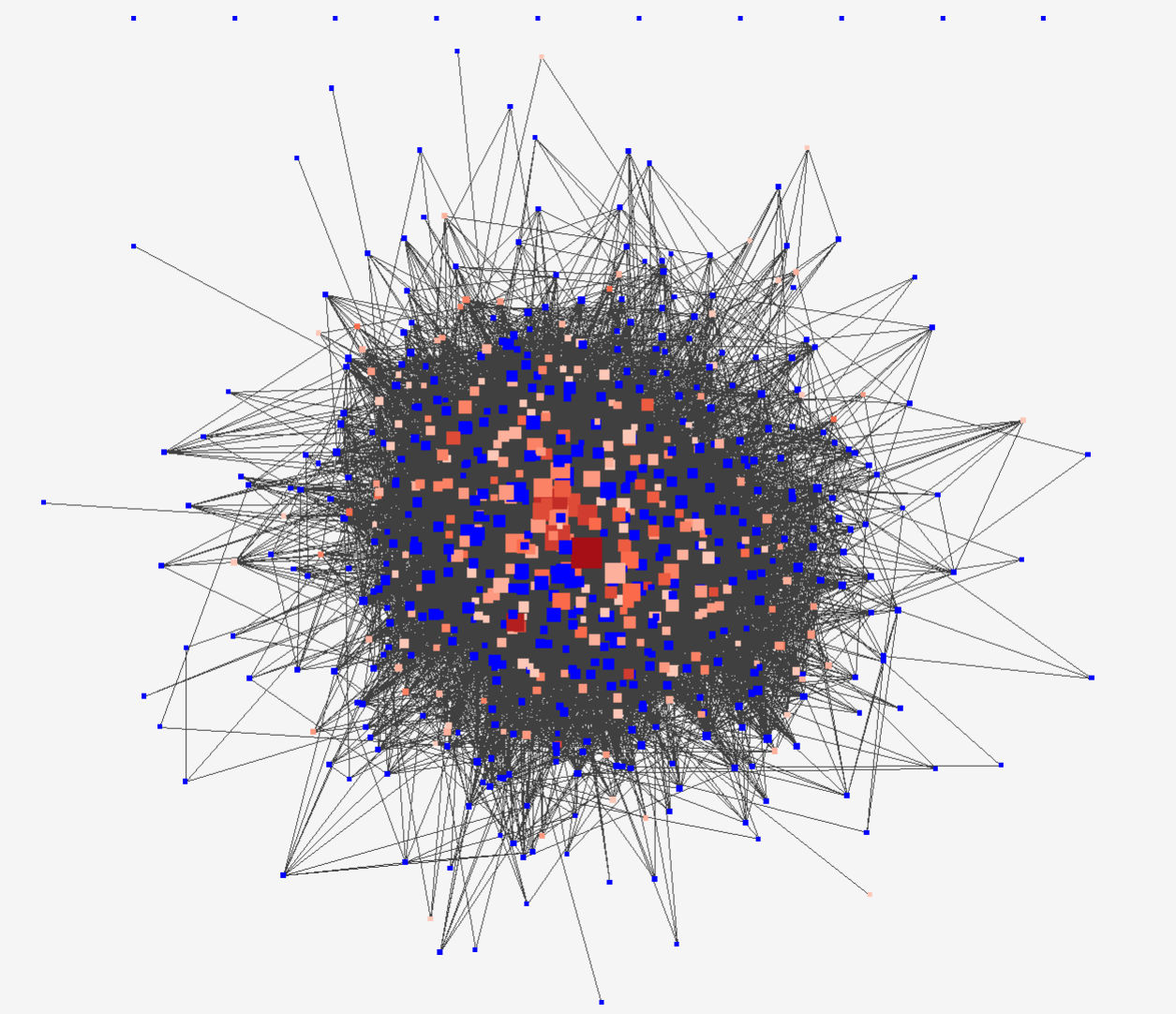

Figure 4: The whole PPI network of 716 gene products as visualised in Cytoscape. Blue nodes represent gene products that have not yet been annotated, red nodes represent gene products that have a hallmark annotation, with darker red indicating more hallmarks being annotated, and the size of a node correlates with its degree.

The gene product with the highest degree was TP53 with a degree of 352, followed by MYC with 281 and AKT1 with 258. Those were also the gene products with the highest centrality measures with a

closeness of 0.659, 0.616 and 0.603, and a betweenness of 0.087, 0.038 and 0.034, respectively. These gene products were clearly hubs in the network. In addition, they were also annotated to many of the hallmarks. TP53 contained 11 annotations among 9 hallmarks (both promoting and suppressing hallmarks 4 and 9), MYC was annotated to 8 hallmarks and AKT1 contained 9 annotations.
The average degree of all gene products in the network was 42.2. For each gene product in the network, the average shortest path length was obtained using NetworkAnalyzer. The average of this measure for all nodes in the network was 2.261. This means that the network is highly interconnected, since on average only 2.261 edges lay between each pair of nodes. Centrality measures were also calculated. The average closeness centrality was 0.443.

In Figure 5 the degree distribution, $P(k)$, is plotted for each possible degree, k, between 0 and 355. It is visible that the degree distribution approximately follows a powerlaw, $P(k) \sim k^{-\gamma}$, with γ, the degree exponent, between 1 and 2. The network has, thus, the properties of a scale-free network, with the hubs playing an important role due to the low γ.

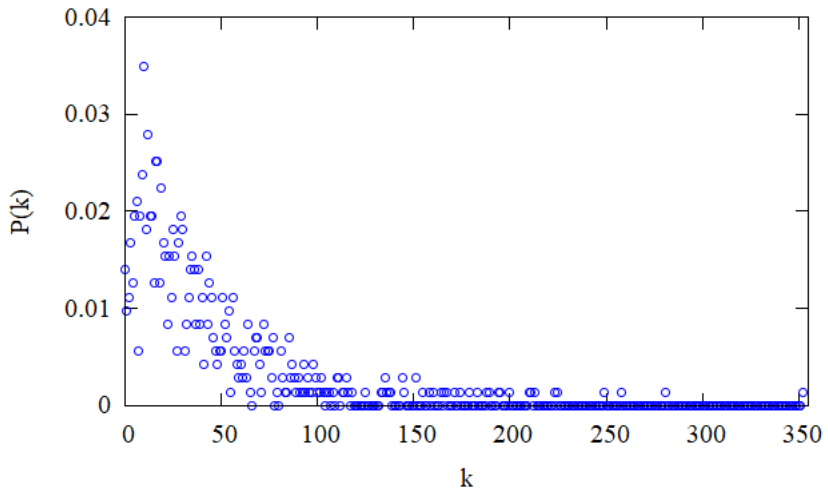

Figure 5: The frequency of occurrence, $P(k)$, of each possible degree k between 0 and 350.

After the hallmark of cancer annotations were added to the network, statistics regarding these annotations were obtained. The network consisted of 267 gene products that had one or more hallmark annotations and 449 gene products without a hallmark annotation. In total, there were 819 hallmark annotations among the 267 annotated gene products, which means that each of these gene products had approximately 3 annotations on average. However, TP53, previously identified as hub, was annotated with 11 hallmarks, which makes this gene product the most annotated gene product in the network. In Figure 6 it is visualised how many gene products are annotated with each hallmark of cancer. Hallmark 3 and 6, invasion and metastasis and resisting cell death, respectively, were both annotated to the most gene products, followed by hallmark 1, sustaining proliferative signaling. Hallmark 3, avoiding immune destruction, and 5, tumour promoting inflammation, have been annotated to the least amount of gene products.

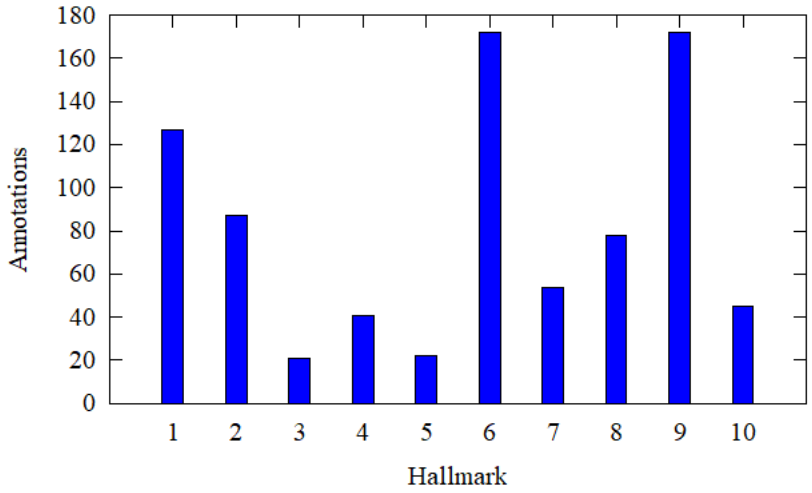

Figure 6: The amount of annotations for the ten different hallmarks among the 267 annotated cancer genes.

3.2 Cluster Statistics

MCODE has identified six clusters of which the average of the weights assigned to the nodes was at least 10. The clusters are shown in Figure 7. It could be seen that the first two clusters were the most densely connected and contained the most annotated gene products. The last two clusters were the least densely connected, but seemed to contain some more densely connected groups. It stands out that cluster 4 was very small, consisting of only nine gene products among which two contained a hallmark annotation.

The statistics of each cluster are shown in Table 3. These statistics also show that the first two clusters are very densely connected, where cluster 5 and 6 are less densely connected. Their highest closeness is, however, higher than the average closeness in the whole network, which indicates that there may be some densely connected groups in these clusters. This was confirmed by the visualisation of the clusters, which showed that there are a couple of groups of gene products in cluster 5 and 6 that show a high density. For this reason, these clusters were still used for further analysis. Cluster 4 was a very small cluster, consisting of only nine nodes that almost formed a fully connected (sub-)graph.

In Figure 8 and 9 the hallmark annotations per cluster are shown. The first two clusters had the highest proportion of annotated gene products and the highest amount of annotations. In Figure 9 the proportions of annotations per hallmark are shown as percentages and compared to those of the whole network. The hallmark annotations seem to be distributed in each cluster, which shows the interconnections between the hallmarks. Some hallmarks are clearly more represented among the annotated gene products than others in some clusters. There is not one hallmark that has a high percentage of annotations in all six clusters (compared to the whole network). Hallmark 8 seemed to be very well represented in cluster 3. However, it should be noted that the total amount of annotations in this cluster, 71, is quite low compared to the amount of annotations in the whole

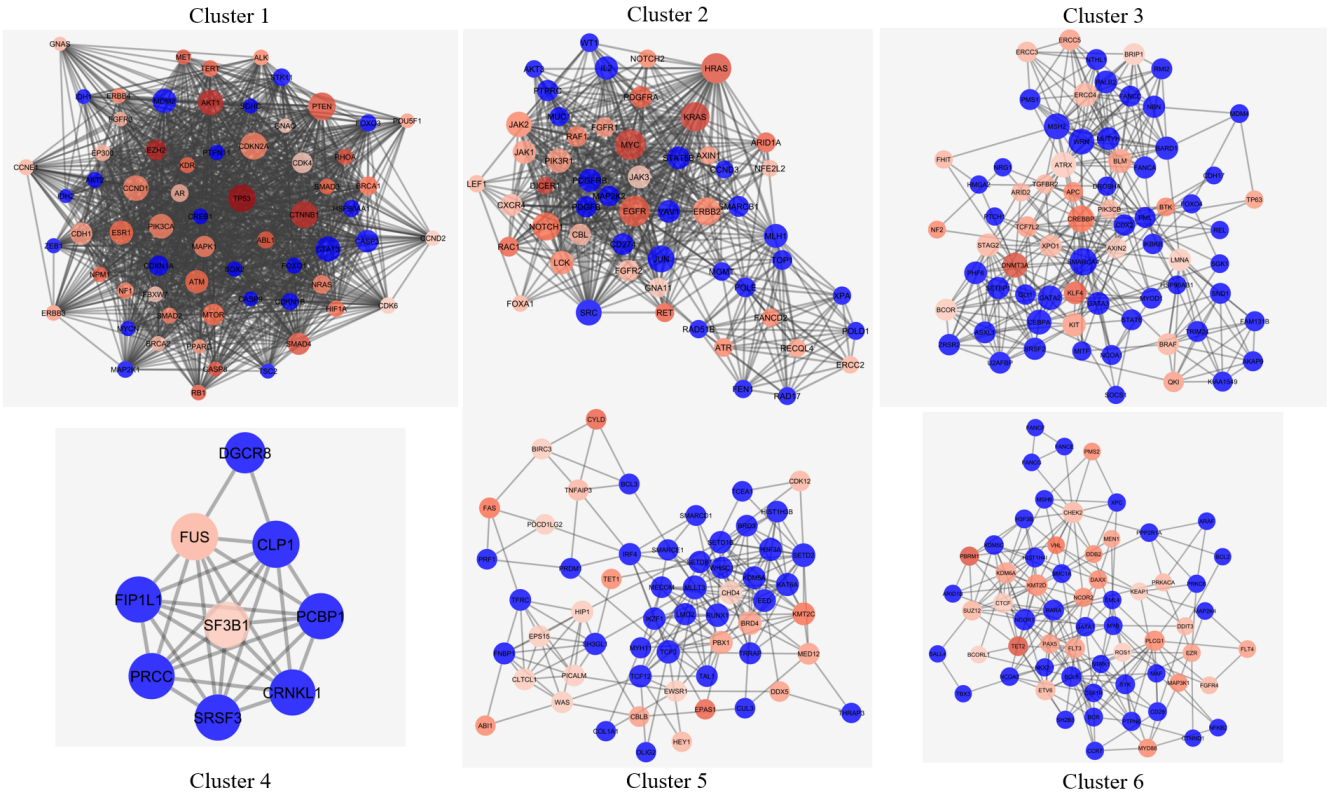

Figure 7: The six clusters, of which the averages of the weights were above the threshold, as visualised in Cytoscape. Blue nodes represent gene products that have not yet been annotated, red nodes represent gene products that have a hallmark annotation, with darker red indicating more hallmarks being annotated, and the size of a node correlates with its degree.

Table 3: Statistics for each of the selected clusters. The first two clusters contained the highest number of edges and the highest average degree k. The highest and lowest degree, together with the highest closeness and average shortest path length for the nodes in each cluster also show the density and interconnections of the clusters.

Cluster	Nodes	Edges	Avg. k	Highest k	Lowest k	Highest closeness	Avg. SPL
1	65	1347	41.45	64	23	1.000	1.352
2	55	553	20.11	44	7	0.844	1.734
3	71	345	9.72	24	2	0.593	2.33
4	9	30	6.67	8	2	1.000	1.167
5	57	199	6.98	15	2	0.544	2.597
6	63	207	6.57	16	2	0.492	2.668

network. As an effect, a single annotation has a greater impact on the percentages. This is also the reason that the percentages of annotations in cluster 4 were not filled in, since this cluster included only three annotations. Furthermore, the distribution of the hallmark annotations only say something about the overrepresentation of specific hallmarks among the annotated gene products. The gene products that have not yet been annotated, however, also play an important role in the clusters. Especially due to the fact that there are more gene products that have not yet been annotated in clusters 3 to 6. Therefore, these statistics can only be used as an indication for which hallmarks are well represented among the annotated gene products in each cluster and to compare to the results of further analysis.

Cluster	Gene products	Annotated	Not Annotated	Annotations
1	65	44 (67,69%)	21 (32,31%)	199
2	55	31 (56,36%)	24 (43,64%)	125
3	71	26 (36,62%)	45 (63,38%)	71
4	9	2	7	3
5	57	23 (40,35%)	34 (59,65%)	60
6	63	27 (42,86%)	36 (57,14%)	85
Total	716	267 (37,29%)	449 (62,71%)	819

Figure 8: The amount of annotated gene products and the total amount of annotations for each cluster and the whole network.

Cluster		H1	H2	H3	H4	H5	H6	H7	H8	H9	H10
1	Annotations:	26	18	8	12	8	38	18	18	38	15
	Percentage:	13,07	9,05	4,02	6,03	4,02	19,10	9,05	9,05	19,10	7,54
	Difference w/ total:	**-2,44**	**-1,58**	1,46	1,02	1,33	**-1,91**	2,45	**-0,48**	**-1,91**	2,04
2	Annotations:	23	6	6	9	5	21	11	9	27	8
	Percentage:	18,4	4,8	4,8	7,2	4	16,8	8,8	7,2	21,6	6,4
	Difference w/ total:	2,89	**-5,82**	2,24	2,19	1,31	**-4,20**	2,21	**-2,32**	0,60	0,91
3	Annotations:	8	9	1	3	0	12	1	16	19	2
	Percentage:	11,27	12,68	1,41	4,23	0,00	16,90	1,41	22,54	26,76	2,82
	Difference w/ total:	**-4,24**	2,05	**-1,16**	**-0,78**	**-2,69**	**-4,10**	**-5,18**	13,01	5,76	**-2,68**
4	Annotations:	0	1	0	0	0	1	0	0	1	0
	Percentage:										
	Difference w/ total:										
5	Annotations:	9	7	2	4	3	13	4	5	10	3
	Percentage:	15,00	11,67	3,33	6,67	5,00	21,67	6,67	8,33	16,67	5,00
	Difference w/ total:	**-0,51**	1,04	0,77	1,66	2,31	0,67	0,07	**-1,19**	**-4,33**	**-0,49**
6	Annotations:	17	7	1	1	3	20	2	9	21	4
	Percentage:	20,00	8,24	1,18	1,18	3,53	23,53	2,35	10,59	24,71	4,71
	Difference w/ total:	4,49	**-2,39**	**-1,39**	**-3,83**	0,84	2,53	**-4,24**	1,06	3,70	**-0,79**
Total	Annotations:	127	87	21	41	22	172	54	78	172	45
	Percentage:	15,51	10,62	2,56	5,01	2,69	21,00	6,59	9,52	21,00	5,49

Figure 9: The amount of annotations of each hallmark for the annotated gene products in each cluster. Also, the percentage annotations of the total amount of annotations for each hallmark is shown and compared to the percentage of annotations for each hallmark where the largest differences for each cluster are marked with a green colour.

3.3 GO Enrichment Analysis

To discover which biological processes are regulated by the gene products of each cluster, GO enrichment analysis was performed. The gene products were input to the PANTHER classification system. The resulting lists of GO terms annotated to the gene products were very long for most clusters, as can be seen in Table 4.

Table 4: Amount of GO terms annotated to the gene products of each cluster (also for terms with p-value below the set threshold) retrieved from the PANTHER classification system.

Cluster	GO terms	GO terms with $p < 0.001$
1	1812	1428
2	1266	1009
3	596	526
4	17	17
5	164	164
6	541	483

Due to the amount of annotations, the lists with GO terms had to be summarised to discover which biological processes were overrepresented in each cluster. This was done using REVIGO, after which this information was visualised with CirGO.

3.4 Linking Biological Processes to Hallmarks of Cancer

The circular plots retrieved from CirGO are shown below. CirGO used the treemaps generated from REVIGO. In a treemap, the summarised list of enriched GO terms is visualised and grouped together under high-level tags. The tag clouds show keywords in the overrepresented GO term descriptions and keywords that correlate with the p-values of GO terms. The tags are visualised in the inner circles of the plots made with CirGO. The outer circles consist of the representatives of the redundant list of enriched GO terms.

Cluster 1

In Figure 10 the diagram from CirGO with the list of non-redundant GO terms enriched to the gene products of cluster 1 is visualised. "Response to hormone" was the group of GO terms that was most represented, with 31.0% of all representative GO terms in this group. In this group there were a lot of GO terms involved in signaling and responses to stimuli. "Regulation of cell communication", "cell surface receptor signaling pathway" and "regulation of signaling" indicate a link between this group of representative GO terms and hallmark 1, sustaining proliferative signaling. Furthermore, with 26 annotations this hallmark is well represented among the annotated gene products, as can be seen in Figure 9.
The second most represented group, which closely followed the first, was "negative regulation of developmental process", with 29.7% of the representative GO terms. Some GO terms in this group, such as "tube development" , "gland development" and "circulatory system development", imply

that hallmark 7, inducing angiogenesis, was overrepresented in this cluster. However, there were too few specific GO terms among the list of representative terms that directly link to angiogenesis. Due to the fact that not all representative GO terms could be visualised in the diagram, the terms "angoigenesis" and "blood vessel development" were searched for in the complete list obtained from REVIGO (available in the supplementary materials). Both terms were not present, and, thus, hallmark 7 was not overrepresented in this cluster. In contrast, there were many terms among this group that could be linked directly to hallmark 2, evading growth suppressors. "Negative regulation of cell cycle", "reproductive structure development" and "negative regulation of cell proliferation", among others, are terms that indicated the representation of this hallmark. In addition, several terms in the first group are involved in response to stimuli, that could be interpreted as the representation of hallmark 2 in this group. With 18 annotations of hallmark 2, the annotated gene products strengthen this representation.
The third most represented group of representative GO terms was "negative regulation of biological process" with a presence among the representative GO terms of 14.7%. This group covered a lot of GO terms that involve the regulation of metabolic processes. Even in the fourth group there were representative terms about the regulation of metabolic processes. These terms could be linked to hallmark 10: deregulating cellular energetics, since this hallmark involves reprogramming of the energetics in order for the cancer cells to proliferate and grow. Moreover, in Figure 9 it is visible that this hallmark was well represented among the annotated gene products in this cluster.
The difference between the remaining terms was too big to link another hallmark to the gene products in this cluster.

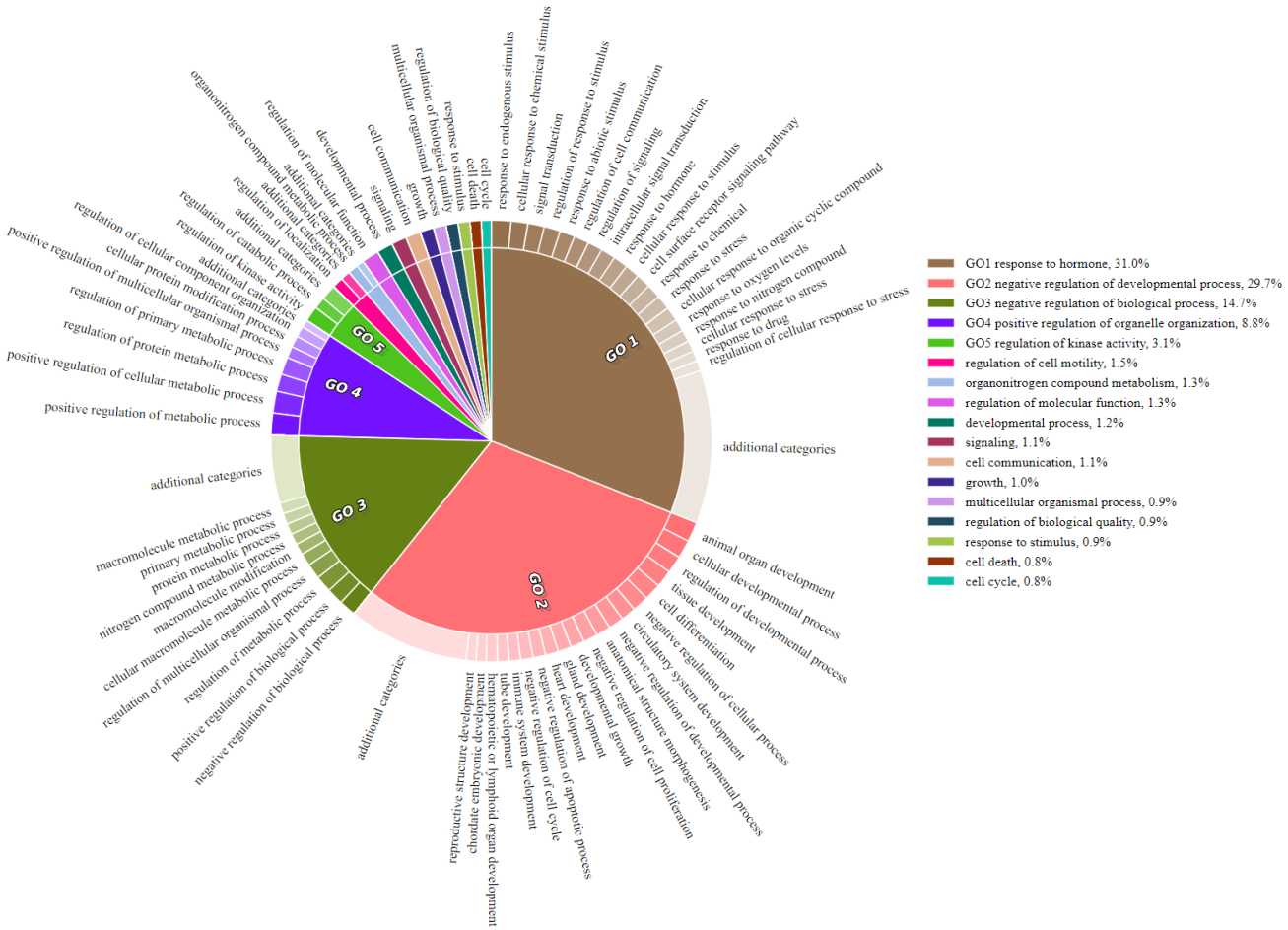

Figure 10: Diagram with the representative GO terms enriched to the gene products in cluster 1. In the legend the labels of the high-level groups are shown including the proportion of the terms that are included in those groups.

Cluster 2

The diagram with the (groups of) representative GO terms for cluster 2 obtained from CirGO is shown in Figure 11. The largest group of GO terms (26.9%) was "transmembrane receptor protein tyrosine kinase signaling pathway". This is a series of signals, where the receptors possess tyrosine kinase activity. Signals concerning growth and differentiation are frequently transmitted in this way, as mentioned in the introduction. In the diagram, a lot of terms that involve the (regulation of and response to) signal transduction are included in this group. This group of GO terms could, thus, be linked to proliferative signaling, hallmark 1. Moreover, in other groups, there also occur GO terms concerning proliferative signaling, e.g. peptidyl-tyrosine phosphorylation in the third group and positive regulation of cell proliferation in the fourth group. Comparing the amount of annotations of this hallmark in cluster 2 to that of the whole network, this hallmark seemed to be well represented among the annotated gene products, as can be seen in Figure 9.
It should be noted that, again, in the first group there are terms included that could be linked to hallmark 2. "Cellular response to growth factor stimulus" and "regulation of response to stimulus" in this group, together with "developmental process" in one of the smaller groups, indicated the representation of this hallmark. However, there were too few (specific) terms that strengthen this representation. Moreover, hallmark 2 was annotated to only 6 gene products in this cluster.
The second group of GO terms was labeled with gland development, including 20.9% of the representative GO terms. With representative GO terms like gland development, tube development/morphogenesis and morphogenesis of an epithelium, this group of GO terms might be linked to hallmark 7, inducing angiogenesis. "Blood vessel development", a direct parent term of "angiogenesis", was also one of the representative GO terms in group 2 (under additional categories). This specific term in combination with the other terms that could be linked to hallmark 7, indicate the representation of this hallmark. In addition, compared to the whole network there were a lot of annotations for this hallmark in cluster 2.
In the third group, there were many terms regarding (the regulation of) metabolic processes, as well in a couple of other groups. This implies a link with hallmark 10, deregulating cellular energetics.

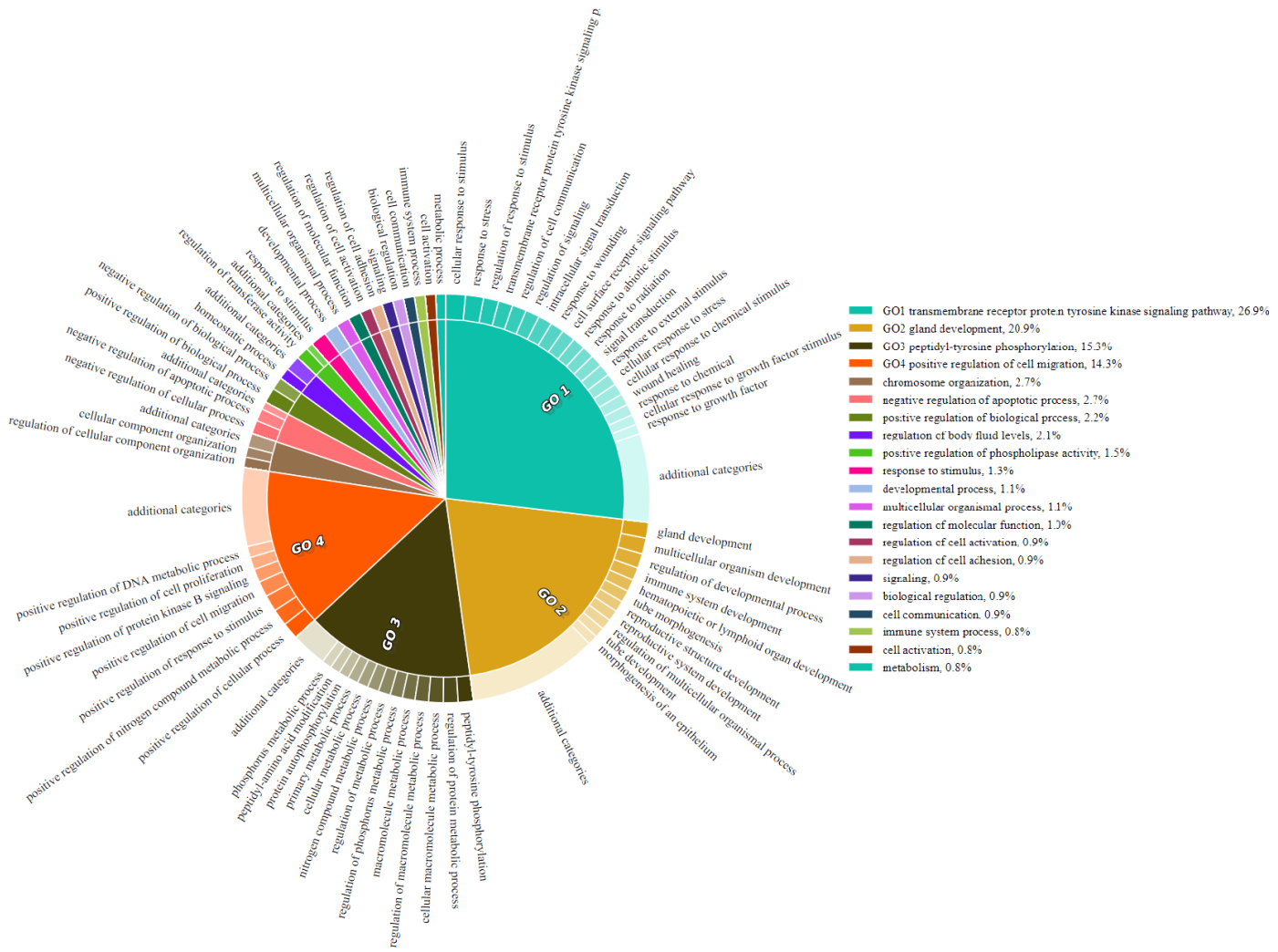

Figure 11: Diagram with the representative GO terms enriched to the gene products in cluster 2. In the legend the labels of the high-level groups are shown including the proportion of the terms that are included in those groups.

Cluster 3

In Figure 12 it is shown that most of the representative GO terms from cluster 3 fall under the group labeled with "cell fate commitment". The capacity of cells to differentiate into particular kinds of cells could not directly be linked to one hallmark. However, several GO terms across different groups in the diagram could be linked to hallmark 6: invasion and metastasis. "Animal organ development/morphogenesis", "system development", "(regulation of) developmental process", "hemopoiesis" and "multicellular organismal process" are all processes that contribute to the extension and penetration by cancer cells to neighbouring cells. It should be noted, though, that most of these terms are relatively general terms. In combination with the fact that 12 of the 26 annotated gene products have an annotation with hallmark 6, it could be said that this hallmark is well represented in the cluster, although the evidence isn't very strong.
There was, however, an explicit overrepresentation of GO terms that are involved in genome instability and mutations, hallmark 8. "Chromosome organization" (group 5), "cellular response to DNA damage stimulus" (group 2) and almost all terms in group 4, which included 17.2% of the representative GO terms, are examples of terms that indicate the overrepresentation of hallmark 8 in this cluster. Furthermore, with more than 22% of the annotations in the cluster being for hallmark 8 (see Figure 9), this hallmark also seemed to be overrepresented compared to the annotations in the whole network.
Biological processes regarding metabolic processes also occur a lot in the diagram, specifically in group 3. This could be linked to the representation of hallmark 10, change of cellular energetics, in this cluster. However, only 2 out of the 26 annotated gene products were annotated with this hallmark.

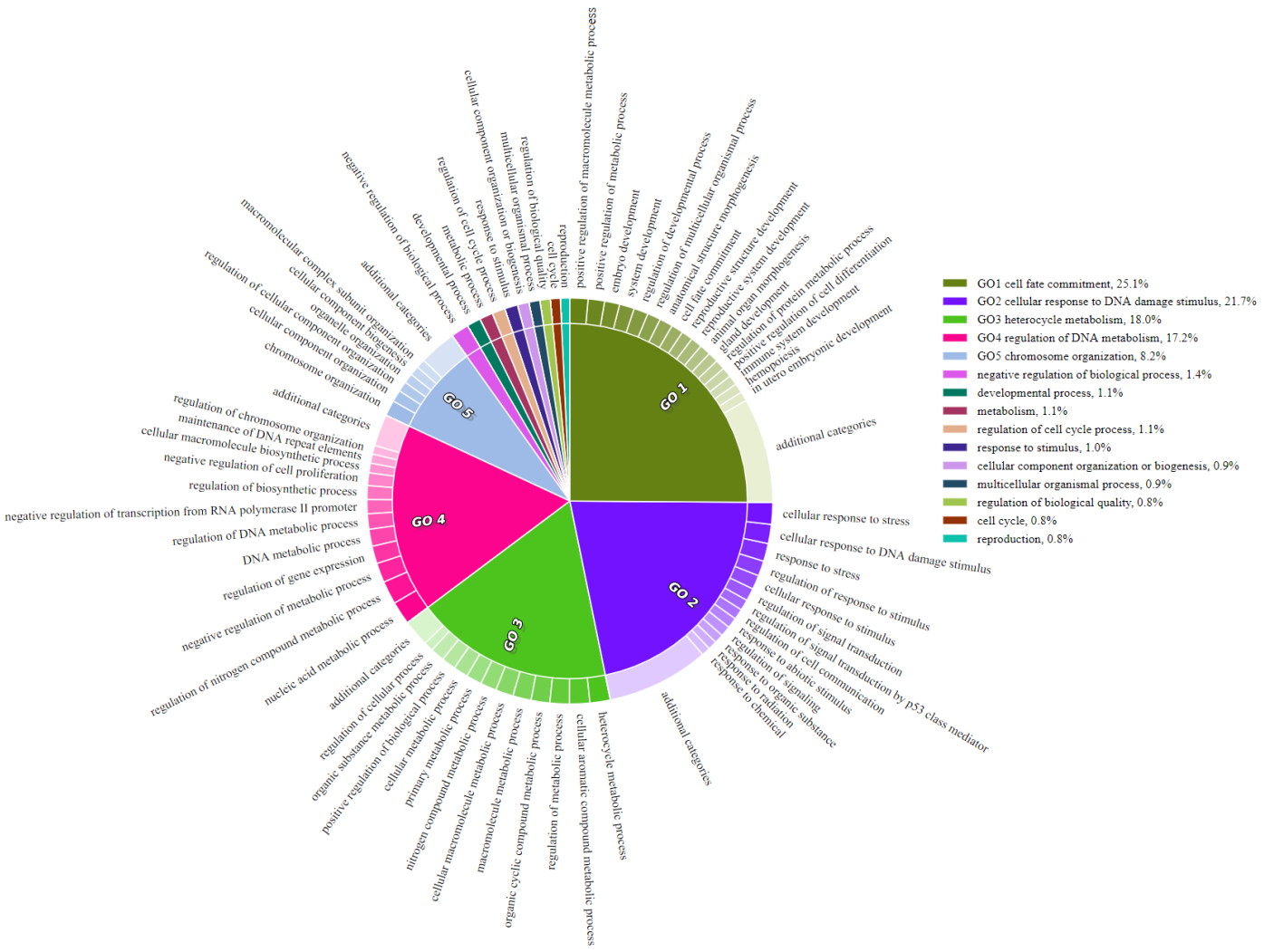

Figure 12: Diagram with the representative GO terms enriched to the gene products in cluster 3. In the legend the labels of the high-level groups are shown including the proportion of the terms that are included in those groups.

Cluster 4

The fourth cluster consisted of only nine gene products with three annotations among two gene products. Therefore, the obtained list of enriched GO terms was not very long and links between different hallmarks could not be discovered.
However, all of the GO terms were children of GO terms that involved either mRNA processing or mRNA splicing. This indicates that mRNA and alterations to mRNA are involved in cancer.

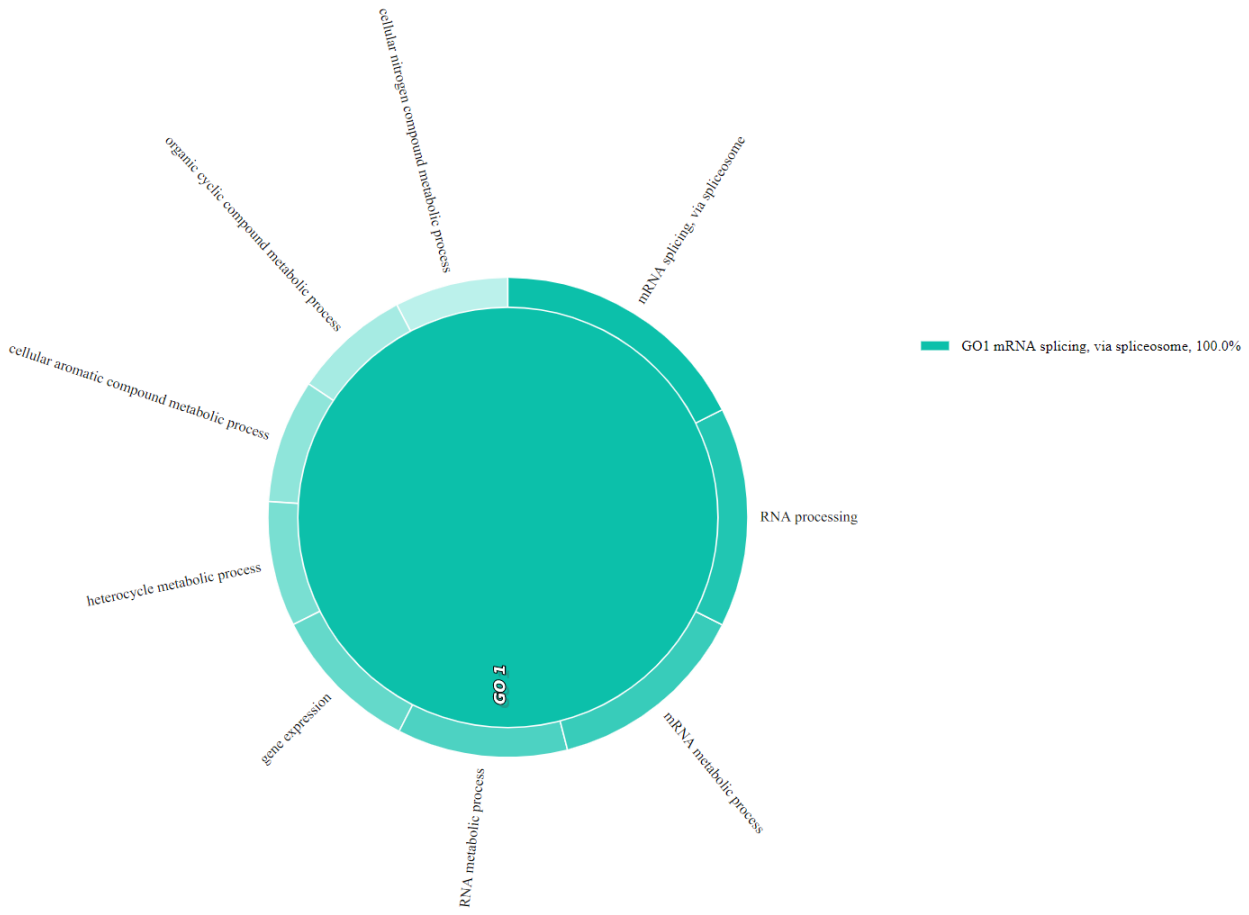

Figure 13: Diagram with the representative GO terms enriched to the gene products in cluster 4. In the legend the labels of the high-level groups are shown including the proportion of the terms that are included in those groups.

Cluster 5

The representative GO terms from the first group, "hemopeiosis" (22.5%), in the diagram in Figure 14 could be linked to different hallmarks. "Immune system development", "cell differentiation", "cell fate commitment" and "anatomical structure formation involved in morphogenesis" are examples of this varied group.
Where the GO terms in the second group seem to be too general too be linked to one of the hallmarks, almost all terms in groups 3, labeled with "transcription elongation from RNA polymerase II promoter", and 4, labeled with "chromatin organisation", could be linked to hallmark 8, genome instability mutations. With 32.5% of all representative GO terms in those groups combined and their linkage to hallmark 8, this seemed to be a clear overrepresentation of this hallmark.
It should be noted that the other terms in the diagram are varied. As in cluster 1, the remaining terms could be linked to several hallmarks, which shows the interconnection between the hallmarks in this cluster.

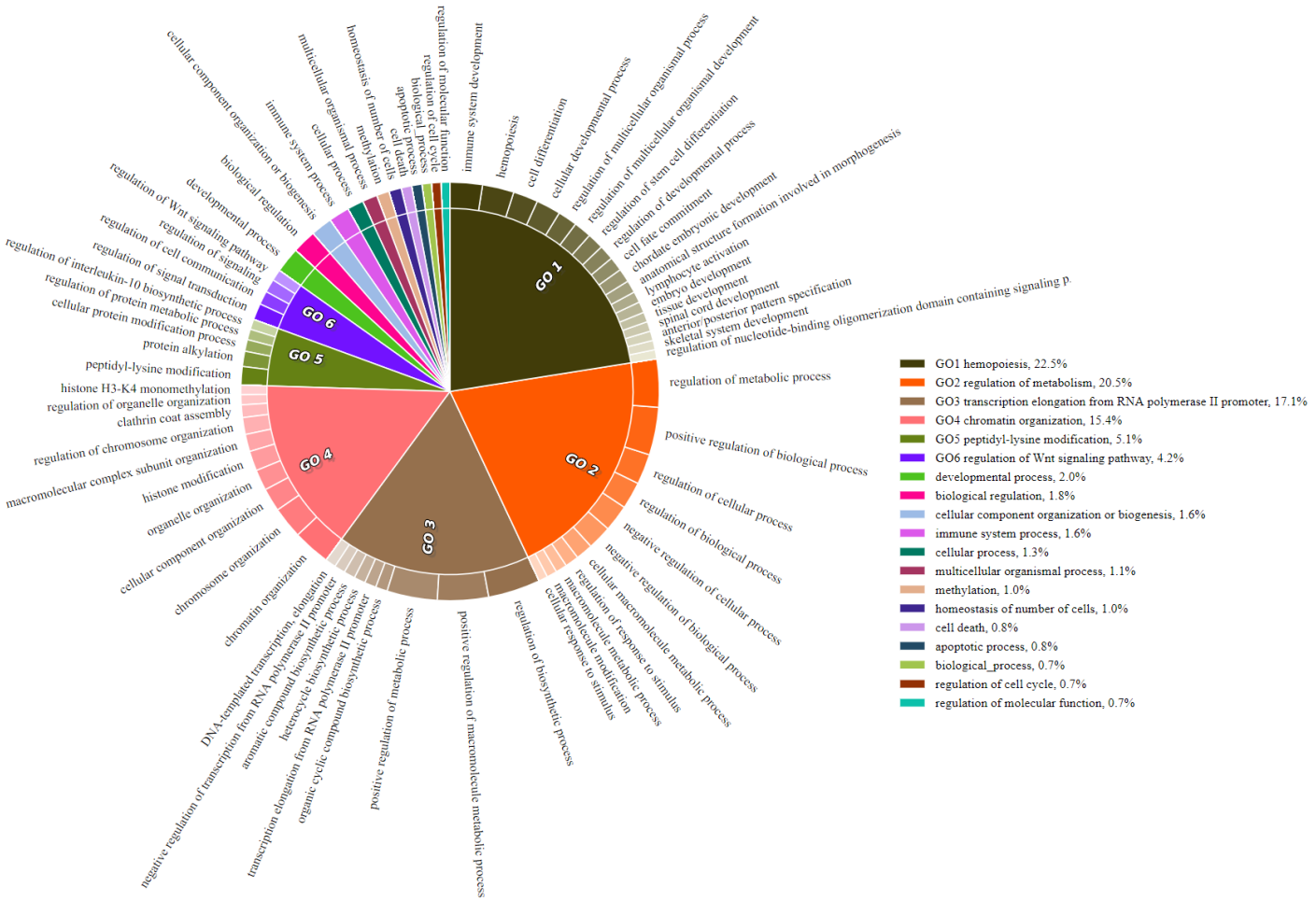

Figure 14: Diagram with the representative GO terms enriched to the gene products in cluster 5. In the legend the labels of the high-level groups are shown including the proportion of the terms that are included in those groups.

Cluster 6

Including 55.4% of the representative GO terms, the first group, labeled with "regulation of MAPK cascade", seemed to represent hallmark 1, sustaining proliferative signaling. There are, namely, many terms that involve signal transduction. However, this group also includes many terms that could be linked to different hallmarks, such as "negative regulation of transcription from RNA polymerase II promoter", "primary metabolic process" and "negative regulation of apoptotic process". This group, thus, shows a slight overrepresentation of hallmark 1 in cluster 6, but also the interconnection between the different hallmarks. The latter is strengthened by the fact that the remaining 44.6% of the representative GO terms also did not show a strong link to one of the hallmarks, but rather to a large part of the hallmarks.

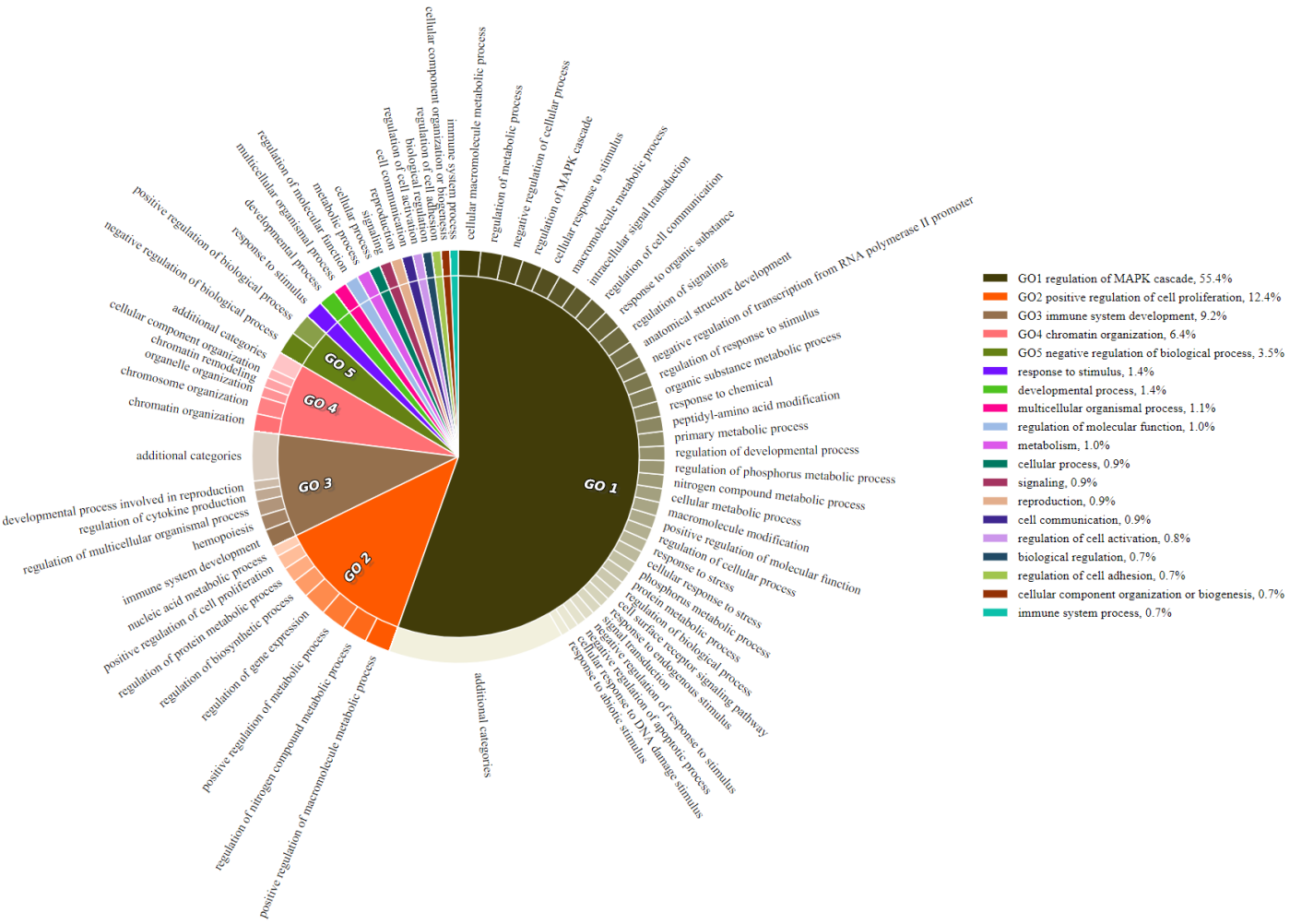

Figure 15: Diagram with the representative GO terms enriched to the gene products in cluster 6. In the legend the labels of the high-level groups are shown including the proportion of the terms that are included in those groups.

3.5 Cliques

For each cluster, several maximal cliques were found using MClique. Since the goal is to predict hallmark annotations for the unannotated gene products, cliques that contain gene products that are annotated with the same hallmark(s) are suitable for analysis. Because the list of discovered maximal cliques is long in some cases, a program in Python (avalaible upon request) was written to find the suitable cliques. In Table 5 the amount of discovered maximal cliques and the amount of maximal cliques where the same hallmark(s) was/were annotated to all the annotated gene products in the clique can be found for each cluster.

Table 5: The amount of cliques identified by MClique for each cluster and the amount of cliques in which the annotated gene products all had an annotation for the same hallmark.

Cluster	Maximal cliques	Fully annotated maximal cliques
1	1333	348
2	185	127
3	72	54
4	1	1
5	63	34
6	55	36

Many cliques show the same unannotated gene products as other cliques. In general, the cliques with different unannotated gene products were selected. Also, each time the largest cliques that cover the not yet analysed unannotated gene products were selected.

3.6 Semantic Similarity between Gene Products

For each selected clique, the semantic similarity between all pairs of gene products was calculated using Wang's method and the best-match average combination method with the GOSemSim package in R. Furthermore, ten random samples of size 15 were selected from the whole network. For these samples, the pairwise semantic similarity was also calculated to compare with the cliques. The average semantic similarity of all gene products in these samples are shown in Table 6 and the full pairwise semantic similarity scores can be found in the supplementary materials.
Because of the fact that the biological processes of gene products are more closely studied and more annotated to gene products, this ontology was used in the first place to discover similarities. However, after similarities are discovered this way, the molecular function and cellular component similarity scores were also looked at before predicting hallmark annotations.
With the lowest and highest average pairwise semantic similarity, regarding biological processes, in the samples being 0.233 and 0.363, respectively, and the average of the samples being 0.308, there is an indication of how similar randomly selected cancer gene products are. This can be used for comparison with the selected cliques and their gene products.
However, this comparison could only be used to discover if the gene products in the cliques were more similar than randomly selected gene products, but not to predict hallmark annotations. To be as certain as possible about the prediction of hallmark annotations for specific gene products, a

Table 6: The averages of the pairwise semantic similarities between the gene products in the random samples using their biological process terms, molecular function terms and cellular component terms.

	Average Similarity		
Sample	BP	MF	CC
1	0.299	0.396	0.635
2	0.251	0.539	0.593
3	0.292	0.589	0.612
4	0.324	0.522	0.522
5	0.233	0.498	0.565
6	0.316	0.468	0.622
7	0.329	0.522	0.597
8	0.350	0.583	0.704
9	0.318	0.567	0.714
10	0.363	0.503	0.560
Average	0.308	0.519	0.612

high threshold of 0.650 was chosen.
It should be noted that the semantic similarities based on the molecular function and cellular component annotations are higher than that of the biological process annotations. This is a consequence of the fact that the biological process ontology contains a lot more terms, as discussed in Section 2.

The unannotated gene products with a semantic similarity score above the threshold are discussed. In order to avoid redundant results, only the unannotated gene products that showed significantly high semantic similarity are discussed and shown in tables. All the analysed cliques can, however, be found in the supplementary materials. The numbers of the cliques follow the enumeration as MClique generated them.

Cliques from Cluster 1

As a result of the fact that the first cluster is the most densely connected, a lot of (relatively large) cliques have been identified by MClique. For 23 cliques, the pairwise semantic similarities between its gene products were calculated and analysed. The results of unannotated gene products that show significantly high similarity with annotated gene products are summarised in Table **??**.

As can be seen in Table 7 there are three different gene products that show significantly high semantic similarity with an annotated gene product in the same clique. AKT2 and MAP2K1 show significant semantic similarity in two different cliques with different hallmark annotations, but with the same gene products. In both cliques, however, their average semantic similarities with all annotated gene products in the cliques are also relatively high, compared to the random samples. CDKN1B and CDKN1A show significantly high semantic similarity with CDKN2A and, in addition, a relatively high average semantic similarity with all other annotated gene products.
Therefore, CDKN1B and CDKN1A could be predicted to be annotated with hallmark 6, invasion

Table 7: The not yet annotated gene products in cliques obtained from the first cluster that have a significantly high semantic similarity with at least one of the annotated gene products in the same clique are shown. The size of those cliques, with the amount of annotated gene products between brackets, are shown together with the hallmark that is annotated to all annotated gene products. In the last column the average of the pairwise semantic similarities between the unannotated gene product and all annotated ones is shown. *Gene product that was not annotated yet, but could be predicted to be annotated and was therefore also used for comparing with the unannotated gene products.

Clique	Size	Hallmark	Gene product	Annotated gene product	SemSim	Avg. SemSim w/ annotated products
20	21(13)	6	AKT2	AKT1	0.658	0.476
			CDKN1B	CDKN2A	0.701	0.463
			MAP2K1	MAPK1	0.713	0.536
34	21(13)	6	CDKN1A	CDKN2A	0.697	0.520
			CDKN1A	CDKN1B*	0.725	
649	17(10)	10	AKT2	AKT1	0.658	0.486
973	16(9)	10	MAP2K1	MAPK1	0.713	0.526

and metastasis. AKT2 and MAP2K1 could be predicted to be annotated with hallmark 6, as well as with hallmark 10, deregulating cellular energetics.

Cliques from Cluster 2

The second cluster was also very densely connected, resulting in a lot of identified cliques. Many of these cliques have been analysed, and the significantly high similarities between unannotated gene products and annotated gene products are denoted in Table 8.

Table 8: The not yet annotated gene products in cliques obtained from the second cluster that have a significantly high semantic similarity with at least one of the annotated gene products in the same clique are shown. The size of those cliques, with the amount of annotated gene products between brackets, are shown together with the hallmark that is annotated to all annotated gene products. In the last column the average of the pairwise semantic similarities between the unannotated gene product and all annotated ones is shown.

Clique	Size	Hallmark	Gene product	Annotated gene product	SemSim	Avg. SemSim w/ annotated products
1	15(9)	1	PDGFRB	EGFR	0.656	0.513
			PDGFRB	HRAS	0.662	
			STAT5B	JAK1	0.668	0.480
17	13(9)	1 & 9	STAT5B	JAK1	0.668	0.502
20	13(8)	1 & 9	PDGFRB	EGFR	0.656	0.585
			PDGFRB	HRAS	0.662	
			PDGFRB	PDGFRA	0.791	
50	12(8)	1 & 9	PDGFB	EGFR	0.657	0.595
			PDGFB	ERBB2	0.667	
122	9(6)	1 & 7	PDGFRB	HRAS	0.662	0.540
67	11(3)	8 & 9	FEN1	FANCD2	0.672	0.629
			FEN1	RECQL4	0.689	
			POLD1	ATR	0.655	0.696
			POLD1	FANCD2	0.718	
			POLD1	RECQL4	0.716	
			POLE	RECQL4	0.691	0.613
			RAD51B	RECQL4	0.655	0.565
			XPA	FANCD2	0.792	0.699
			XPA	RECQL4	0.692	
104	10(4)	8 & 9	FEN1	ERCC2	0.661	0.637
			POLD1	ERCC2	0.693	0.696
			XPA	ERCC2	0.717	0.703
125	9(7)	9	SMARCB1	ARID1A	0.662	0.391

PDGFRB has shown significant similarity with three different gene products in clique 20, where the annotated gene products were all annotated with hallmark 1 and 9. The average semantic similarity between PDGFRB and all other gene products in this clique was also high, 0.585. It also occured

in the clique annotated with hallmark 1 and 7, but the average semantic similarity between this gene product and all annotated gene products was lower here.
STAT5B shows significantly high semantic similarity with JAK1 in clique 1 and 17. Since the average semantic similarity between STAT5B and the annotated gene products in cluster 17 was higher than in cluster 1, this gene product is likely to be annotated with hallmark 1 and 9, sustaining proliferative signaling and resisting cell death, respecitvely.
Furthermore, PDFB has also shown high semantic similarity with gene products annotated with these hallmarks.
Clique 67 showed a lot of high semantic similarity scores. FEN1, POLD1 and XPA were significantly similar to several gene products. POLE and RAD51B showed significantly high similarity with RECQL4. These unannotated gene products also showed very high semantic similarity among eachother. Finally, in clique 104, also annotated with hallmark 8 and 9, it is shown that FEN1, POLD1 and XPA show significant similarity with ERCC2.
In clique 125, SMARCB1 is shown to be significantly similar with ARID1A. However, this is not the case with all annotated gene products, as the average of the semantic similarity with these gene products is 0.391.

Cliques from Cluster 3

The cliques identified in the third cluster are small compared to the previously discussed cliques, as can be seen in Table 9. However, several unannotated gene products in the cliques showed significantly high semantic similarities with annotated gene products.
First of all, FANCC showed a similarity of 0.658 with ERCC4 in clique 3, of which the three annotated gene products had an annotation with hallmark 8. The average similarity with the other two annotated gene products included was 0.616. This was considered high enough to predict an annotation with hallmark 8 for FANCC.
PALB2 had a significantly high similarity with all three annotated gene products in this clique. Moreover, PALB2 had a maximum similarity with BLM and ERCC4 and a similarity of 0.902 with BRIP1. WRN showed high similarity with BLM, 0.736, and with all annotated gene products on average, 0.641. Furthermore, WRN had a maximum similarity with PALB2, which was previously identified as being highly similar with all annotated gene products.
In clique 5, which was also annoted with hallmark 8 among all annotated gene products, NBN showed a significantly high semantic similarity with BLM and BRIP1, 0.720 and 0.661 respectively. This gene product also showed a high average similarity with all annotated gene products and a maximum similarity with PALB2.
Clique 7 consisted of several highly similar gene products. PMS1 showed significantly high similarity with all three annotated gene products. The average similarity with those gene products was 0.949. MUTYH showed significantly high similarity with ERCC4 and maximum similarity with PMS1. WRN also was considered highly similar to one of the annotated gene products and to PMS1. Both MUTYH and WRN also showed high similarity on average with the three annotated gene products, compared to the average similarity between random gene products in the network.
RMI2 scored a significantly high similarity of 0.771 with BRIP1 in clique 8, of which the three annotated gene products had an annotation for hallmark 8. Including the other annotated gene products, RMI2 also scored a high average similarity. Moreover, it scored a similarity of 0.879 with

Table 9: The not yet annotated gene products in cliques obtained from cluster 3 that have a significantly high semantic similarity with at least one of the annotated gene products in the same clique are shown. The size of those cliques, with the amount of annotated gene products between brackets, are shown together with the hallmark that is annotated to all annotated gene products. In the last column the average of the pairwise semantic similarities between the unannotated gene product and all annotated ones is shown. *Gene product that was not annotated yet, but could be predicted to be annotated and was therefore also used for comparing with the unannotated gene products.

Clique	Size	Hallmark	Gene product	Annotated gene product	SemSim	Avg. SemSim w/ annotated products
3	9(3)	8	FANCC	ERCC4	0.658	0.616
			PALB2	BLM	1.000	0.967
			PALB2	BRIP1	0.902	
			PALB2	ERCC4	1.000	
			WRN	BLM	0.736	0.641
			WRN	PALB2*	1.000	
5	8(3)	8	NBN	BLM	0.720	0.669
			NBN	BRIP1	0.661	
			NBN	PALB2*	1.000	
7	8(3)	8	PMS1	ERCC3	0.963	0.949
			PMS1	ERCC4	0.963	
			PMS1	ERCC5	0.920	
			MUTYH	ERCC4	0.662	0.624
			MUTYH	PMS1*	1.000	
			WRN	ERCC3	0.676	0.626
			WRN	PMS1*	0.963	
8	8(3)	8	RMI2	BRIP1	0.771	0.658
			RMI2	PALB2*	0.879	
21	6(2)	1 & 6 & 9	PHF6	CREBBP	1.000	1.000
			PHF6	DNMT3A	1.000	

PALB2, previously shown to be very similar with several gene products annotated with hallmark 8. Finally, clique 21 consisted of six gene products of which only two were annotated. However, PHF6 showed maximum similarity with both of these gene products, which, in combination with the fact that gene products in a clique are very likely to be related, indicated that this gene product could be annotated with hallmark 1, 6 and 9.

Cliques from Cluster 4

The fourth cluster was very small and, as a result, only one maximal clique could be identified. This clique consisted of eight out of the nine gene products in the cluster. The two annotated gene products didn't have a similar hallmark annotation. As a consequence, the prediction of hallmark annotation didn't seem to be possible. However, CRNKL1 and PCBP1 did show significantly high similarity with SF3B1 and with eachother, as can be seen in Table 10 (where the column with the average simliarity between the unannotated gene product and the annotated gene product was not filled in due to the fact that the annotated gene products didn't have a similar annotation). In this way, CRNKL1 and PCBP1 were considered to be annotated with hallmark 9, the annotation of SF3B1.

Table 10: The not yet annotated gene products in cliques obtained from cluster 4 that have a significantly high semantic similarity with at least one of the annotated gene products in the same clique are shown. The size of those cliques, with the amount of annotated gene products between brackets, are shown together with the hallmark that is annotated to all annotated gene products. In the last column the average of the pairwise semantic similarities between the unannotated gene product and all annotated ones is shown. *Gene product that was not annotated yet, but could be predicted to be annotated and was therefore also used for comparing with the unannotated gene products.

Clique	Size	Hallmark	Gene product	Annotated gene product	SemSim	Avg. SemSim w/ annotated products
1	8(2)	9	CRNKL1	SF3B1	0.747	
			PCBP1	SF3B1	0.736	
			CRNKL1	PCBP1*	0.789	

Cliques from Cluster 5

As a result of the fact that cluster 5 was less densely connected than the first three clusters, there were also less and smaller cliques identified in this cluster. However, some interesting results were found. SMARCE1 and SMARCD1 both showed significantly high similarity with gene products annotated with hallmark 8, CHD4 and SHD4 respectively, across two cliques. Furthermore, they scored a similarity among eachother of 0.927, and were, thus, both considered to be annotated with hallmark 8.
In clique 7, BRD3 showed a similarity of 0.843 with BRD4, which was annotated with hallmarks 1, 4 and 9. Due to the fact that these gene products belong to the same family, a similar hallmark annotation could be predicted.

Table 11: The not yet annotated gene products in cliques obtained from cluster 5 that have a significantly high semantic similarity with at least one of the annotated gene products in the same clique are shown. The size of those cliques, with the amount of annotated gene products between brackets, are shown together with the hallmark that is annotated to all annotated gene products. In the last column the average of the pairwise semantic similarities between the unannotated gene product and all annotated ones is shown. *Gene product that was not annotated yet, but could be predicted to be annotated and was therefore also used for comparing with the unannotated gene products.

Clique	Size	Hallmark	Gene product	Annotated gene product	SemSim	Avg. SemSim w/ annotated products
40	3(1)	8	SMARCE1	CHD4	0.771	0.771
44	3(1)	8	SMARCD1	SHD4	0.714	0.714
			SMARCD1	SMARCE1*	0.927	
7	4(1)	1 & 4 & 9	BRD3	BRD4	0.843	0.843

Cliques from Cluster 6

Finally, several cliques in cluster 6 were identified and the semantic similarities between its gene products were calculated. In clique 2, KDM5C showed significantly high similarity with KDM6A. The average semantic similarity between KDM5C and the four gene products annotated with hallmark 6 was 0.464. Based on the fact that KDM5C and KDM6A belong to the same family and have a high semantic similarity, KDM5C was predicted to be annotated with hallmark 6.
ARID1B in clique 6 showed a very high similarity of 0.836 with KMT2D. Furthermore, the average semantic similarity between ARID1B and the three annotated gene products in this clique. As a result, ARID1B was predicted to be annotated with hallmark 6.

Table 12: The not yet annotated gene products in cliques obtained from the first cluster that have a significantly high semantic similarity with at least one of the annotated gene products in the same clique are shown. The size of those cliques, with the amount of annotated gene products between brackets, are shown together with the hallmark that is annotated to all annotated gene products. In the last column the average of the pairwise semantic similarities between the unannotated gene product and all annotated ones is shown.

Clique	Size	Hallmark	Gene product	Annotated gene product	SemSim	Avg. SemSim w/ annotated products
2	6(4)	6	KDM5C	KDM6A	0.672	0.464
6	5(3)	6	ARID1B	KMT2D	0.836	0.529

4 Conclusion and Discussion

In this thesis, the hallmarks of cancer were investigated using a network of cancer genes. This was done with the goals of discovering connections in terms of overlaps between the hallmarks and predicting hallmark annotations for cancer genes.

After analysing the topology of the PPI network of the 716 gene products it could be concluded that the network was very densely connected, as follows from the average degree of 42.2 and small average shortest path lengths. Most of the 267 annotated gene products had an annotation with more than one hallmark, which indicated the possibility of connections between hallmarks. The cluster statistics also showed dense connections between the gene products. Multiple hallmarks were covered in each cluster, which again showed possible interconnections between the hallmarks on basis of the annotated gene products. Some hallmarks were more represented than others in each cluster, indicating connections between specific hallmarks. To determine whether this was the case, the gene products that have not yet been annotated were taken into account by linking the biological processes, regulated by all gene products in each cluster, to hallmarks and in this way determine if and which hallmarks were overrepresented and, thus, linked in each cluster.

4.1 Connections between Hallmarks

In biology, gene products in clusters are likely to form complexes and, thus, the clusters of related gene products were used to discover connections between hallmarks. This was done by performing Gene Ontology enrichment analysis, after which the lists of enriched GO terms were summarised using REVIGO. By visualising the remaining representative GO terms with CirGO, the results showed several overrepresented GO biological processes that could be linked to different hallmarks. In this way, overlaps and, thereby, connections between hallmarks were discovered.

First of all, the first cluster showed overrepresentation of GO terms that could be linked to sustaining proliferative signaling, evading growth suppressors and deregulating cellular energetics, hallmarks 1,2 and 10, respectively.
The GO enrichment analysis that was performed on the second cluster showed partially similar overrepresented biological processes as compared to the first cluster. Here, however, the evidence for a link between the overrepresented GO biological processes and hallmark 2 was not strong enough to call this hallmark overrepresented. Hallmark 1, sustaining proliferative signaling, hallmark 7, inducing angiogenesis, and hallmark 10, deregulating cellular energetics, did seem to be overrepresented in this cluster.
The representative GO terms enriched to the gene products in the third cluster could, for a large part, also be linked to hallmark 10, deregulating cellular energetics. Furthermore, there was also an overrepresentation of representative GO terms that could be linked to hallmark 8, genome instability and mutations. Hallmark 6, activating invasion and metastasis, also seemed to be well represented by the gene products in this cluster, but the evidence was considered too weak to call this hallmark overrepresented.
Cluster 4 was very small and was, thus, enriched with a small amount of GO terms. Nonetheless, those GO terms were interesting since they were all involved in mRNA, especially in its splicing and processing. This implies that mRNA and/or the processing and splicing of it plays a role in cancer.
Cluster 5 and 6 were less densely connected than the other clusters. Cluster 5 did show an overrepresentation of terms that could be linked to hallmark 8, genome instability and mutations. Cluster 6 showed a slight overrepresentation of GO terms that could be linked to hallmark 1. The GO terms in those clusters were, however, too varied to show an overrepresentation of biological processes that could be linked to multiple hallmarks of cancer. This might be a result of the interconnections between all (or a large part of) the hallmarks. However, this could a consequence of the fact that the gene products in this cluster are not as related as in the other clusters, which would cause more varied processes in which the gene products are involved.

Interpreting these results and referring back to the first research question, several connections between the hallmarks of cancer have been discovered. Hallmark 1, sustaining proliferative signaling, and 10, deregulating cellular energetics, showed this connection in both cluster 1 and 2. Furthermore, on basis of the results from cluster 1, these hallmarks also showed a connection with hallmark 2, evading growth suppressors, and on basis of cluster 2, also a connection with hallmark 7, inducing angiogenesis, seems to exist. The gene products in those clusters were very related to each other as a result of the network topology. As a consequence, the overlaps between hallmarks 1, 2 and 10 and between hallmarks 1, 7 and 10 indicate that cancer genes that are involved in one of those hallmarks are directly or indirectly involved in the other connected hallmarks because of their

relatedness to genes involved in the other hallmarks. For sustaining proliferative signaling and evading growth suppressors (hallmarks 1 and 2), this connection seems very logical, since both capabilities are processes involved in the growth and proliferation of cancer cells. However, for hallmark 7, inducing angiogenesis, this connection is a less obvious. The connection between the three hallmarks and hallmark 10, deregulation of cellular energetics, implies that the genes that regulate growth and angiogenesis are also connected with the adjustment of the metabolism to provide enough energy to succeed in growing and creating blood vessels.
The found connections between the hallmarks and more importantly between the genes involved in those hallmarks, could be of importance for determining which capabilities are already acquired by cancer cells and for the treatment of cancer. When it is found that cancer cells have acquired one of the connected hallmarks, the other hallmarks are very likely to also be acquired by those cells. Furthermore, action could then be taken against the processes involved in those connected hallmarks, instead of against only one hallmark.

It would be interesting to use these discovered overlaps to find out if there is a general order in which the hallmarks of cancer arise in the multi-step process of cancer development. This might be done by including pathway information of the gene products in the network. In this way, a directed PPI network could be constructed, where pathways that show involvement in different hallmarks might show overlaps.

4.2 Hallmark Annotation Predictions

For each identified cluster, cliques of gene products were found. Since gene products in a clique can interact with all other gene products in the same clique, they show high relatedness regarding the processes in which they are involved. To strengthen this and to discover significantly high similarities between gene products in the cliques, the pairwise semantic similarities between gene products in each clique were calculated. On basis of these results, gene products of which the role in cancer is not yet known, i.e. unannotated gene products, were predicted to be annotated with hallmarks of gene products in the same clique with which they showed significantly high similarity. In this way, hallmark annotations were predicted for several genes, listed in Table 13.
The second goal of this thesis was to predict hallmark annotations. Using network analysis, existing hallmark annotations and a semantic similarity measure, 41 annotations among 27 genes were predicted. These hallmark annotations could be confirmed in research, by discovering the specific biological functions of these genes in cancer tissue. Predictions for 27 out of the 449 not yet annotated genes is not a lot, but the confidence of these predictions is very high due to the stringent cut-off for the semantic similarities.

Table 13: Genes for which a hallmark annotation was predicted.

Gene	Predicted hallmark annotation
CDKN1B	6
CDKN1A	6
AKT2	6 & 10
MAP2K1	6 & 10
PDGFRB	1 & 9
STAT5B	1 & 9
PDGFB	1 & 9
FEN1	8 & 9
POLD1	8 & 9
POLE	8 & 9
RAD51B	8 & 9
XPA	8 & 9
FANCC	8
PALB2	8
WRN	8
NBN	8
MUTYH	8
PMS1	8
RMI2	8
PHF6	1 & 6 & 9
CRNKL1	9
PCBP1	9
SMARCD1	8
SMARCE1	8
BRD3	1 & 4 & 9
KDM5C	6
ARID1B	6

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA: A Cancer Journal for Clinicians*, vol. 69, no. 1, pp. 7–34, 2019.

[2] A. Efeyan and M. Serrano, "p53: Guardian of the Genome and Policeman of the Oncogenes," *Cell Cycle*, vol. 6, no. 9, pp. 1006–1010, 2007. PMID: 17457049.

[3] Y. Yin and W. H. Shen, "PTEN: a new guardian of the genome," *Oncogene*, vol. 27, pp. 5443–5453, 2008.

[4] ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, "Pan-cancer analysis of whole genomes," *Nature*, vol. 578, no. 7793, pp. 82–93, 2020.

[5] T. Zhan, N. Rindtorff, J. Betge, M. P. Ebert, and M. Boutros, "CRISPR/Cas9 for cancer research and therapy," *Seminars in Cancer Biology*, vol. 55, pp. 106 – 119, 2019. Translational Cancer Genomics.

[6] D. Hanahan and R. A. Weinberg, "The Hallmarks of Cancer," *Cell*, vol. 100, no. 1, pp. 57–70, 2000.

[7] D. Hanahan and R. A. Weinberg, "Hallmarks of Cancer: The Next Generation," *Cell*, vol. 144, no. 5, pp. 646 – 674, 2011.

[8] W. Zhang and H. T. Liu, "MAPK signal pathways in the regulation of cell proliferation in mammalian cells," *Cell Research*, vol. 12, p. 9–18, 2002.

[9] J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, P. Fish, B. Harsha, C. Hathaway, S. C. Jupe, C. Y. Kok, K. Noble, L. Ponting, C. C. Ramshaw, C. E. Rye, H. E. Speedy, R. Stefancsik, S. L. Thompson, S. Wang, S. Ward, P. J. Campbell, and S. A. Forbes, "COSMIC: the Catalogue Of Somatic Mutations In Cancer," *Nucleic Acids Research*, vol. 47, no. D1, pp. D941–D947, 2018.

[10] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes, "The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers," *Nature Reviews Cancer*, vol. 18, pp. 696–705, 2018.

[11] M. W. Gonzalez and M. G. Kann, "Chapter 4: Protein interactions and disease," *PLOS Computational Biology*, vol. 8, no. 12, pp. 1–11, 2012.

[12] P. Uetz, L. Giot, and G. Cagney, "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae," *Nature*, vol. 403, pp. 623–627, 2000.

[13] R. Albert, "Scale-free networks in cell biology," *J Cell Sci.*, vol. 118(Pt 21), pp. 4947–4957, 2005.

[14] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular systems biology*, vol. 3, p. 88, 2007.

[15] C. D. Nguyen, K. J. Gardiner, and K. J. Cios, "Protein annotation from protein interaction networks and gene ontology," *Journal of Biomedical Informatics*, vol. 44, no. 5, pp. 824 – 829, 2011.

[16] G. Wu, X. Feng, and L. Stein, "A human functional protein interaction network and its application to cancer data analysis," *Genome Biology*, vol. 11(R53), 2010.

[17] G. Kar, A. Gursoy, and O. Keskin, "Human cancer protein-protein interaction network: A structural perspective," *PLOS Computational Biology*, vol. 5, no. 12, pp. 1–18, 2009.

[18] E. Ravasz and A.-L. Barabási, "Hierarchical organization in complex networks," *Phys. Rev. E*, vol. 67, p. 026112, 2003.

[19] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker, "Cytoscape 2.8: new features for data integration and network visualization," *Bioinformatics*, vol. 27, no. 3, pp. 431–432, 2010.

[20] N. T. Doncheva, J. H. Morris, J. Gorodkin, and L. J. Jensen, "Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data," *Journal of Proteome Research*, vol. 18, no. 2, pp. 623–632, 2019.

[21] B. Snel, G. Lehmann, P. Bork, and M. Huynen, "STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene," *Nucleic Acids Res.*, vol. 28(18), pp. 3442–3444, 2000.

[22] Y. Assenov, F. Ramírez, S. Schelhorn, T. Lengauer, and M. Albrecht, "Computing topological parameters of biological networks," *Bioinformatics*, vol. 24(2), pp. 282–284, 2008.

[23] A. H. Y. Tong, B. Drees, G. Nardelli, G. D. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, M. Quondam, A. Zucconi, C. W. V. Hogue, S. Fields, C. Boone, and G. Cesareni, "A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules," *Science*, vol. 295, no. 5553, pp. 321–324, 2002.

[24] C. Pizzuti and S. E. Rombo, "Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods," *Bioinformatics*, vol. 30, no. 10, pp. 1343–1352, 2014.

[25] Y. Cho, W. Hwang, and A. Zhang, "Identification of overlapping functional modules in protein interaction networks: Information flow-based approach," pp. 147–152, 2006.

[26] G. Bader and C. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics*, vol. 4(2), 2003.

[27] C. Pizzuti and S. E. Rombo, "Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods," *Bioinformatics*, vol. 30, no. 10, pp. 1343–1352, 2014.

[28] Gene Ontology Consortium, "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Research*, vol. 32, no. $\text{suppl}_1, pp.\ D258 - -D261, 2004$.

[29] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler, "The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology," *Nucleic Acids Research*, vol. 32, pp. D262–D266, 2004.

[30] H. Mi, A. Muruganujan, J. Casagrande, and P. Thomas, "Large-scale gene function analysis with the PANTHER classification system," *Nat Protoc.*, vol. 8(8), pp. 1551–1566, 2013.

[31] C. Pesquita, D. Faria, A. O. Falcão, P. Lord, and F. M. Couto, "Semantic similarity in biomedical ontologies," *PLOS Computational Biology*, vol. 5, no. 7, pp. 1–12, 2009.

[32] A. Ayllón-Benítez, F. Mougin, J. Allali, R. Thiébaut, and P. Thébault, "A new method for evaluating the impacts of semantic similarity measures on the annotation of gene sets," *PLOS ONE*, vol. 13, no. 11, pp. 1–22, 2018.

[33] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," *CoRR*, vol. abs/cmp-lg/9511007, 1995.

[34] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, (San Francisco, CA, USA), p. 296–304, Morgan Kaufmann Publishers Inc., 1998.

[35] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *CoRR*, vol. cmp-lg/9709008, 1997.

[36] A. Schlicker, Domingues, and R. F.S., "A new measure for functional similarity of gene products based on gene ontology," *BMC Bioinformatics*, vol. 7, 302, 2006.

[37] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees.," *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[38] I. Kuznetsova, A. Lugmayr, and S. Siira, "CirGO: an alternative circular way of visualising gene ontology terms," *BMC Bioinformatics*, vol. 20, 84, 2019.

[39] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, 2007.

[40] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang, "GOSemSim: an R package for measuring semantic similarity among GO terms and gene products," *Bioinformatics*, vol. 26, no. 7, pp. 976–978, 2010.

[41] R. Gentleman, V. Carey, and D. Bates, "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, R80, 2004.

[42] C. Pesquita, D. Faria, and H. Bastos, "Metrics for GO based protein semantic similarity: a systematic evaluation," *BMC Bioinformatics*, vol. 9, S4, 2008.