Anàlisi *in silico* de malalties: des de les mutacions fins les xarxes biològiques

Eduard Porta Pardo

Programa de doctorat en biomedicina
Universitat de Barcelona

Anàlisi *in silico* de malalties: desde les mutacions fins les xarxes biològiques

Memòria presentada per Eduard Porta Pardo per optar al grau de doctor per la Universitat de Barcelona

Institut de Medicina Predictiva i Personalitzada del Càncer

Directora	**Co-director**	**Tutor**	**Doctorand**
Ana Maria Rojas Mendoza	Ildefonso Cases Díez	Albert Tauler Girona	Eduard Porta Pardo

Moltes gràcies!

A la tropa de l'institut (Sergi, Jaume, Uri, Albert, Guille, Víctor, Elena, Lourdes…), perquè ja són molts els anys que ens tenim els uns als altres.

Als biotecnloeggs (Vengo, Muesly, Catix, Alba, Bernat, Sergio, Bet, Alba, Marie, Marta, Esther, Andrea, Cristina, Joaquin…) perquè ells millor que ningú són conscients de l'esforç que comporta tot plegat.

Als de l'IMPPC, perquè entre ping-pongs (Carles, Josep, Bernat, Gabriela, Quim…) i xerrades sobre com arreglar el món (Aida, Victor, Judith, Lorena, Guerau, Ana, Ernest…) es treballa molt més a gust.

A Ana y Alonso, por creer en mi cuando mi reto era el "Beginning Perl for Bioinformatics" (al final no ha salido mal del todo el experimento eh? jejeje).

Als meus pares i germà, per tot el que han fet per mi.

A la meva preciosa dona, la Daniela, per haver estat al meu costat quan més ho necessitava, per ser la meva companya fidel i per estar a punt de fer-nos el regal de la nostra vida.

I a tu, que encara no hi ets, però la teva mare i jo t'esperem amb tot la il·lusió del món!

Table of Contents

1. Introduction

1.1. Genes, mutations and diseases

The origins of modern genetics date back to 1866, when Johann Mendel published his work on pea plants and described the basic patterns of inheritance and several seminal concepts such as phenotype, dominance and recessiveness. Some years later it was established that these principles could be applied to the human inheritance of several traits, including diseases such as alkaptonuria[1] (the first genetic disease identified ever, by Sir Archibal Garrod in 1902). However, it wasn't until 1943 that Avery, MacLeod and McCarty identified the DNA as the cellular component holding the genetic information[2]. This discovery, along with that of the DNA's composition by Chargaff in 1948[3] and the DNA's structure in 1953 by Francis, Watson and Crick[4], set the ground for Baglioni to discover alterations in the globin gene causing sickle-cell anemia, the first mutations at the protein level associated to a disease in 1962[5].

The identification of the mutations and genes causing disorders at the DNA level was difficult at first. For example, the first published repository of genetic disorders, *"Mendelian Inheritance in Man"* by Victor McKusick[6], had over 1400 entries but no autosomal loci in its first edition. The development of various biological technologies, including DNA sequencing, southern, northern or western blotting and DNA recombination among others, facilitated the discovery of several types of disorders caused by both germline and somatic mutations.

There are several types of mutations depending on the type and extend of the DNA alteration (table 1). The largest mutations are the loss or gain of complete chromosomes, followed by inversions deletions and duplications of chromosomal regions (that usually involve several genes), or translocations of DNA fragments from one chromosome to another. There are also various types of mutations that involve smaller regions of DNA, such as deletions or insertions of few bases or the substitution of one base for another. This group of smaller mutations can be classified according to the location of the mutation in the gene body into coding and non-coding mutations.

Non-coding mutations are those mutations that are located in genomic regions that are not translated into protein. These mutations are usually associated to disease by disrupting genomic regions that regulate gene expression, such as promoters or enhancers. Another pathogenic mechanism is the disruption of splicing sites that impede the proper splicing of transcripts and, thus, their correct translation into proteins.

Coding mutations affect the region of genes that is translated into protein. Deletions and insertions in coding regions are usually pathogenic by altering the reading-frame of the transcript. Yet, some other remarkable mechanisms have been described, such as the repeated in-frame insertion of triplets that has been associated to neurodegenerative diseases like Huntington's disease[7]. Those diseases are usually caused by the in-frame expansion of a codon (CAG). In fact, clinical features of these diseases, such as their age of onset or their severity, are correlated with the number of expanded triplets[8].

Coding point substitutions can have various effects in the protein. Due to the degeneracy of the genetic code they may simply alter the DNA sequence but not its translation, in which case they are referred to as "synonymous" or "silent". These mutations, while usually are not associated to disease, can alter the splicing of the mRNA[9], the protein translation rate[10] or the RNA structure[11], sometimes leading to the appearance of some disorders. Another option is that they introduce a stop codon. In this case they are called "truncating" or "nonsense" mutations and are usually pathogenic by disrupting the protein. The last possibility is that they change the aminoacid that the codon is coding, in which case one talks about "missense" mutations.

Missense mutations can be pathogenic through several mechanisms. For example, among others, they may alter the proper folding of a protein, disrupt the catalytic site of an enzyme, impede the formation of disulfide bonds, modify an interaction region or disrupt motifs of post-translational modification, such as phosphorylations or glycosylations.

Table 1.- Types of mutations and associated diseases

Extension of DNA affected	Name	Example Diseases
Large	Chromosome gain/loss	Down's syndrome[12], cancer[13], triple X syndrome[14]
	Inversion	Holoprosencephaly spectrum disorder[15]
	Deletion	Wolf-Hirschhorn Syndrome[16], 11q-syndrome[17]
	Translocation	Chronic myeloid leukemia[18]
Small	Insertion	Huntington's disease[19], Tay-Sach's disease[20]
	Deletion	Hypercholesterolemia[21], familial adenomatous polyposis[22]
	Silent	Cystic fibrosis[9,11], Treacher-Collins syndrome[23]
	Nonsense	Beta-thalassemia[24], Breast and Ovarian cancer[25]
	Missense	Sickle-cell anemia[26], Noonan syndrome[27]

It is important to notice that though all the types of mutations can be associated to a disease, most mutations are benign. For example, the most common form of mutations, single-base substitutions, occur at a frequency of about 1 mutation per 200-1000 nucleotides[28], which means that every human genome contains around 6 million such mutations. In order to differentiate between the two, *mutation* is normally used to describe those alterations that are disease-associated and *variation* is used to describe those alterations that are benign.

1.2. Bioinformatics and disease

The publication of the first draft of the human genome[29,30] and further technological developments allowed scientists to use genome-wide technologies, such as microarrays, proteomics, GWAS or whole-genome or exome sequencing. All these genome-wide technologies have proven useful in pinpointing pathogenic mechanisms and identifying pathways involved in a wide variety of phenotypes including cancer[31,32], Crohn's disease[33] or schizophrenia[34]. This is reflected by the exponential growth of the number of entries in two of the most common-used repositories devoted to human diseases: OMIM and COSMIC (figure 1).

The typical output of the aforementioned technologies is usually a large list of genes or mutations that are potentially associated to a given phenotype, which may not be the actual list of those disease-associated. For example, in the case of GWAS, the output is a list of single nucleotide polymorphisms (SNPs) ordered according to their p value of association to the phenotype[35]. Given that not all the SNPs are actually explored in a GWAS, associated SNPs might be simply pointing to a region in linkage disequilibrium where the actual causal mutation is located or even be false positives. This highlights the need of further evidence in order to accurately identify the causal mutation from the whole list. However, since it is unfeasible to obtain experimental evidence for all the genes and mutations from these lists, computational approaches to prioritize them or suggest molecular hypotheses on their underlying pathogenic mechanisms have become a must.

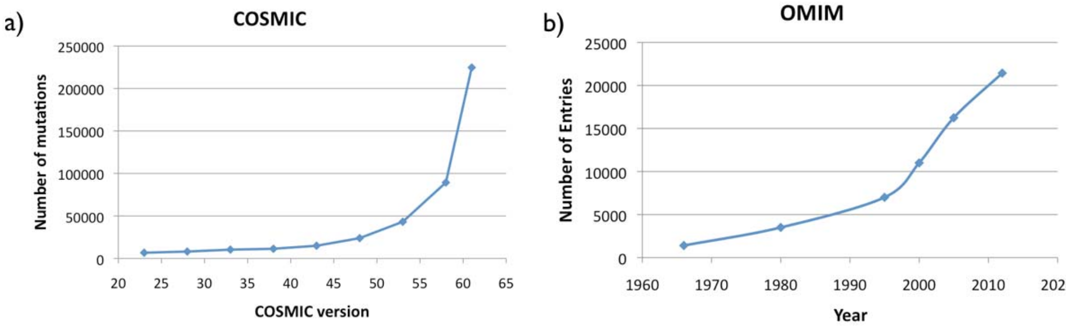


Figure 1.- Data regarding disease associations has experienced an exponential growth in recent years. (a) Number of mutations in each version of the COSMIC database. (b) Number of entries per year in OMIM.

The contributions of computational biology to the study of diseases and their associated genes and mutations are extensive. For example, several groups have developed algorithms that are able to sort pathogenic from neutral mutations based on several features. These include, but are not limited to, their structural properties[36,37], the degree of conservation along the evolution of the affected position[38], the propensity of the mutation to cause changes in the protein stability[39] or a combination of several of these properties[40,41].

An experimental technique that has been proven very useful in identifying the pathogenic mutations involved in mendelian[42] or complex[43,44] disorders is whole-genome or exome sequencing. Several types of cancer have also been sequenced (table 2). In this latter case a biomedical problem similar to that of the analysis of GWAS results arises: the identification of driver mutations. Tumor mutations can be classified into driver or passenger according to their overall contribution to the apparition of cancer. Driver mutations are those that are critical for a tumor to develop, whereas passenger mutations are those cancer-neutral variations retained during the evolution of the tumor[45].

Recent sequencing of genomes from tumor samples have revealed that the number of missense mutations per tumor ranges between 40 and 600 (table 2), though some cases of up to 4000 missense mutations have been described[46]. In order to identify the few driver mutations in each tumor the usual approach is to identify genes that are heavily mutated -by statistical means- along several tumor samples when compared to a random distribution of the mutations[47], or compare the ratio of non-synonymous to synonymous mutations in each gene[48].

The underlying idea in both cases is that, since these genes are systematically affected by missense mutations in cancer, their alteration must be a key event in cancer development. On the other hand, those genes mutated in only a few tumor samples are more likely to contain passenger mutations. Interestingly, a similar approach has been recently used to identify protein domains that are largely mutated in cancer, including the kinase domain, MH2, Miro or APC[49].

Detailed *in silico* studies of mutations can also lead to the generation of hypotheses regarding their molecular pathogenic mechanisms. Shan *et. al.* used computational molecular simulations to discover that the mutation L834R in the EGFR gene, associated to cancer, is pathogenic because it stabilizes an intrinsically disordered region of the protein[50]. This stabilization causes the domain to overdimerize, leading to abnormal activation of proteins downstream in the pathway. Similarly, Fröhling *et. al.* used a combination of *in vitro* and *in silico* experiments in order to identify driver mutations in the kinase FLT3 in acute myeloid leukemia[51]. They used predictions by SIFT[38] and Pmut[52] algorithms to complement results obtained from *in vitro* assays. Bioinformatic results correctly predicted 3 of the 9 driver mutations, whereas in 2 other cases the results of both programs did not agree.

Analysis of pathogenic mutations and genes as a group has also led to important discoveries of the biology of diseases. Torkamani *et. al.* where able to identify regions of the kinase domain that are enriched in disease-related mutations when compared to background missense SNPs[53]. They related this phenomenon to differences in the degree of conservation along the regions and in their overall contribution to the function of the domain. Interestingly in another paper they also found differences in the location within the domain of mutations causing cancer and those causing any other disease[54].

Table 2.- Average number of missense mutations per tumor identified in different types of tumor from several cancer genome sequencing projects

Type of cancer	Average missense mutations per tumor	References
Breast	60	Sjöblom[47]
Colorectal	44 - 170	Sjöblom[47], Nehrt[49]
Large B-cell lymphoma	16	Pasqualucci[55]
Prostate	50-4000	Kumar[46]
Melanoma	643	Hodis[56]
Lung	540	Lee[57]
Ovarian	40	TCGA[58]
Pancreatic	41	Jones[59]
Glioblastoma multiforme	28	Parsons[60]

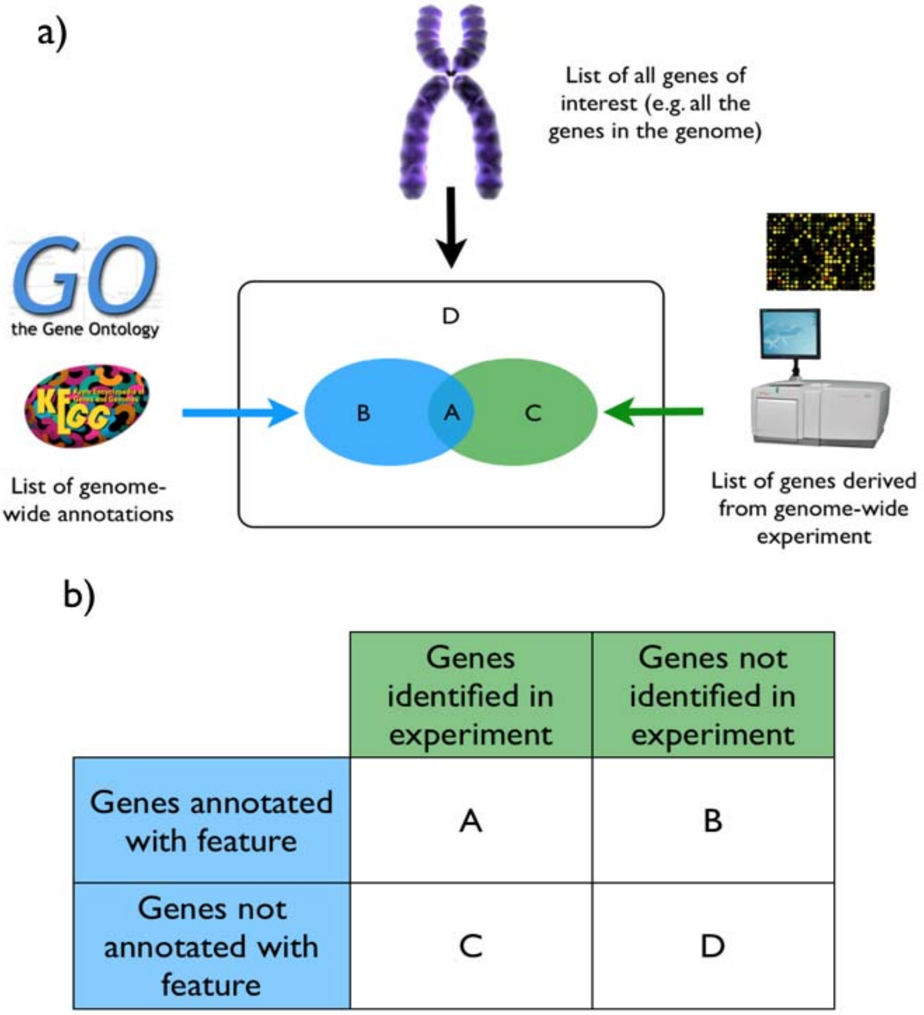
Another fascinating example is the recent work by Vavouri *et. al*[61] describing the correlation between the content of intrinsically disordered regions in a protein and dosage-sensible genes. They observed that genes that are harmful when overexpressed tend to have a high content of protein intrinsic disorder. Their proposed mechanisms is that overexpression of intrinsically disordered proteins is likely to result in interaction promiscuity and appearance of unspecific interactions, thus, causing a malfunction of the cell. Notably, cancer-related genes tend to have longer intrinsically disordered regions[62].

1.3. Enrichment Analysis

Some of the first computational approaches developed to generate hypotheses and analyze genome-wide data were based in enrichment analysis. These approaches rely on the extensive annotation of the human genes with descriptions of the biological features of genes, such as their function or their involvement in biological pathways among many others.

The main idea of the enrichment analysis is that the list of genes derived from a genome-wide experiment may be statistically biased in some of their biological properties (annotations) when compared to a list of genes of interest (usually the whole genome). These biased biological properties will then give a hint on the underlying biology involved in the studied phenotype (figure 3).

In order to identify the biased features the enrichment analysis takes as input genome-wide annotations and a list of genes suspected to be related to the disease (derived from a genome-wide experiment). Then, in its simplest implementation, the Singular Enrichment Analysis (SEA), the analysis compares for each term of the controlled vocabulary the number of genes in the list derived from the experiment that have the annotation with the number that would be expected if the list was chosen randomly from the genome, usually performing a hypergeometric or Fischer's test.


a)
List of all genes of interest (e.g. all the genes in the genome)
GO
the Gene Ontology
KEGG
List of genome-wide annotations
D
B
A
C
List of genes derived from genome-wide experiment
b)

	Genes identified in experiment	Genes not identified in experiment
Genes annotated with feature	A	B
Genes not annotated with feature	C	D

Figure 3.- Description of enrichment analysis to study genome-wide data. (a) General schema of the analysis. First of all a background model is needed (e.g. all the genes in the genome). Then, the distribution of annotations in the background model and in the list of genes of our interest obtained from the genome wide experiment (for example, all the genes above a certain fold-change from a microarray experiment) is compared using statistics such as Fischer's test. This is done by creating the contingency table shown in (b) and applying some method to correct for multiple testing, such as Bonferroni correction

Enrichment analysis relies on extensive gene annotations of biological features. Given that the human genome contains around 20.000 genes, algorithms automating this analysis soon became popular. However, in order to automate this analysis, the annotations must be computer-friendly, thus it is particularly important that they are unique and univocally identified. This need quickly turned scientists to the use of controlled vocabularies to annotate genes.

A controlled vocabulary is a list of terms that describe a realm of knowledge. Each term of the vocabulary can be univocally identified, so that no confusions or miss-annotations can be made. An example of a biologically relevant controlled vocabulary is KEGG[63], which, among other things, provides the terms to describe several protein pathways. Each pathway and gene in KEGG has a unique identifier, which allows its users to recognize it unambiguously (for example, the "Glycolysis pathway" is always identified as "hsa:00010").

While controlled vocabularies solved the automation problem of enrichment analysis, its use has one limitation: when performing genome-wide analysis the p values required for an association to be statistically significant are very high. This is because multiple testing corrections are needed, which makes lack of statistical power an issue. One way to circumvent this problem is to use ontologies to annotate genes. An ontology in computer science is a type of controlled vocabulary defined by a set of terms that describe the domain of knowledge (the different biological processes in a cell for example) and the relationships between them. It can be represented by a directed acyclic graph in which nodes are the terms that belong to the domain of knowledge and edges represent the relationships between the terms (figure 4).

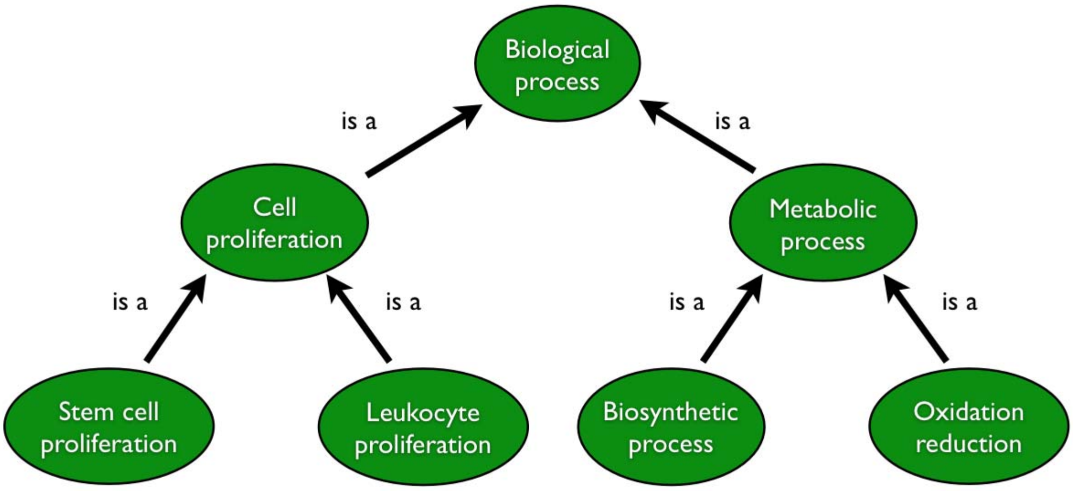


Figure 4.- Representation of a part of the gene ontology using a directed acylic graph. Bubbles represent different terms of the ontology. Arrows go from "child" to "parent" terms. Free text next to arrows explains the type of relationship. Notice that every child term is automatically annotated with all its parents.

This structure of ontologies allows the identification of non-obvious associations by propagating the annotations using the relationships between terms described in the ontology[64]. Propagating the annotations along the ontology increases the statistical power[65] of the analysis. For example if a gene is annotated with the Gene Ontology[66] term "stem cell proliferation" (GO:0072089) we can infer also the annotation of this same gene with the GO term "cell proliferation" (GO:0008283) since the ontology defines that GO:0072089 "is a" GO:0008283 (figure 4). The most used ontology to perform SEA is the Gene Ontology. One of the reasons for that is probably that it has been extensively used to annotate human genes, both manually and electronically, and thus, it has a wide coverage of the genome.

While the basic enrichment analysis, SEA, has proven useful in a variety of contexts several modifications have been developed trying to use more information derived from the experiment. These algorithms are usually referred to as Gene Set Enrichment Analysis[65] (GSEA) and use all the genes and their associated values obtained in a genome-wide experiment (e.g. the fold-change in a microarray experiment) to perform the analysis and obtain the enriched terms[67].

The main idea of the GSEA is that instead of using only the genes with the highest signal (for example, those with a p value < 0.05), it uses the whole list of genes ranked according to their value in the experiment. This allows the use of information from genes that would be simply discarded in the SEA. The maximum enrichment score (MES) of each term of the controlled vocabulary is calculated using the ranked list of genes from the experiment, whereas the enrichment p values of each MES are obtained by comparing the rank with random distributions of the same genes[67] or parametric statistical approaches such as the Z-score[68,69].

Some groups have also used the information from the ontology to improve the results obtained with the SEA. These approaches, usually referred to as Modular Enrichment Analysis (MEA)[65], take advantage of the information regarding the relationships among terms that is intrinsic to the ontology. The main improvement of these algorithms is that by grouping the enriched terms according their biology, in the form of modules, biological patterns that wouldn't be captured or difficult to identify by SEA may emerge[70,71].

The extensive annotation of human genes with GO terms and their use in enrichment analyses with data derived from genome-wide experiments, have certainly provided some insights into the biology of diseases. For example, López-Bigas *et. al.* observed differences in the Biological processes and Molecular functions of genes associated to different types of diseases[72]. For example, they observed that genes involved in transport are overrepresented in metabolic disorders, but underrepresented in cancer. Hence, SEA has been particularly useful in the case of complex phenotypes, where focusing on single genes can be misleading since epistatic interactions are overlooked. Moreover, the aggregation of information makes the ontology useful in cases where epistatic associations may exist such as atherosclerosis[73] or inflammatory bowel disorders[74].

Cancer is likely the disease that has profited the most from the use of ontologies to analyze its lists of associated genes in enrichment analyses[75,76,77] where various GO terms, such as "Apoptosis", "DNA repair", "Intracellular signaling", have been identified to be overrepresented in cancer-causing genes and other GO terms, such as “Transporter” or “metabolism” are underrepresented. These associations have been used as predictors to identify new cancer-related genes or to infer new functional annotations in known cancer-associated genes[77].

Enrichment analysis can be performed using any ontology. However, though there are over 100 different ontologies listed in the OBO foundry[78], few attempts to perform enrichment analysis using other ontologies than GO have been done. In one of them Tirrell and collaborators created a framework able to perform enrichment analysis in any set of genes using any type of ontology[79]. This same framework has been used later to perform an enrichment analysis using the DO[80]. In that experiment disease ontology terms were mapped to human genes using the NCBO annotator service and afterwards they performed an enrichment analysis using a set of genes that were annotated with the GO term “aging”. By doing this they were able to identify terms of the DO that were associated to aging-related genes, such as “Alzheimer’s disease”, “Insulin Resistance” or “Atherosclerosis” among others.

Another example is the identification of a group of diseases with fewer mutations than expected within O-glycosylation sites[81]. In this example, mutations in proteins were related to disease terms extracted from the Unified Medical Language System (UMLS) and their properties where compared to those of a group of benign mutations from Swissprot. Nevertheless, recent work by Roque *et. al.* has shown that the normalization of clinical data with ontologies can provide meaningful insights into complex aspects of diseases such as their interrelationships and comorbidity[82].

The description and widespread availability of several sets of protein-protein interactions[83] led to the general realization that proteins usually don’t perform their functions alone, but interacting with other proteins. Recently, some groups have developed algorithms that integrate this information with the enrichment analysis[84,85]. In order to do so they represent protein-protein interactions using networks, and take advantage of several of their mathematical properties [86]. These approaches, unlike basic SEA, are not limited to genes with annotations or overlapping gene sets.

1.4. Networks in biology

A network (G) is formally defined in mathematics by the following equation:

G(V,E)

Where V are the nodes, or vertices, of the network and E are the edges connecting the nodes. They are usually used to represent, analyze and interpret complex relational data. They have been proven useful in the study of fields as diverse as social sciences[87], semantics (ontologies are term networks[66]), economics[88], or biology[89]. In these cases a node in a network represents an entity of our interest (a protein, for example), whereas the edges represent any relationship between the nodes (a physical interaction, correlation of expression between two genes etc.).

Nodes and edges have some attributes. For example, edges can have different **weights** (usually in this case more important/relevant edges have higher weights than less important ones) and either be **directed** (if they have a direction, or are one-way) or **undirected** (if they do not have a direction and thus are two-way). All these concepts have also implications when calculating the **shortest path** between nodes (the path between two nodes that has the minimum weight, calculating this weight as the sum of all edges involved in the path).

Properties can be defined for nodes, such as their **degree** (the number of edges connected to a node) their **clustering coefficient** (the ratio between the actual connections between all node's neighbors and those that are possible) or their **betweenness**. This last concept is defined as the fraction of shortest paths of a network where a node appears. Networks have several attributes that can be used to characterize and describe them too (table 3).

Table 3.- General attributes of networks

Attribute	Definition
Size	The number of nodes in a network
Density	The ratio between the number of actual edges and the number of possible edges
Average degree	The average number of edges per node in a network
Average path length	The average number of steps between two nodes of the network
Diameter	The longest of all the shortest paths of a network
Average clustering coefficient	The average clustering coefficient of the nodes in a network

According to these attributes, several models of networks have been defined. These include regular networks, the Random graph model[90] (often referred to as the Erdös-Reyni), the Small world model[91] (or Watts-Strogatz) or the Scale-free model, defined by Barabasi and Albert[92].

There are several differences between the models. For example, in a regular network all nodes have the same degree, whereas in random networks the degree distribution of the nodes follows a normal distribution[90] thus, nodes with very high or very low degrees are scarce or even absent. Networks following the Watts-Strogatz model are built from regular networks with *a posteriori* random rewiring of a subset of their edges. This confers the network some interesting properties, such as a higher clustering coefficient than random networks, but higher connectivity than regular networks[91]. The degree distribution of these networks depends on the percentage of rewired edges, the higher the percentage the more similar to a normal distribution and the smaller the clustering coefficients.

Finally, in the scale-free network the degree distribution of its nodes follows a power law, where most nodes have very few edges and a very few nodes have the most edges[92]. This model is currently used to describe a great variety of things, such as the connections between different web sites, the worldwide distribution of flights and airports or the physical interactions between proteins in a cell[93]. An important concept that appears in scale-free networks is the "hub". These are the nodes with the highest degree and, because of that, are critical in the connectivity of the network since their removal greatly increases the average path length[94].

Figure 5 shows one undirected example of each model of network and some of their properties are summarized in table 4. All the networks have a size of 100, however their densities (which correlates with the number of edges) are different, being the scale-free model the one with the lowest density (and the lowest number of edges), which is reflected in the average node degree of each network. However, despite the lower number of edges and average degree, the average path length and diameter of the scale-free model are not different to those of Erdös-Renyi or Watts-Strogatz. This is due to the presence of hubs in the scale-free model, which centralize the paths of the network, allowing the shorter connections between nodes with few edges.

Several types of networks have been used to describe complex biological data such as food-webs[95] or neuronal networks[96], most attention has turned to networks describing molecular data. For example, protein interaction networks are currently described for several organisms, from *Homo sapiens*[83] to *Drosophila melanogaster* or *Caernohabtidis elegans*[97]. These are usually undirected networks, and all of them have shown a scale-free distribution, suggesting an evolutionary role of this type of organization of the interactome[98].

Not only global network properties of protein-protein interaction networks are preserved through evolution, but also the organization of certain sub-networks or "modules" with critical functions for the cell[98,99]. Networks representing metabolic data have also been created and do have scale-free properties too[89]. In this case nodes usually represent metabolites whereas edges connect metabolites sharing reactions[100].

Transcriptional regulation networks are used to represent information regarding gene regulation[101]. Therefore, unlike protein-protein interaction networks, they are usually directed with edges going from the transcription factor to the target regulated gene[101,102]. Interestingly, this network in *Saccharomyces cerevisae* is scale free when taking into account only the outgoing edges for all nodes (the out-degree), whereas they follow an exponential function model when analyzing the in edges only (in-degree)[103]. This suggests that regulation of the same gene by multiple transcription factors is less likely to occur than regulation of multiple genes by a single transcription factor[104].

Table 4.- Attributes of three networks following different models

Parameter	Regular	Erdös-Renyi	Watts-Strogatz	Scale-free
Size	100	100	100	100
Density	0.061	0.053	0.060	0.040
Av. degree	6	5.2	5.96	3.94
Av. path length	3.99	2.91	2.88	2.97
Diameter	7	6	5	5
Av. clustering	0.2	0.037	0.214	0.142

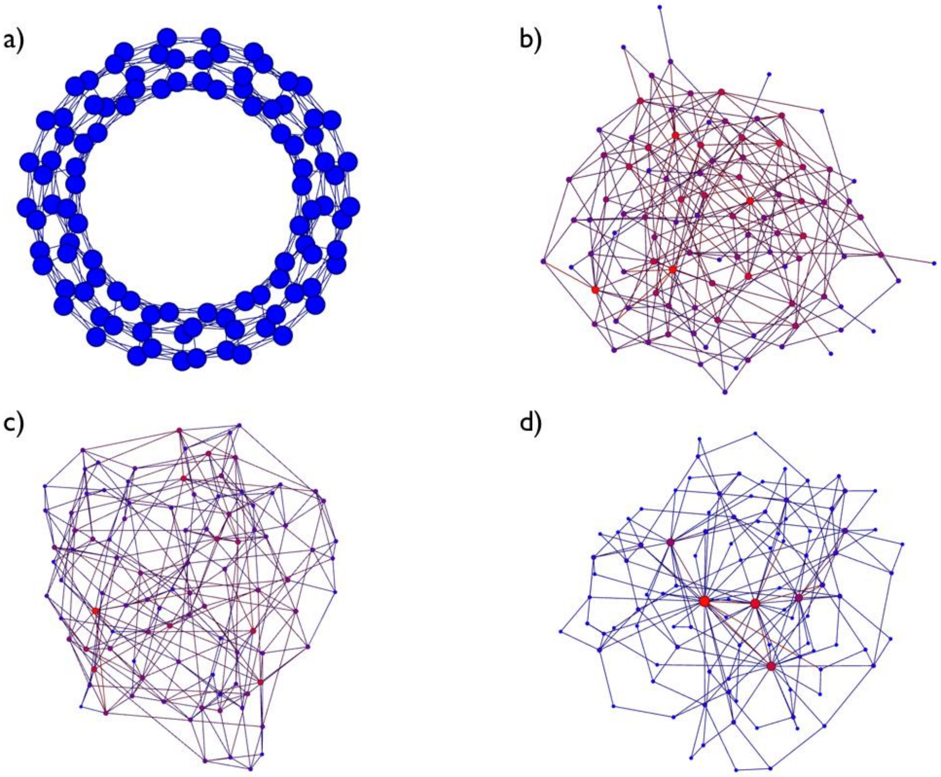


Figure 5.- Graphical representation of three networks of the same size (100) but created following different models. Nodes are colored and sized according to their degree from bigger and red (nodes with higher degree) to smaller and blue (lower degree). From left to right: Regular network, Ërdos-Renyi, Watts-Strogatz and Scale-free. Notice the lower number of edges in the scale-free network and the presence of hubs (bigger nodes colored red) connecting most nodes.

One last interesting example of the application of network science in biology is the analysis of disease networks. These networks connect diseases according to different features such as sharing genes[105] or comorbidity[106,82]. One striking feature of the network by Goh *et. al.* is that most (68%) diseases are connected to other diseases and a significant proportion (40%) form a single component, which means that there is indeed a genetic connection between the majority of disorders[105].

1.5. Diseases and networks

Recent developments in network science and mathematics have also had implications into our understanding of how diseases arise. From the "one gene, one protein, one function" hypothesis[107], with the description of the interactome and the complex interrelationships between proteins, scientists now also take into account the influence of the network organization as factor to describe the interrelationship between genotype and phenotype[108].

One great example on how interactome data can help in understanding genotype/phenotype relationships is the recent work by Zhong *et. al.*[109]. They hypothesized that mutations in the same gene might be associated to different disorders depending on the type of network perturbation that they were causing. From the point of view of interaction networks mutations may cause two types of alterations with different consequences: removal of a node or removal of an edge(s) (figure 6).

Removal of a node could be caused by nonsense mutations that cause non-viable or very small protein products (figure 6). In this case, since the protein is not present in the network, all the edges disappear which implies that none of the interactions of that protein can happen and, thus potentially affecting all the functions of neighboring proteins. On the other hand, missense mutations affecting an interaction surface, or nonsense mutations that leave a partial protein product, only disrupt some of the interactions of that protein. This affects only the function of some of the neighboring proteins, which can lead to different phenotypes.

Zhong *et. al.* tested this hypothesis by mapping 50.000 mutations causing mendelian disorders[109]. They indeed observed different phenotypes for node-removal and edge-removal mutations. Moreover they were also able to identify differences in the phenotypes associated to mutations in the same gene but affecting different domains, thus putatively affecting different edges of the protein.

However, the most important contribution of network science to the biology of diseases is probably the identification of new disease-related genes. Since the observation that genes encoding hubs tend to be essential[93], several groups have tried to answer whether disease genes have specific locations in molecular networks. There is evidence, for example, that cancer proteins tend to have more connections when compared to the rest of the genome[110,111]. Genes associated to mendelian disorders also have more protein-protein interactions[112] and lower clustering coefficients[113] than the rest of the proteins of the genome.

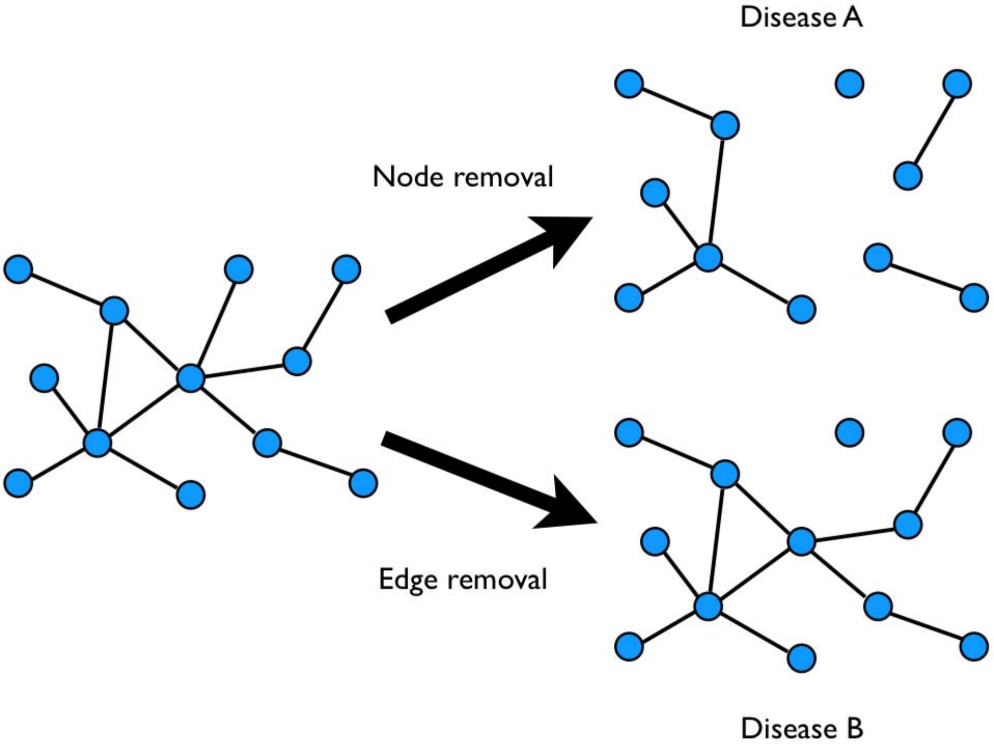


Figure 6.- Diseases as result of network perturbations. Mutations affecting the same gene may be associated to different diseases depending on the network perturbation they cause: the consequences of removing a whole node (e.g. a nonsense mutation at the beginning of the transcript) might be different than the consequences of a missense mutation affecting a single domain.

These results, along with the observation that genes encoding similar functions tend to be located together in the interactome[114] and that genes associated to the same or similar diseases tend to interact more than expected[105], led to the development of several algorithms to predict disease-associated genes based on network features. These features include the closeness to other disease genes[115], the disease centrality of the gene[116], or its local topology signature[117].

Closeness-based methods, also called “guilty-by-association” methods, though they are not controversy-free[118], are becoming very popular and several approaches have been proposed. The simplest method consists in ranking a list of candidate genes according to the number of direct neighbors (DN) in the interactome (distance 1) associated to the disease of interest[119]. This same idea has been successfully applied to protein function prediction[120], and variations exploring neighbors in distances 2 or 3 from the source node have also been developed[121]. In this last case a weighting schema proportional to the distance to the source node was also applied, so that annotations from genes at distance 3 had lower weight than annotations at distance 2 (the same between distance 2 and 1).

Other closeness-based methods rely of the identification of disease-modules. These methods arise from the observation that functionally related genes tend to form locally dense sub-networks[122,123]. Sub-networks can be identified by looking for groups of nodes with higher than expected number of edges between them and fewer than expected edges outside the module. These methods can be computationally challenging, though several algorithms have been developed recently[124,125,126].

There is also another group of algorithms exploiting the closeness to other diseases genes. These are based on diffusion of information through the network and they fully-exploit the local network structure, and thus, are usually more successful than direct-neighbor or module-based approaches[115]. Algorithms in this category include random walk with restart (RWR), diffusion kernels (DK)[127] or propagation flow[128]. In the case of RWR[127] the algorithm calculates the steady-state probability of ending in each node of the network when walking randomly through the network. Candidate disease genes are ranked according to this probability. Higher probabilities indicate higher connectivity to genes known to be associated to the disease, and thus, it is more likely that the candidate gene is also associated to the disorder. The walker starts from any node associated to the disease of interest and goes from one node to another randomly. Another interesting feature of this algorithm is that, in each iteration, there is a probability "p" of going back to the source node.

Methods based on the similarity of the local topology of the gene are based on the idea that genes with similar local topologies can have similar functions[129]. The local topology of a gene can be captured using graphlet-based vectors. A graphlet is an induced non-isomorphic subgraph of a larger network. The difference between a graphlet and a motif is that, while both are subgraphs, a graphlet must contain all the edges between its nodes that appear in the larger network (that's why it is induced), whereas a motif mustn't, since it is only a partial subgraph[130].

Another important concept is that, in order to capture local topologies, one must take into account the difference between the positions within the graphlet. For example, in the 3-node graphlet forming a line, G1 (figure 7), there is a difference between being at the extreme of the line or in the middle. To identify genes associated to melanoma, Milenkovic *et. al*. counted the number of times that each node in the network appeared in each position of every graphlet between 2 and 5 nodes. With this data they created for every node a vector with 72 dimensions (one for every position in each graphlet). Then, candidate genes were ranked according to the similarity of their vector to vectors from genes associated to melanoma[117].

Networks derived from interaction data have been also combined with networks derived from other sources of information to identify disorder-related genes. Yang *et. al.*[131] used a RWR in a mixed network to perform their predictions. Their joint network consisted in a combination of two networks, one derived from interaction data and another one derived from protein-complex data, that were connected by edges connecting proteins belonging to each complex. Similar approaches combining networks from interaction and phenotypic similarity data have also been developed[132,133,134,135].

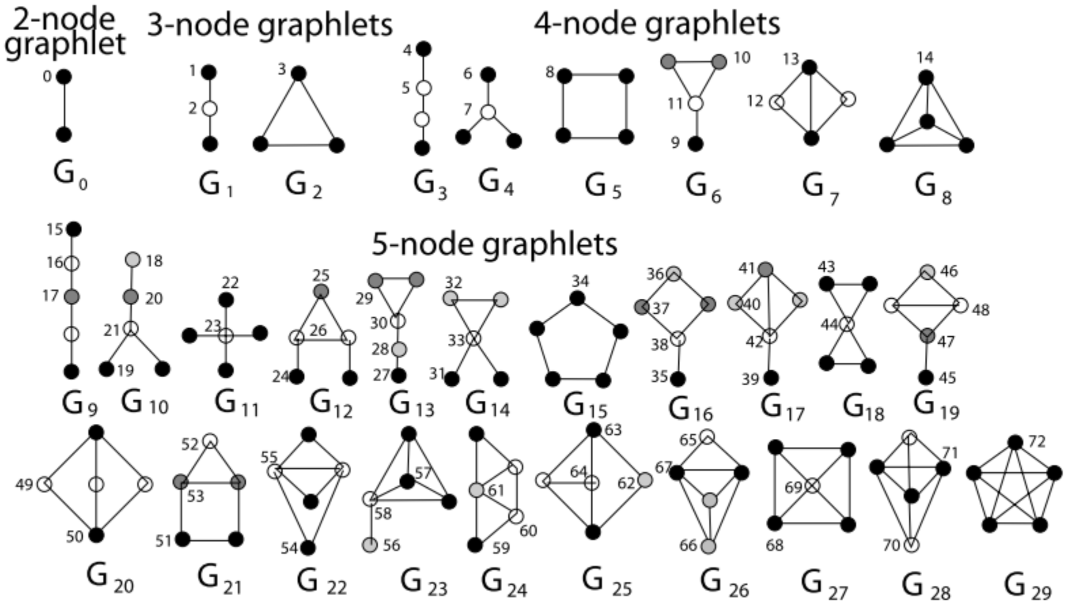


Figure 7.- The thirty different graphlets that can be constructed with 2, 3, 4 or 5 nodes. Different non-symetrical positions within each graphlet are also highlighted and numbered. This figure is from the publication by Milenkovic *et. al.*[115]

Another approach consists in integrating the information from the protein-protein interaction network with that from other sources such as the GO, expression data from microarrays or ESTs or metabolic and pathway data from KEGG among many others. The main differences between these approaches rely on how they integrate the information. Chen *et. al.* calculated the score derived from a DK algorithm of each candidate gene in several networks and kept the highest[136] for each gene after normalizing all the scores. On the other hand, Li and Patra[137] and Lage *et. al.*[138] used ranking statistics to combine the rankings of candidate genes obtained from multiple data sources. Other approaches are the use of support vector machines classifiers[139] or the creation of functional networks using naïve Bayesian classifiers[140,141].

A nice example of the application of functional networks to disease is the work by Lee *et. al.*[142]. To create their network they used 21 different types of information, such as mRNA coexpression, protein-protein interactions or protein complexes for human proteins and their orthologs in mouse, yeast, *C. elegans* or fly. Then, they trained a Bayesian classifier using one gold-standard of true interacting proteins and a set of negative interactions. This classifier yielded 476,399 functional interactions between 16,243 human genes. An important feature of this network is that functional interactions, the edges of the network, are weighted according to their supporting evidence. While they initially used this network to predict phenotypic effects in *C. elegans*[143], their success encouraged them in using it to predict new disease-associated genes from GWAS data for diseases such as Chron's disease or type II diabetes[142].

1.6. Summary

In the era of “omics” data, the use of computational approaches to store, integrate and analyze biological information is becoming a priority, particularly in the field of biomedicine and the study of diseases. Bioinformatics methods have been successfully applied to numerous problems derived from this data explosion, such as the integration of experimentally-derived raw data with other sources of biological information in order to analyze it, the identification of features specific for biologically relevant sets of genes (such as those related to disease) or the prioritization of long lists of genes and mutations potentially associated to different phenotypes.

In this thesis we will develop a new relational database of genes and mutations associated to disorders where annotations will be mapped to ontologies. By doing so, we will overcome some limitations of existing databases, such as their lack of normalization of annotations. This will provide us an optimal framework to investigate the use of ontologies and enrichment analysis to identify disease-specific mutation features that, hopefully, will help us in understanding some aspects of the underlying molecular biology of these diseases. Finally, we will explore whether networks derived from different types are better are predicting different diseases. Moreover, we will also test several combinations of these networks in order to see if they perform better than the networks alone.

2. Materials and methods

2.1. Data handling

Design and development of the relational databases

Two databases were designed during this thesis. Both databases were created, managed and queried using the relational database system MySQL (version 5.1.39). The first one, OCG, was created to store data regarding disease-associated mutations and the corresponding protein features. The second one, CCBG, was created to manage proteins, their associated diseases and different types of relationships between proteins.

OCG

This database is based on the schema from the COSMIC database v44 (November 2009). However, while we reused some of their tables, we have not used those containing information about samples and patient data. We also created some other tables to fit in information on mutation and protein features. The final database contains 21 tables. A general schema of the final database can be found in figure 8. All table's descriptions can be found in the supplementary material 9.1.

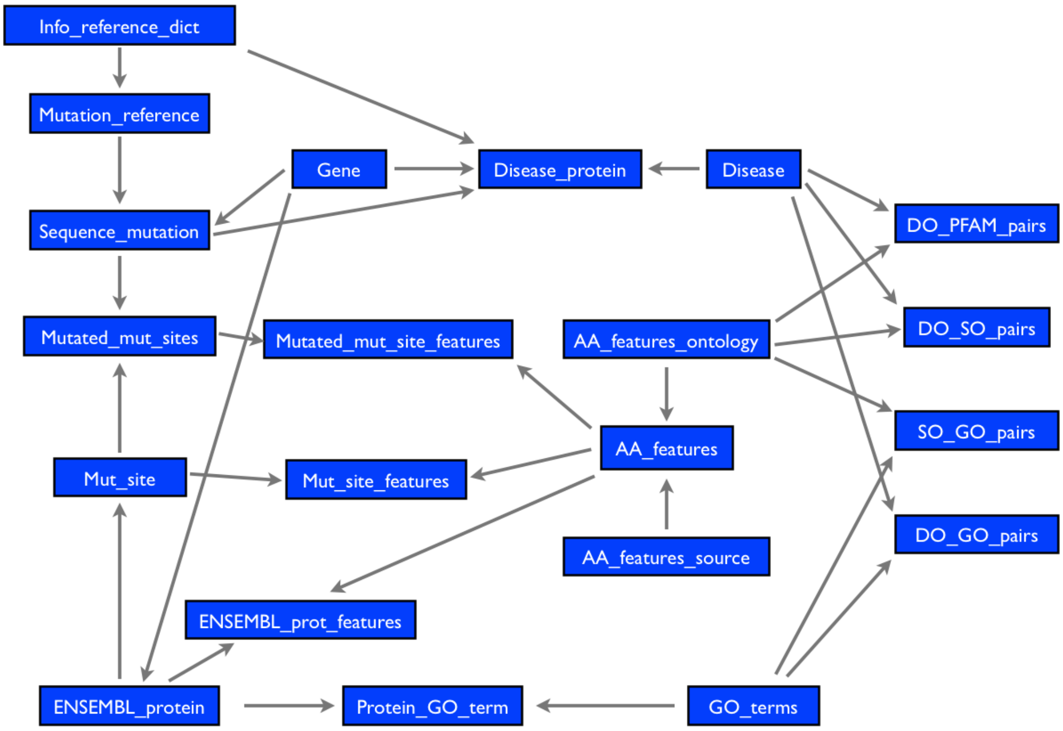


Figure 8.- OCG tables and their relationships. Arrows indicate one-to-many relationships

CCBG

This database contains information on proteins, associated diseases and several types of relationships among the proteins. The final database contains 12 tables and its schema is described in figure 9. The descriptions of all the tables can be found in the supplementary material 9.2.

Since the main purpose of the database is to retrieve different types of protein networks as fast as possible, all the entries in the table storing relationships among proteins, “Genes_related”, are duplicated. In one of the entries the first gene of the relationship is one of the members of the pair and in the other entry is the other gene. By doing this we only have to make one query instead of two to retrieve all the neighbors of a protein in a network. We are well aware that this could introduce some problems as it makes the table more prone to inconsistencies; this is why this table is only modified once, and by a script, when all the information is entered.

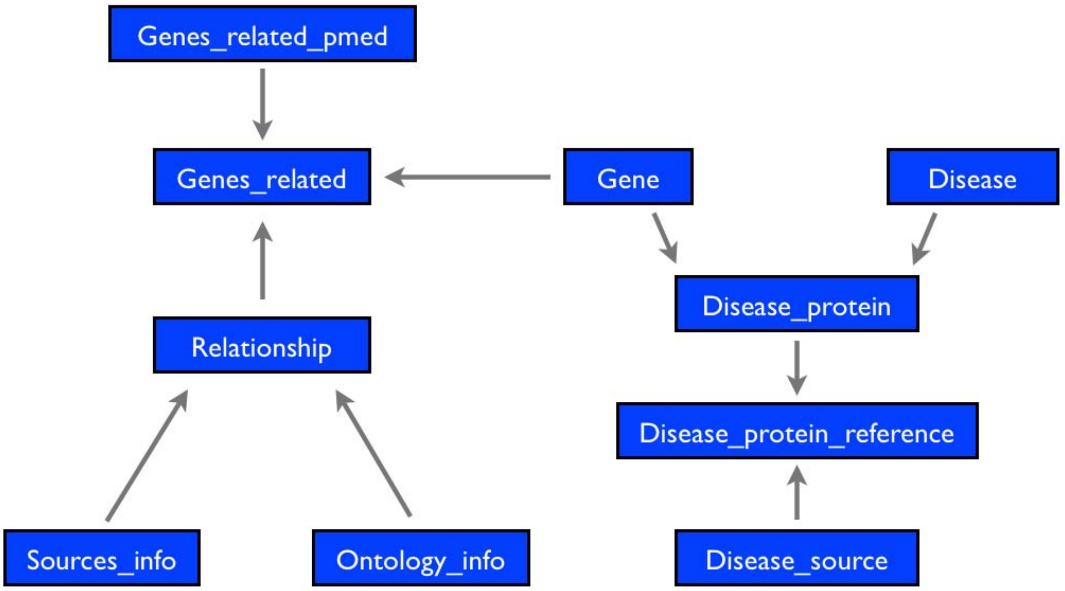


Figure 9.- CCBG tables and their relationships. Arrows indicate one-to-many relationships.

Ontology Lookup Service

All the ontologies used in this thesis were stored and queried in a relational database provided by the Ontology Lookup Service[144] (OLS). This service, provided by a project of the European Bioinformatics Institute, maintains a database with many biomedical related ontologies, included those used in this thesis: the “Gene Ontology”[66] (GO), the “Sequence Ontology”[145] (SO) and the “Disease Ontology” (DO). Weekly releases of whole database dumps are available through EBI’s FTP server. We used the latest version available on 14/04/2011.

The database contains mainly two tables named “Term” and “Term_path”. Table “Term” contains the definition of all the terms of all the ontologies while the

“Term_path” table contains information regarding relationships among terms. A more detailed description of this database can be found in the supplementary material 9.3. In order to fulfill our needs we added some terms and their corresponding relationships to the Sequence Ontology (SO). These are terms regarding different types of post-translational modifications and aminoacid-biased regions (table 5).

Table 5.- SO terms manually added to the ontology. Terms highlighted in red are children of “Post-translationally modified region” (SO:0001089), in dark blue of “sumoylation” (SO:1), in light orange of “palmitoylation” (SO:2), in light brown of “Phosphorylation” (SO:4), in light purple of “glycosylation (SO:3), in yellow of “compositionally biased region” (SO:0001066), in dark brown of “aminoacid enriched region” (SO:13) and in dark orange of “stretch” (SO:14)

Term ID	Translation	Term ID	Translation	Term ID	Translation
SO:1SO1	sumoylation	SO:28SO1	Ser-rich	SO:54SO1	PolyThr
SO:2SO1	palmitoylation	SO:29SO1	Pro-rich	SO:55SO1	Gln/Pro-rich
SO:3SO1	glycosylation	SO:30SO1	PolyHis	SO:56SO1	Leu-rich
SO:4SO1	phosphorylation	SO:31SO1	His-rich	SO:57SO1	Met-rich
SO:5SO1	sumoylation type I	SO:32SO1	Tyr-rich	SO:58SO1	Arg/Asp/Glu/Lys-rich
SO:6SO1	sumoylation type II	SO:33SO1	Pro/Ser-rich	SO:59SO1	PolyPro
SO:7SO1	PKC phosphorylation	SO:34SO1	Ala-rich	SO:60SO1	PolyAla
SO:8SO1	palmitoylation type III	SO:35SO1	PolyPhe	SO:61SO1	PolyIle
SO:9SO1	palmitoylation type I	SO:36SO1	PolyLys	SO:62SO1	Gln/Gly-rich
SO:10SO1	o_glycosylation	SO:37SO1	Arg/Asp-rich	SO:63SO1	Gln/His-rich
SO:11SO1	palmitoylation type II	SO:38SO1	Gln-rich	SO:64SO1	Cys/His-rich
SO:12SO1	n_glycosylation	SO:39SO1	PolyMet	SO:65SO1	Glu/Pro-rich
SO:13SO1	aminoacid enriched_region	SO:40SO1	Gly/Pro/Ser-rich	SO:66SO1	Gln/Glu/Pro-rich
SO:14SO1	stretch	SO:41SO1	Pro/Ser/Thr-rich	SO:67SO1	PolyAsn
SO:15SO1	Ala/Gly/Ser-rich	SO:42SO1	Glu-rich	SO:68SO1	PolySer
SO:16SO1	Arg/Glu-rich	SO:43SO1	PolyGlu	SO:69SO1	Ala/Gly-rich
SO:17SO1	Arg/Lys-rich	SO:44SO1	Glu/Pro/Ser/Thr-rich	SO:70SO1	Ala/Pro-rich
SO:18SO1	Arg-rich	SO:45SO1	PolyGly	SO:71SO1	Gln/Gly/Ser/Tyr-rich
SO:19SO1	Thr-rich	SO:46SO1	Lys-rich	SO:72SO1	Glu/Lys-rich
SO:20SO1	Asp/Glu-rich	SO:47SO1	PolyLeu	SO:73SO1	Ser/Thr-rich
SO:21SO1	Lys/Ser-rich	SO:48SO1	Gly/Leu-rich	SO:74SO1	PolyAsp
SO:22SO1	Arg/Ser-rich	SO:49SO1	Ala/Gly/Pro-rich	SO:75SO1	Asp/Ser-rich
SO:23SO1	Cys-rich	SO:50SO1	Glu/Ser-rich	SO:76SO1	Gly-rich
SO:24SO1	PolyGln	SO:51SO1	Gly/Pro-rich	SO:77SO1	Asp/Glu/Lys-rich
SO:25SO1	Gly/Ser-rich	SO:52SO1	Asp-rich	SO:78SO1	PolyVal
SO:26SO1	Ala/Asp-rich	SO:53SO1	PolyArg	SO:79SO1	Arg/Gly-rich
SO:27SO1	Gly/Thr-rich				

Extraction of information from OMIM and GAD text files to OCG and CCBG

OMIM[6] and GAD[146] text files were downloaded in 21/04/2010 and 16/03/2010 respectively. We designed a parser written in Perl programming language (version 5.10) to identify the genes, mutations and disease terms in both text files. We used the Perl module DBI (version 1.609) as interface to the relational databases and the module Bio::ENSEMBL to use the ENSEMBL[147] Perl API.

Both parsers first establish a connection to the corresponding database (OCG or CCBG) and to ENSEMBL. Then, the parser reads the file until a gene symbol, a disease and, in the case of OCG also a mutation, are found. At that moment, the information and the corresponding relationships between the elements are stored in the proper database.

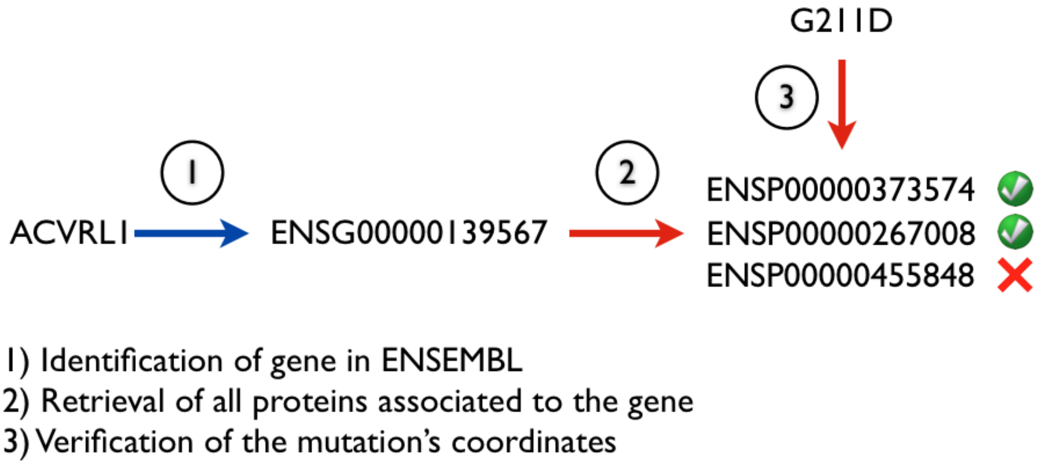

G211D
3
1
2
ACVRL1
ENSG00000139567
ENSP00000373574 ✓
ENSP00000267008 ✓
ENSP00000455848 ✗
1) Identification of gene in ENSEMBL
2) Retrieval of all proteins associated to the gene
3) Verification of the mutation's coordinates

Figure 9.- Identification and verification of genes and mutations found in text files using ENSEMBL API. Arrows in blue indicate steps involved in both, CCBG and OCG. Arrows in red indicate steps involved only in OCG.

In order to do so, the parsers first look for the gene symbol in the text files. When the symbol is identified, a query is made to ENSEMBL to find the proper ENSEMBL identifier for the gene (figure 9, step 1). If a single gene is retrieved the parser moves to next step, however if more than one gene is found (which is usual for example if the gene is located in the HLA region of chromosome 6), the proper gene is manually identified from all the possibilities (usually less than 5).

Then, the parsers search the DO term that best matches the disease words in the text file (figure 10). In order to do so they first look for the disease terms as found in the text in the DO. If there is a single match it keeps the DO identifier. If no matches are found they compare the words found in the text file with all the terms in the DO and keep those DO terms above a certain threshold. The edition score is calculated using the Perl module Text::Levenshtein (figure 10). If more than one match is found at this stage, the parsers iteratively look the parent of the terms that contains the most initial terms (i.e.: if the two matches are “gastrointestinal adenoma” and “gastrointestinal lymphoma” we keep the parent of these two “gastrointestinal system cancer”) and keep its DO identifier. At this point, in the case of CCBG the gene’s symbol, ENSEMBL identifier and DO term identifier of the associated disease are stored and the parser moves to the next entry.

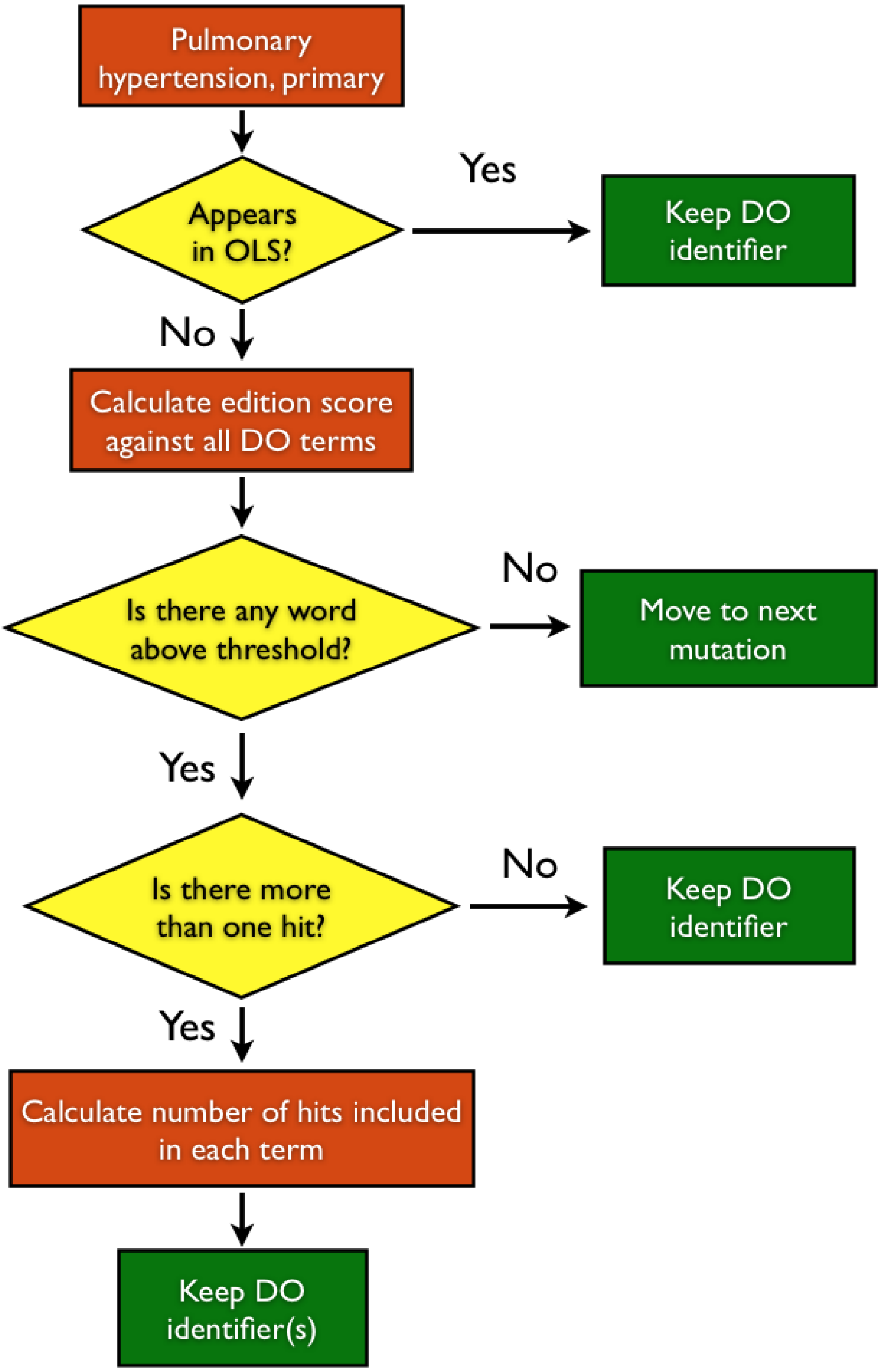

Pulmonary hypertension, primary
Appears in OLS?
Yes
Keep DO identifier
No
Calculate edition score against all DO terms
Is there any word above threshold?
No
Move to next mutation
Yes
Is there more than one hit?
No
Keep DO identifier
Yes
Calculate number of hits included in each term
Keep DO identifier(s)

Figure 10.- Flowchart of the parser to identify the DO term from the description of the disease found in the text files.

Since in CCBG we are only interested in gene-disease relationships this is the final step of the parser. However, in the case of OCG, we want to also store mutation data. In order to do so the parser takes two additional steps (figure 9, steps 2 and 3, marked in red). These consist in fetching all the coding transcripts of the ENSEMBL gene and their corresponding protein sequences. Then the parser checks in which of the protein isoforms the identified mutation is plausible. When a protein isoform matches the mutation's coordinates, the gene symbol, its ENSEMBL identifier, the DO term, the mutation and the isoform's ENSEMBL identifier are stored in the database.

Extraction of information from COSMIC database

An Oracle dump containing the whole COSMIC[148] database version 44 was downloaded from the Sanger Institute FTP server on 25/11/2009. This dump was converted to MySQL format using DBConvert. COSMIC stores the histological description and the anatomical location of the samples where the mutations are found. The histology does not add a lot of information in version 44, since most of the mutations (>90%) are described as "carcinoma". Thus we inferred the type of cancer for each sample from its anatomical location. Since the number of anatomical locations is relatively small (220), when possible, we manually mapped all the anatomical location to the best matching DO term.

We tried to identify all the COSMIC genes in ENSEMBL. For each gene identified, its associated DO terms where extracted, as explained above, from the anatomical information of the mutated tumor samples. The gene symbol, its ENSEMBL identifier and the associated DO terms were stored in the databases. Again, in the case of CCBG this is the final step, since we were only interested in gene-disease associations.

In the case of OCG, the program additionally fetches all the coding transcripts for each gene correctly identified in ENSEMBL and tries to map all the non-synonymous mutations described in COSMIC to them. When a mutation was successfully mapped to an isoform, the gene symbol, its own ENSEMBL identifier and that of the protein isoform, the mutation description and the DO term inferred from the anatomical location of the sample were stored in OCG.

Mutated protein features

For all the proteins in OCG we extracted different sequence features from both, publicly available resources and dedicated software. All the features were manually mapped to their corresponding PFAM or SO term and associated to the mutation site in OCG.

We have added novel SO terms because we believe that important differences with biological impact are not reflected in the SO (table 5, mentioned in section 2.1). For instance, an "aminoacid-biased region" or "cryptic repeat" is a region of a protein that shows enrichment in a given amino acid. While it is a compositionally biased region (and, thus, it has been added as a children of this term in the SO), we have made a distinction between this type of bias and “homopolymeric stretches”. These are regions that are made of linear repetition of a single amino acid (this term has also been added as a son of “compositionally biased regions” as well).

The rationale to make this distinction is because polymorphisms in homopolymeric stretches have been associated to several types of diseases such as Huntington’s disease[7] or oculopharyngeal muscular dystrophy[149]. The difference between these two, according to Uniprot, is that “aminoacid tandem repeats” are homopolymers of a single aminoacid with less than one interruption per five aminoacids, whereas “cryptic repeats” are larger regions that show a lower level of bias, but in which the aminoacid composition also differs significantly from that expected at random. We added also subtypes of “aminoacid-biased region” and “aminoacid tandem repeats”. For example, “Serine-rich regions” are “aminoacid-biased regions” where the amino acid bias is caused by an excess of serines.

ENSEMBL

Using the ENSEMBL Perl API we extracted information regarding compositionally biased region and PFAM domains for all the mutation sites. Mutation sites falling within compositionally biased regions where associated to its corresponding SO term identifier (SO:1066SO1), whereas those located inside a PFAM domain were associated to the domain’s PFAM identifier. Finally, we also extracted all the GO terms associated to the mutation site protein.

Uniprot

We queried Uniprot through its Das server to fetch several types of information on the mutation sites using the Perl module Bio::Das::Lite. However, since mapping ENSEMBL isoforms to Uniprot proteins is not direct we used the ENSEMBL API to find, for each ENSEMBL protein, the corresponding Uniprot ID. Once a Uniprot protein id was found, we tried to remap a window of 10 aminoacids centered in the mutation site from the ENSEMBL protein to that Uniprot protein. If the window could be mapped, the Uniprot coordinates for the mutation-site were calculated and the features extracted and mapped to the mutation, otherwise the feature was only extracted and stored in the table “ENSEMBL_prot_features”. We extracted information from Uniprot regarding several types of compositional bias, post-translational modification sites, transmembrane regions, disulfide bonds, signal peptides and secondary structure.

Predictors

We used a series of predictors to extract different types of features around the mutation sites. These predictors were either run locally or queried through Internet using different Perl modules. Table 6 summarizes this information.

It's important to notice that only wild-type sequences were used to make the prediction, thus we have no information on the mutated sequences.

Table 6 .- Software used to predict mutation features

Predictor	Feature	Run
PsiPred[150]	Secondary structure	Local
CSS Palm[151]	Palmitoylation sites	Local
Sumo SP[152]	Sumoylation sites	Local
Sherloc[153]	Sub-cellular localization	Local
NetPhos[154]	Phosphorylation sites	Das Server
NetOGlyc[155]	O-glycosylation sites	Das Server
NetNGlyc	N-Glycosylation sites	Das Server
FoldIndex[156]	Unstructured regions	Server API

2.2. Identification of disease-specific features

In order to identify features statistically associated to specific diseases we performed a two-tailed Fisher's test comparing the distribution of the mutations between pairs of ontology terms. These pairs are formed by a term from the DO and another term from either the GO, the SO or PFAM.

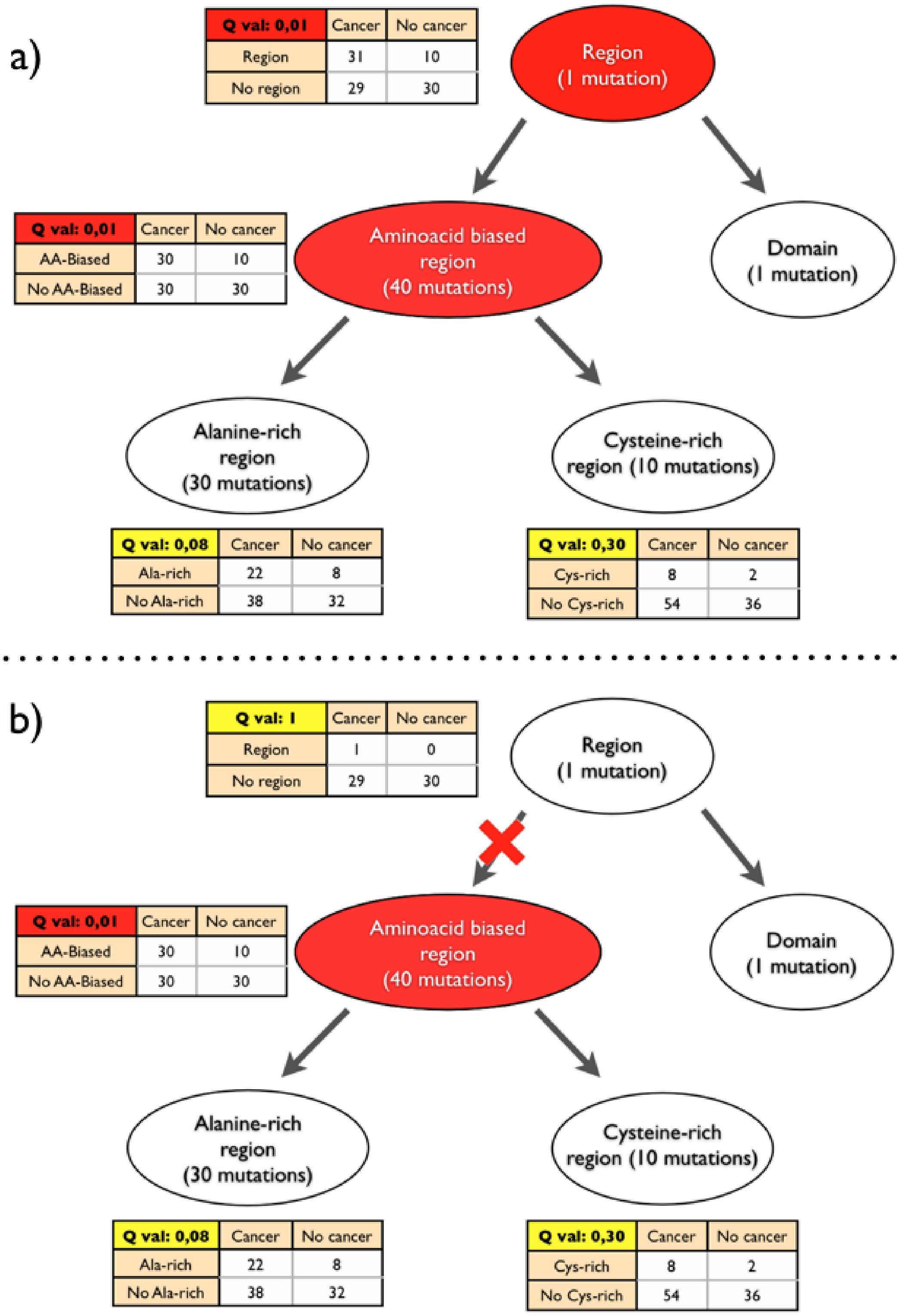
Enrichment analysis

Two terms of two different ontologies are related if they share at least one mutation. We used the ontology structure to aggregate the information from children to parents so that we would be able to detect non-obvious associations. We have corrected multiple testing results using Bonferroni to avoid false positives. We have also identified three other possible sources of non-informative associations in our analysis.

First, in any ontology there is a *strong interdependence of the different terms*. This is an intrinsic property of any ontology and therefore difficult to overcome. If two terms in our analysis show a statistically significant association it is possible that their respective parents are also associated. However, this particular association between the parents may be an artifact caused by extending the information from the children. In order to solve this issue we have implemented the "elim" algorithm (figure 11) described by Alexa and collaborators[157], where information from a child to its parent is not aggregated if the child is already associated to a given feature. This way we keep only the most informative associations (the ones further from the root of the ontology and, as a consequence, that involve the most specific terms).

Secondly, there is indeed a *bias in the number of mutations depending on the gene*. Popular genes, which are usually associated to extensive studied diseases, have been scanned for mutations more often than those that have not been associated to those diseases. This could introduce a bias in the features associated to a disease. For example, the TP53 gene is one of the most studied cancer genes. It has 189 different missense mutations according to our data. If all its mutations would share one particular feature, that feature could be considered as "associated to cancer" even if it is only present in that single gene. In order to solve this problem we only kept associations involving more than 3 genes.

Finally, the *absence of a given feature in certain diseases* though statistically meaningful, might not be biologically relevant. In those cases where a feature never appears in a set of disease-associated mutations, this particular feature is most likely irrelevant for that certain disease and therefore the association is uninformative. Otherwise, if the feature is present in some mutations, but less than expected, it is more likely that the association is biologically relevant. This issue has been corrected by introducing the 3 genes association threshold aforementioned.



Q val: 0,01	Cancer	No cancer
Region	31	10
No region	29	30

Q val: 0,01	Cancer	No cancer
AA-Biased	30	10
No AA-Biased	30	30

Q val: 0,08	Cancer	No cancer
Ala-rich	22	8
No Ala-rich	38	32

Q val: 0,30	Cancer	No cancer
Cys-rich	8	2
No Cys-rich	54	36

Q val: 1	Cancer	No cancer
Region	1	0
No region	29	30

Q val: 0,01	Cancer	No cancer
AA-Biased	30	10
No AA-Biased	30	30

Q val: 0,08	Cancer	No cancer
Ala-rich	22	8
No Ala-rich	38	32

Q val: 0,30	Cancer	No cancer
Cys-rich	8	2
No Cys-rich	54	36

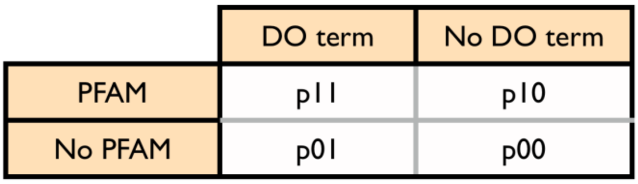
Figure 11.- Use of the ontologies to aggregate the information. Arrows go from parents to children. Terms filled in white are not associated to the disease of interest, whereas those filled in red are associated. When propagating the information from children to parents (a) when a term is associated to a disease, its parents are usually also associated because the same mutations are being analyzed. In (b), thanks to the "elim" algorithm, once associated, mutations do not propagate to the parents and, thus, less informative associations are avoided.

Identification of mutation patterns across different PFAM domains and DO terms

We next interrogated whether similar DO terms would be enriched or depleted of mutations in similar PFAM domains. In order to answer this question we calculated the odds ratio (OR) of all the DO/PFAM pairs using our mutation data. The odds ratio is calculated using the following equation:

OR = p11*p00/(p01*p11)

Where the p00, p01, p10 and p11 represent the values of each cell of the following contingency table:



	DO term	No DO term
PFAM	p11	p10
No PFAM	p01	p00

In order to analyze the data we created a heatmap using the R package “gplots”. The rows and columns are clustered using a hierarchical clustering algorithm and the cells are colored according to the logarithm of the OR from red (negative values) to green (positive values).

Since there are over 1000 DO terms in our database and considering that most of them are related to each other, the interpretation of a heatmap including all of them would be difficult to interpret. Moreover, some DO terms are associated to a handful of mutations, so we would only have data for few PFAM domains.

Therefore, we designed an approach to select an optimal group of DO terms. We established an adaptive mutation threshold that we ranged between 50 and 300 in intervals of 5 mutations. Then, for each threshold we calculated which DO terms have a number of mutations above the threshold and are the most distant from the root (all those that have no children with mutations above the threshold). Once this set of “root” DO terms was generated, we calculated its mutation coverage, since the DO terms should represent at least the 90% of the mutations in the database).

We finally obtained an optimal threshold of 295 mutations and generated the heatmap with the DO terms derived from this threshold.

Association with the mutations or association with the genes

We next addressed if the associations found between SO and DO terms were caused by a bias in the properties of the genes or by a true bias in the properties of the mutations. Then, we calculated for each feature associated to cancer mutations whether such association was maintained when only considering genes having the feature. By doing this we removed the possible bias caused by the gene properties: all genes have the property, thus, any difference in the mutation rate will be truly disease dependent and not a sampling problem.

To do so we extracted all the genes having a given feature, regardless of its mutation state, and divided them into two groups: cancer-associated and no cancer-associated. Then, we recalculated to statistical significance of all the associations using these unbiased groups.

Confirmation of the associations in an independent dataset

In order to confirm the relevant associations of SO terms with cancer-associated mutations found in the enrichment analysis we downloaded a set of missense mutations from the Cancer Genome Atlas (TCGA)[158]. This other dataset of cancer mutations includes 18834 missense mutations from 814 cancer samples of acute myeloid leukemia, colorectal cancer, glioblastoma, ovarian cancer and rectum adenocarcinoma.

Genes and mutations were validated, mapped to ENSEMBL identifiers and stored in the database as explained in section 2.1 with the only difference that the phenotypes were manually mapped to DO terms.

Information regarding the 6 relevant SO terms found in the enrichment analysis was retrieved as described in section 2.1. We then used this information to perform a two-tailed Fischer test comparing the abundance of each feature in this set of cancer mutations against the abundance in the set of mutations associated to other diseases.

Intra-ontology associations

Since a single mutation can map to more than one SO term, some SO terms might be giving us the same information than others and, thus, be redundant. We established the independence of information between the different SO terms by looking for higher than expected co-occurrences of features in mutations. To do so we performed a two-tailed Fisher's test comparing every pair of SO terms that mapped to the same mutations. We corrected the results for multiple testing using the Bonferroni correction.

2.3. Network-based prioritization of disease genes

We extracted information from different public resources on four different types of relationships among human genes: physical interactions, coexpression in healthy tissues, belongingness to the same biological pathway and paralogy.

For each type of relationship we defined a network **G(V,E)** where V is the set of nodes and E is the set of edges. The set of nodes V of each network comprises all the genes for which at least one relationship of the given type is defined. The edges, E, of each network, G, are the relationships extracted from public resources. It is important to notice that:

- A gene V_i can be present in a network G_i, but absent in another G_j, if a relationship is defined for network G_i but not for G_j
- All the genes V_i that belong to network G_i, have at least one edge
- All the edges in all the networks are undirected

Interaction network

We used the file "BINARY_PROTEIN_PROTEIN_INTERACTIONS.txt" (downloaded from the HPRD[159] web site on 25/10/2011) to create this network. This is a tab-delimited file that defines in each row a physical interaction between two proteins and provides information on the experiments were the interaction was found and the Pubmed identifier(s) of the publication(s) of the experiments. All the interactions of these networks are identified with the term "CCBGID:1".

A Perl script was created in order to identify both members of the interaction in ENSEMBL using its API. An interaction was stored only when both genes were identified and their ENSEMBL identifiers were kept. All the papers describing the associations were also kept. Self-interactions or multimerizations were not stored. The original file contains 36849 interactions between 9451 proteins Of these, we could successfully extract 7605 interactions (21%) between 3852 proteins (41%).

Coexpression network

Data for this network comes from files from the BioGPS[160] website downloaded on 23/09/2011. The files contain gene expression data from 79 human tissues measured using two different chips: HG-0133A and GNF1 from Affymetrix. This data has been normalized using the gcRMA algorithm[161].

We did not use expression values from whole tissue and disease samples as they might add noise to the data, and we only kept expression values from those probes that could be mapped to ENSEMBL genes through its API. For those genes that mapped to more than one probe, their final expression value was calculated as the average of all probes. Moreover, we used only those probes that had a gcRMA value above 8 in at least one tissue, since values below that threshold are considered to be background fluorescence.

We then calculated the Spearman's rank correlation coefficient between all gene pairs. An edge between two genes is defined in this network when both genes show an R correlation value above 0.71 (which corresponds to an R^2 value of 0.5), regardless of the slope of the correlation (thus we have edges connecting correlated -positive slopes- and anti-correlated -negative slopes- genes). All the interactions are defined in the database with the word "CCBGID:2". The original file contains expression data for 14455 genes, which means that there are 208.947.025 potential interactions in this network. However, after applying all the aforementioned filters and calculations, the final network contains 8056 nodes and 790113 edges.

Pathways network

Edges in this network are defined when two genes are sequentially connected in a pathway. All the edges are identified in the database by the word "CCBGID:3". Data for this network comes from Reactome[162]. The file "homo_sapiens_interactions.txt" was downloaded from the Reactome[162] site on 16/09/2011. This tab-delimited file contains in each line different identifiers of two genes and their type of relationship. Reactome defines 4 types of relationships: "reaction", "neighbouring_reaction", "complex" and "indirect_complex". In this case we used those pairs labeled as "reaction" or "sequential_reaction" and, again, we kept only those were we could retrieve ENSEMBL identifiers for both genes. The original file contains 76438 pairs between 3979 genes. Of these, we successfully parsed 72817 relationships (95%) between 3790 genes (95%).

Paralogs network

To generate this network we used the homology-derived data from ENSEMBL, extracted via its API on 22/02/2012 using a Perl script. An edge is defined in this network when two human genes are defined as paralogs in ENSEMBL. Only data for protein coding genes was stored. This network contains 14580 genes connected by 91022 edges. All the interactions that belong to this network are labeled in the database with the identifier "CCBGID:4".

2.4. Performance of the different networks and algorithms

We chose 5 different diseases to evaluate the performance of each network and each algorithm: “cancer”, “colorectal cancer”, “simple genetic diseases”, “diabetes” and “neurodegenerative diseases”. We chose these 5 diseases because there are at least 20 genes associated to each one of them in each network and their underlying biology is different from each other.

We assessed the predictive power for each network and algorithm in each disease by calculating the area under the ROC curve (AUC) obtained in a 5-fold cross validation experiment. Briefly, we divided each group of disease-associated genes in 5 different sets and used 4 of these sets to try to predict the last one. This was repeated 5 times, one to predict each group. The AUCs were calculated using the ROCR[163] R package.

2.5. Network algorithms

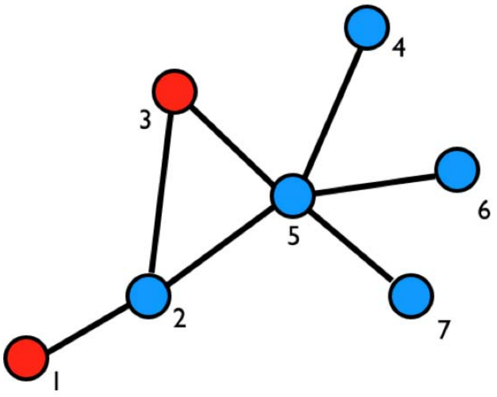
We evaluated five different algorithms to predict in a set if candidate disease genes that might be associated to a disease: “direct neighbor” (up to distances 1, 2 and 3), “diffusion kernels” and “random walks with restart”.

Direct neighbor (DN)

This algorithm ranks candidate genes according to the number of neighbors up to distance D that are associated with the phenotype of interest. In order to do so, the algorithm sums, for each candidate gene, the number of nodes associated to the disease of interest up to a predefined maximum distance (Dmax):

$$Score(A) = \sum_{i \in I} \sum_{D=1}^{D\max} \frac{i}{D}$$

Where A is the candidate gene, I is the set of disease-associated genes, D is the distance to gene A and Dmax is the maximum distance. Notice also that the weight given to the disease-associated neighbors is inversely proportional to the distance to the candidate gene.



Dmax = 1	Gene 2	Gene 5	Gene 4
Dist 1	2	1	0
Score	2	1	0

Dmax = 2	Gene 2	Gene 5	Gene 4
Dist 1	2	1	0
Dist 2	0	0.5	0.5
Score	2	1.5	0.5

Dmax = 3	Gene 2	Gene 5	Gene 4
Dist 1	2	1	0
Dist 2	0	0.5	0.5
Dist 3	0	0	0.33
Score	2	1.5	0.83

Figure 12.- Illustration of the weighted DN algorithm. Nodes in red represent disease-associated genes. The table on the right summarizes the scores obtained for 3 candidate genes (nodes 2,5 and 4) when going to distance 3. Notice that, while gene 2 is connected to gene 3 also at distance 2 (through the path 2-5-3), it is not taken into account when analyzing that distance because this relationship has been analyzed at distance 1.

We studied the performance of the algorithm using maximum distances of 1, 2 and 3. A neighbor of the candidate gene is only taken into account once in the analysis in the shortest distance where it appears. This is exemplified in figure 12. In this figure we can see that gene 3 (associated to the disease) is found at distances 1 (edge 2-3) and 2 (edges 2-5-3) from candidate gene 2. Only the distance 1 is taken into account for our analysis, regardless of the maximum distance explored.

It is important to also notice that the weight of the disease genes is proportional to the distance where they are found. For instance, in figure 12, gene 1 adds a score of 0.5 to gene 5, since it is located at distance 2 from that gene. Something similar happens with gene 4 and gene 1, which only ads 0.33 since it's located at distance 3 from that gene.

Random Walk with Restart (RWR)

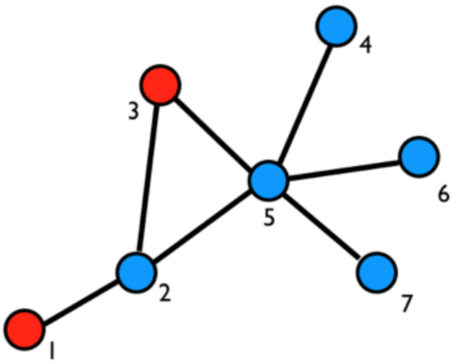
The random walk on graphs is defined as an iterative walker's transition from its current node to a randomly selected neighbor starting at a given source node. We used a variant of this approach as defined by Kohler *et. al.*[127] that allows the restart of the walk at each time with probability r. The random walk is described by the equation:

$$p_{t+1} = (1 - r) \cdot W \cdot p_t + r \cdot p_0$$

Where W is a column-normalized adjacency matrix of the graph, p_t is a vector in which the i-th element holds the probability of being at node i at time t and p_0 is the initial probability vector. For each set of disease-associated genes, D, elements of p_0 are defined as:

1/|D| for each gene d∈D
0 otherwise

Figure 13 shows a small network and its associated column-normalized adjacency matrix.



$$\begin{pmatrix} 0 & 0.33 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0.5 & 0 & 0.2 & 0 & 0 \\ 0 & 0.33 & 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0.33 & 0.5 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.2 & 0 & 0 \end{pmatrix}$$

Figure 13.- The example network and the colum-normalized adjacency matrix. Notice that this matrix is not symmetric.

The first iteration (p_1) in such network with a restart probability “r” of 0.25 would be calculated by:

$$p_1 = (1-0.25)\cdot\begin{pmatrix}0 & 0.33 & 0 & 0 & 0 & 0 & 0\\ 1 & 0 & 0.5 & 0 & 0.2 & 0 & 0\\ 0 & 0.33 & 0 & 0 & 0.2 & 0 & 0\\ 0 & 0 & 0 & 0 & 0.2 & 0 & 0\\ 0 & 0.33 & 0.5 & 1 & 0 & 1 & 1\\ 0 & 0 & 0 & 0 & 0.2 & 0 & 0\\ 0 & 0 & 0 & 0 & 0.2 & 0 & 0\end{pmatrix}\cdot\begin{pmatrix}0.5\\ 0\\ 0.5\\ 0\\ 0\\ 0\\ 0\end{pmatrix} + 0.25\cdot\begin{pmatrix}0.5\\ 0\\ 0.5\\ 0\\ 0\\ 0\\ 0\end{pmatrix} = \begin{pmatrix}0.13\\ 0.56\\ 0.13\\ 0\\ 0.18\\ 0\\ 0\end{pmatrix}$$

The walker in the first iteration, starting from both disease genes, has moved to the neighboring nodes 2 and 5. However, notice that there is a non-null probability that the walker goes back to the disease genes.

In the second iteration (below) the walker has a non-null probability of being in every node. The two disease-related genes, nodes 1 and 3, have now increased their score because the walker can go back to them either through the normal iteration (from nodes 2 and 5) or through the restart mechanism:

$$p_2 = (1-0.25)\cdot\begin{pmatrix}0 & 0.33 & 0 & 0 & 0 & 0 & 0\\ 1 & 0 & 0.5 & 0 & 0.2 & 0 & 0\\ 0 & 0.33 & 0 & 0 & 0.2 & 0 & 0\\ 0 & 0 & 0 & 0 & 0.2 & 0 & 0\\ 0 & 0.33 & 0.5 & 1 & 0 & 1 & 1\\ 0 & 0 & 0 & 0 & 0.2 & 0 & 0\\ 0 & 0 & 0 & 0 & 0.2 & 0 & 0\end{pmatrix}\cdot\begin{pmatrix}0.13\\ 0.56\\ 0.13\\ 0\\ 0.18\\ 0\\ 0\end{pmatrix} + 0.25\cdot\begin{pmatrix}0.5\\ 0\\ 0.5\\ 0\\ 0\\ 0\\ 0\end{pmatrix} = \begin{pmatrix}0.26\\ 0.17\\ 0.29\\ 0.03\\ 0.19\\ 0.03\\ 0.03\end{pmatrix}$$

Ideally, the score for each candidate gene **A** should be a-th element of p_∞. However, since it is not feasible to calculate the steady-state probability vector p_∞, we approximated its value by iterating p_t until the L^1 norm between p_t and p_{t+1} fell below 10^{-6}, which is usually before 50 iterations. In order to explore longer distances of the networks we used a low restart probability of 0.1.

Influence of the different parameters in the performance of the RWR

After evaluating each of the 5 methods and comparing their performances, we decided to use only the RWR for the next part of the project because:

a) It usually performs better than the DN algorithm
b) It is usually much faster than the DN and DK algorithms, particularly for large networks

Given that the algorithm depends mainly on 3 factors, the restart probability, the initial probability vector and the matrix, we decided to evaluate the influence of each of these 3 factors in the performance of the algorithm.

In order to assess the influence of the restart probability we calculated the AUC of all the disease/network combinations varying the restart probability between 0.1 and 0.9 in 0.1 intervals.

In the case of the initial probability vectors, for each network and each disease we created 10 random vectors with the same number of starting nodes and measured their capability to predict 1/5 of the disease genes using each network.

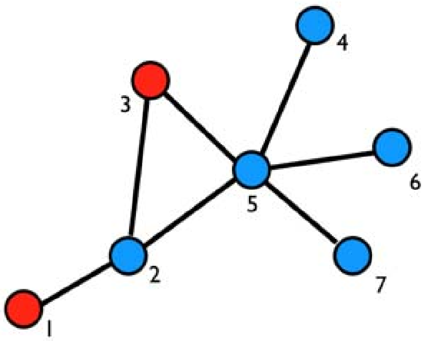
Finally, we designed a similar experiment for the matrixes: we generated 10 random matrixes and, using the initial probability vectors, we tried to predict 1/5 of our disease genes.

Diffusion Kernel (DK)

The idea of the diffusion kernels of graphs consists in simulating a random walk without restart (explained in the following section) with and infinite number of infinitesimally small steps, which would be similar to p_∞ in a random walk without restart. It is used to simulate the diffusion of an element introduced in a given node through the graph. In our case we are using this concept to evaluate how the information introduced in a candidate gene diffuses through the network, as a proxy to how close two genes are. Mathematically, the diffusion kernel K of a graph G is defined as:

$$K = e^{-\beta L}$$

Where β is a parameter that controls the magnitude of the diffusion and L is a matrix defined as L = D-A. D is the diagonal matrix containing the nodes' degrees and A is the adjacency matrix of G. Figure 14 shows a network and its corresponding A, D, L and K matrixes.



$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

$$D = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Figure 14.- Example network and its associated adjacency and node-diagonal matrixes.

Thus, in this the L matrix (D-A) would be:

$$L = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 2 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & -1 & -1 & -1 & 5 & -1 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}$$

And the diffusion kernel (K) with β = 0.1:

$$K = \begin{pmatrix} 0.9091 & 0.0824 & 0.0042 & 0.0001 & 0.0039 & 0.0001 & 0.0001 \\ 0.0824 & 0.7524 & 0.0820 & 0.0039 & 0.0716 & 0.0039 & 0.0039 \\ 0.0042 & 0.0820 & 0.8267 & 0.0040 & 0.0751 & 0.0040 & 0.0040 \\ 0.0001 & 0.0039 & 0.0040 & 0.9088 & 0.0752 & 0.0040 & 0.0040 \\ 0.0039 & 0.0716 & 0.0752 & 0.0752 & 0.6239 & 0.0752 & 0.0752 \\ 0.0001 & 0.0039 & 0.0040 & 0.0040 & 0.0752 & 0.9088 & 0.0040 \\ 0.0001 & 0.0039 & 0.0040 & 0.0040 & 0.072 & 0.0040 & 0.9088 \end{pmatrix}$$

In this case, given a set of genes, D, associated to a disease, the score of a candidate gene A is defined as:

$$\boldsymbol{Score\ (A) = \sum_{i \epsilon I} K_{\,i}(A)}$$

Where $K_i(A)$ is the value in row “i” (the row with the values of a disease-associated gene), column “A” (the column of the candidate gene) of matrix K. In our example, the final score for genes 2, 5 and 4 would be:

Score (2) = $K_{1,2} + K_{3,2}$ = 0.0824 + 0.0042 = **0.0866**
Score (5) = $K_{1,5} + K_{3,5}$ = 0.0039 + 0.0751 = **0.0790**
Score (4) = $K_{1,4} + K_{3,4}$ = 0.0001 + 0.0042 = **0.0043**

Gene 2 would be ranked as the best candidate because its DK value is the highest.

2.6. Combination of the networks

There have been some attempts to combine different types of information in a single network[164] or different networks between them[131]. We then explored whether the combination of networks outperforms the networks alone when trying to predict disease-associated genes. We also tested whether this could depend on how the networks are combined. We evaluated 4 different ways to combine the networks:

Bayesian inference

This method consists in learning the *a priori* probabilities of a gene to be associated to a disease or not given some parameter and labeled positive and negative examples. In our case we used 3/5 of the disease-associated genes in each network to start a RWR and calculate the closeness of the rest of the genes in the network to the disease genes.

As labeled positive genes we used 1/5 of the disease genes, and as negative genes we used the rest of the genes in the network that are not associated to the disease of interest. The next step is to calculate the mean and standard deviation of the scores of these two groups. With this information we estimated the probabilities of the remaining 1/5 of disease genes to be associated to the disease or not using the following equation:

$$P(Belonging\ to\ group) = \frac{1}{\sqrt{2 \cdot \pi \cdot sd^2}} \cdot e^{-\frac{(score-mean)^2}{(2 \cdot sd^2)}}$$

Where “sd” and “mean” are the standard deviation and the mean of the scores of each group (disease-associated and non disease-associated) and “Score” is the score of the candidate gene. This equation estimates the probability of a certain observation to belong to a given group given the distribution of other observations that belong to this group (defined by its mean and standard deviation).

For each candidate gene we repeated this operation in each network where it appears. We then calculated the integrated probabilities of all candidate genes to belong to each group. In order to do so, we multiplied the probabilities obtained in each network and the global probability to be a disease gene or not. The final score of the candidate genes is the ratio between the integrated probabilities of the gene to be associated to the disease with that of not being associated to it.

Juxtaposition

This method simply consists in adding the nodes and edges of a network $G_i(V_i, E_i)$ to another network $G_j(V_j, E_j)$:

$$\boldsymbol{G_{i \cup j} = (V_i \cup V_j, E_i \cup E_j)}$$

Notice that in this case all the edges have the same weight, regardless of the network they come from or the number of networks they appear in.

Simple addition

This approach is similar to the previous one, however in this case the edges are weighted according to the number of networks they appear in. If we are combining two networks, all the edges that only appear in one of them will have a weight of 1 and all the edges that appear in both will have a weight of 2.

$$\boldsymbol{G_{i+j} = (V_i + V_j, E_i + E_j)}$$

Weighted addition

In this case we first weight the edges of each network in an arbitrary scale between 0 and 20. Then, we make the addition of the networks as in the case of "Simple addition" and evaluate the combination of weights by calculating the AUC. Thus, the network is defined as:

$$\boldsymbol{G(V_W, E_W) = \sum_{i=1}^{4} (w_i \cdot V_i, w_i \cdot E_i)}$$

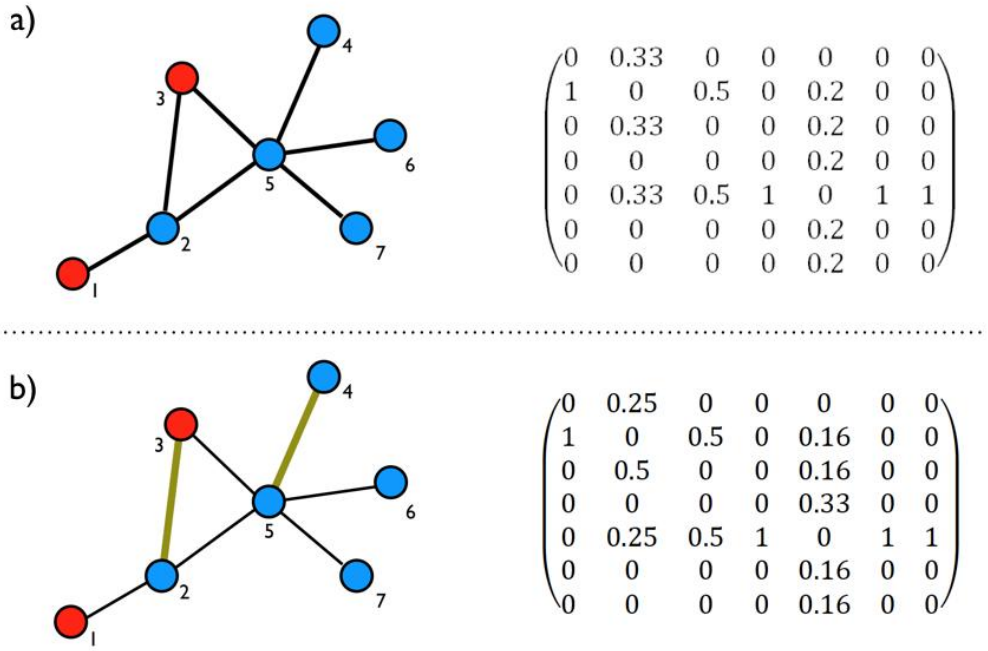
Where "i" is the network identifier, w_i is the optimized weight for network "i", and V_i and E_i are the nodes and edges of that network.

In order to find the optimal combination of weights we thought about evaluating all the possible network combinations using a predefined set of weights in the aforementioned range (e.g. 1, 5, 10, 15, 20). However, given that the solution space is very large and that the optimal solution for one disease is unlikely to be the same for another disease, we decided to optimize the weights for each network by using a simulated annealing over 200 iterations.

The simulated annealing varies one parameter each time in a scale that is proportional to the number of iterations that have already happened (at the beginning the changes are likely to be larger than at the end), and to the variation of the AUC between the two last iterations (if there has been little variation in the AUC, the parameter is more likely to have a bigger variation).

RWR in weighted matrixes

The networks derived from the simple and the weighted addition methods have weighted edges. In order to use this information in the RWR the generation of the matrix is slightly different than for unweighted networks. In this case the transition probabilities from one node to another instead of being purely random are proportional to the weight of the edge connecting the two. This is shown in figure 15 where edges 2-3 and 4-5 now have a weight of 2, whereas all the other edges in the network have a weight of one:



$$\begin{pmatrix} 0 & 0.33 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0.5 & 0 & 0.2 & 0 & 0 \\ 0 & 0.33 & 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0.33 & 0.5 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.2 & 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 0.25 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0.5 & 0 & 0.16 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0.16 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.33 & 0 & 0 \\ 0 & 0.25 & 0.5 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0.16 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.16 & 0 & 0 \end{pmatrix}$$

Figure 15.- Comparison between unweighted (a) and weighted (b) networks and their associated transition matrixes

Figure 15a shows the transition matrix when the weights are not taken into account. In this case, all the transition probabilities are the same and equal to:

$$\boldsymbol{P(t_{i,j}) = \frac{1}{N_i}}$$

Where P($t_{i,j}$) is the transition probability between nodes "I" and "j" and N_i is the number of edges from node "i". As shown in figure 15b, when the weights are taken into account the transition probabilities are directly proportional to the weight of the edge and calculated by:

$$\boldsymbol{P(t_{i,j}) = \frac{w_{i,j}}{\sum_{h=1}^{N} w_h}}$$

Where P($t_{i,j}$) is the transition probability between nodes "I" and "j", $W_{i,j}$ is the weight of the edge between the two nodes, "w_h" is the weight of the edge between nodes "I" and "h" and N is the set of nodes connected to "i".

Identification of "driver" genes

We finally tested the performance of one of our combined networks to predict "driver" genes in cancer. Given that it is quite complex to define whether a gene is a true cancer driver or not, we decided to use one of the most accepted approaches and used mutation frequency data to define a list of driver cancer genes.

Then, we downloaded the mutation frequency data in colorectal cancer from COSMIC version 61. This table contains for each of the 17660 genes that have been found mutated in samples of colorectal cancer (15031 of which can be mapped to ENSEMBL genes), the number of times that the gene has been scanned for mutations and the number of times the gene has been actually identified as mutated.

We then defined a list of true driver genes by selecting those that have been found mutated 15 or more times in colorectal cancer according to COSMIC. This yielded a total of 482 genes. In order to be able to use all the genes in our database as seed, we discarded those that were associated to cancer or colorectal cancer according to our data, so we would only rank "new" driver cancer genes. Of the total 482 genes, 353 where not associated to colorectal cancer and 252 where not associated to any kind of cancer at all in out database. We then evaluated the capability of our combined network H-P-R (result from adding the HPRD, Paralogy and Reactome networks) to predict these 252 and 353 genes using as seed to start the RWR all the genes from our database associated to cancer and CCR genes respectively.

Boosted RWR

Some groups have tried to use external data to improve the results obtained by guilty-by-association approaches. For example, in a recent publication Lee *et. al.* used GWAS data to improve the results obtained when predicting disease-related genes in using their functional network[142].

With the purpose of increasing the predictive power of our method in colorectal cancer we decided to make use of mutation frequency data to generate the initial probability vector. In order to do so, each element of the vector has a probability that is proportional to the number of times the gene has been found mutated in colorectal cancer by applying the following equation:

$$\boldsymbol{P(i)} = \frac{\boldsymbol{N_i}}{\boldsymbol{\sum_{j=1}^{M} N_j}}$$

Where P_i is the i-th element of the vector, N_i is the number of times the gene has been found mutated in colorectal cancer and M is the set of mutated genes in CCR and N_j is the number of CCR mutations found in gene "j". We evaluated the performance of this boosted RWR by trying to predict the 252 and 353 driver genes aforementioned.

Evaluation of the methods

In order to evaluate the performance of the standard and the boosted RWR at predicting driver genes we decided to use 2 different methods:

- Area Under the Curve: As for all the other methods, we calculated the AUC using the ROC curve

- Rank enrichment: We created, for each driver gene, groups of both 10 and 25 genes containing one driver gene and the 9 or 24 closest genes in the genome to the driver gene. We then ranked all the genes in each group according to their scores and observed the rank of the true driver gene

3. Objectives

The main goal of this thesis is:

To identify new disease-associated genes and mutations and their disease mechanisms

In order to accomplish this objective we propose the following work plan:

3.1. Development of a relational database of disease-associated genes and mutations

Available databases containing information about disease-associated mutations do not provide the adequate framework to address our objective. Some are text-based and do not use controlled vocabularies nor ontologies[6,146] to describe the diseases, or rather focus in a single disease[165]. To circumvent this, we wanted to create a new relational database containing information on disease-associated genes and mutations. The database should include automatically extracted features for both, genes and mutations. All terms describing diseases or biochemical properties should be mapped to the corresponding ontologies or controlled vocabularies.

3.2. Identification of disease-specific mutated features

There have been several attempts to identify features associated to pathogenic mutations[166,167]. While each approach has its own biological assumption, such as the degree of conservation through evolution[38] or the protein's structure[37], most of them[168,169] are based in the comparison of the properties of benign single nucleotide polymorphisms (SNPs) and those of disease-associated mutations. While this approach is useful to identify which properties make a mutation pathogenic, subtle differences between mutations associated to different phenotypes might be missed. In order to overcome this limitation, we propose to compare the properties of mutated features in different phenotypes to reveal phenotype-specific mechanisms that help understanding the underlying biology of the diseases.

3.3. Network based prioritization of disease-associated genes

Methods based on the use of biological networks to prioritize lists of putative disease-related genes are becoming very popular[135,142,127]. Though there are several types of networks (representing protein-protein interactions[159] or functional relationships[164] for example) and diverse algorithms depending on different features of the network (such as closeness[127], centrality[170] or topological similarity of the nodes[117]) most approaches are one-size-fits-all in the sense that can't be adjusted to the specific particularities of each phenotype. In this work we also aim to provide a schema of selecting the optimal algorithms, networks and combinations of networks to predict distinct diseases.

4. Results

4.1. Design and development of OCG

As mentioned above, existing public databases of associations between diseases and genes or mutations have certain limitations. For example, some are text-based, which makes it very difficult to perform complex queries, others lack controlled vocabularies and most are focused to single phenotypes.

In order to overcome some of these limitations we created a relational database containing information on disease-associated missense mutations and their biochemical properties. We focused on missense mutations because, unlike nonsense mutations or deletions and insertions, their pathogenic mechanism is in principle more feasible to interpret. Moreover, the lack of standardized description makes descriptions of insertion and deletion mutations extremely hard to correctly interpret and map.

In terms of mutations the database contains 9276 mutations in 2716 genes (figure 16), which is a similar value to that of OMIM (9760 mutations in 1870 genes) and COSMIC v44 (7361 mutations in 2630 genes). We have extracted 12429 mutation-disease associations from OMIM, COSMIC and GAD. The disease coverage in our database is wide and contains 569 different disease terms (supplementary table 1), which grows to 1195 DO terms when including indirect relationships (the 569 terms mentioned before and all their parents).

Regarding biochemical properties we have included 88 terms from SO , 122 including indirect associations. There is at least one SO term for 8529 mutations (figure 16) and a total of 50898 mutation-feature associations.

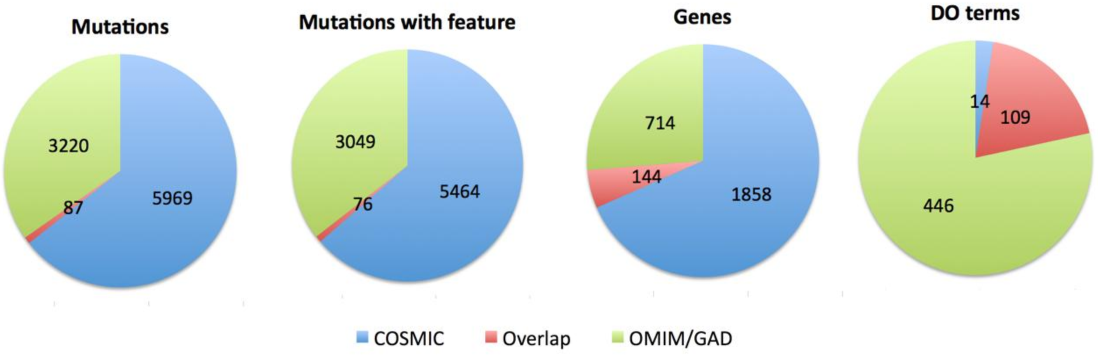


Figure 16.- Summary of the information stored in OCG grouped according to its origin. There is little overlap between the mutations from OMIM/GAD and those that come from COSMIC

Mutations come mainly from COSMIC[165], (6056) and OMIM[6] (3222), therefore the vast majority of the them are associated to cancer or mendelian diseases. While it is important to incorporate mutations associated to complex diseases (the ones that come from GAD[146]), the structure of this repository makes extremely difficult its automatic extraction due to the ambiguous description of the mutations. As a result only 102 mutations were extracted and included in the database from the flat text file provided by the authors[146].

The coverage of the original data varies depending on the repository. For example, while we could successfully extract 6036 missense mutations form the original 7361 present in COSMIC v44 (82%), the numbers for OMIM are smaller: 3222 (33%) out of the 9760 missense mutations in the original text file. In the case of GAD, given that missense mutations do not have a standard description, it is very difficult to estimate the original number of missense mutations. If we take only into account the missense mutations that have their own field in the tabulator-separated file and that are marked as truly associated, there are 1855 possible mutations (which include also insertions, deletions and truncating mutations), thus, the 102 mutations that we recovered, represent at least the 5% of the total.

In order to properly understand the limitations of our data, we estimated the accuracy of the disease annotations. This is an important step, since all the terms describing the diseases associated to the mutations from OMIM and GAD have been automatically mapped to the Disease Ontology[171] (DO) by in-house built parsers. We have calculated that the parsers have an accuracy of 85% by manually checking 100 entries in the database. In the case of COSMIC, the accuracy is 100%, since we manually mapped the DO terms from the anatomical locations of the tumor samples. We successfully identified a DO term for 158 anatomical locations of the total 220 (72%).

Given that 3307 mutations come from OMIM/GAD, and that we have an accuracy of 85%, only about 500 mutations of the whole database are mapped to a wrong DO term. Since we have manually mapped the mutations from COSMIC (which represent 63% of the total number of mutations in OCG), we assume that we have 100% accuracy in those annotations, thus, we estimate that only about 5.4% out of the total 9276 mutations are mapped to a wrong DO term.

The overlap between OMIM/GAD and COSMIC is small and covered by 87 mutations and 144 genes (figure 16) which reflects the different nature of inherited and somatically acquired mutations. The number of overlapping genes is higher than the number of mutations because there are genes for which a mutation is described in OMIM or GAD and another different mutation is described in COSMIC. Overlapping mutations were usually identified in OMIM either as somatic (for example ABL1 F311L, MIM number 189980) or associated to dysplasia or cancer-related syndromes (for instance, APC I1307K, which is associated to familial adenomatous polyposis - MIM number 611731-).

Consistency of the data

In order to check the consistency the data in OCG we looked for diseases, mutations or genes with abnormally high associations in the database by calculating the distribution of the mutations along the genes and the DO terms.

Regarding the distribution of DO terms along the mutations (figure 18), we observed that it follows a logarithmic distribution. The vast majority of the mutations (85%) are associated to a single DO term and 99% of them to five or less. The main outliers were different mutations in KRAS G12 (mutated to different amino acids) and BRAF V600E, which are mutations identified in several different types of cancers in COSMIC.

The same distribution for the mutations is observed along the disease space (figure 17). Of the 569 terms from the DO that have one mutation associated, 76% have less than ten mutations. The exceptions are cancer-related terms such as “colorectal cancer” (DOID:9256) with 1037 mutations , “lung carcinoma” (DOID:3905, with 926), “breast carcinoma” (DOID:3459, with 890) or “skin carcinoma” (DOID:3451, with 768).

Finally, the distribution of mutations along the genes (figure 19) also follows a logarithmic distribution, with 57% of the genes having only one mutation and 88% carrying five or less mutations. The exceptions are genes such as PTEN (299 mutations), EGFR (228), CDKN2A (222), VHL (216) or TP53 (189) that have a strong association with cancer and, thus, have been scanned for mutations more times than other genes associated to less studied diseases.

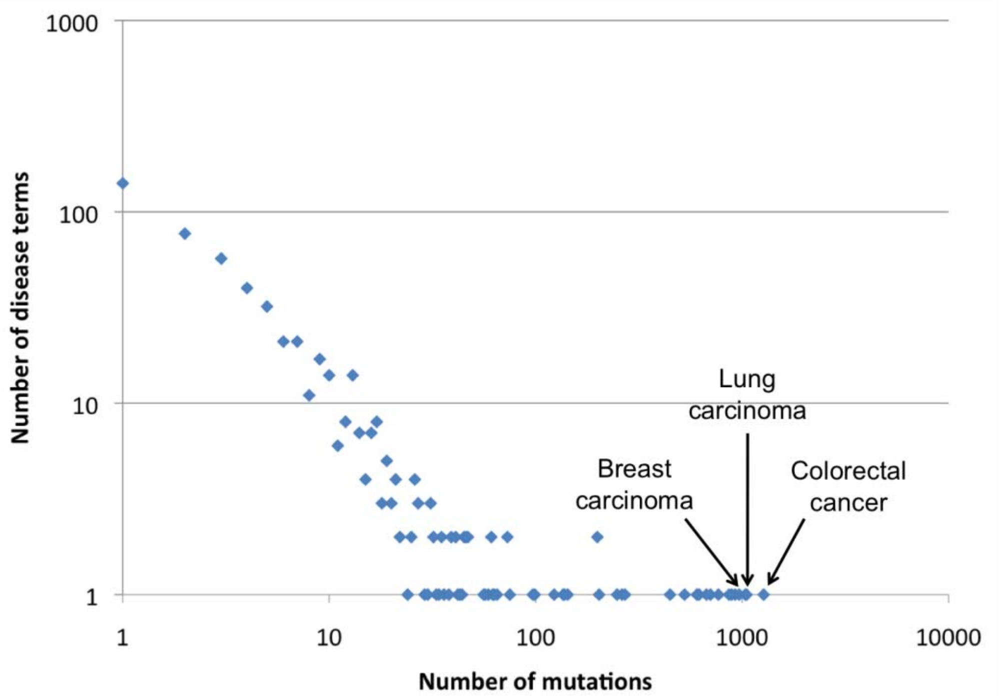


Figure 17.- Number of mutations per DO term. Terms describing cancer phenotypes have the highest number of mutations.

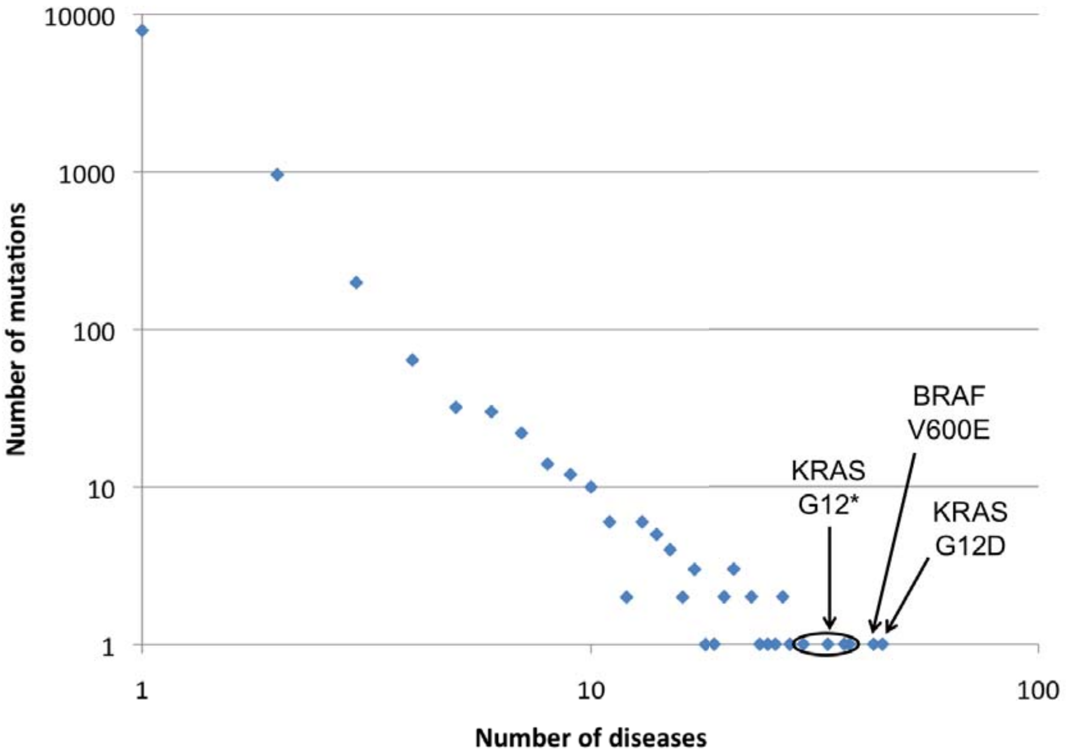


Figure 18.- Distribution of the number of diseases per mutation. Mutations in G12 of KRAS gene (to various alleles) and V600E in BRAF showed the highest number of disease associations.

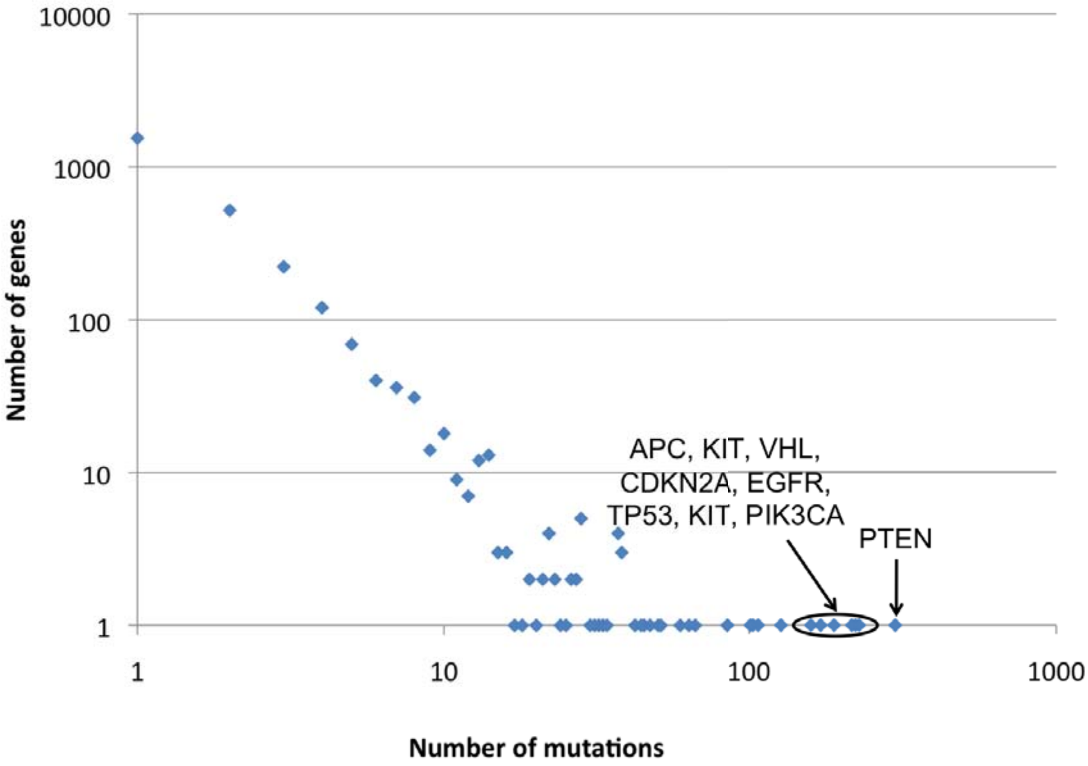


Figure 19.- Number of mutations per gene. Genes strongly associated to cancer show the highest number of mutations.

Next, we reviewed the origin of the mutations that were associated to two or more phenotypes. Of the 1629 mutations identified in two or more phenotypes, 1281 were found only in COSMIC. These are probably associated to different types of cancer (for example, KRAS G12D, associated to “sarcoma” and “breast carcinoma” among 43 other DO terms).

There are 84 mutations that come from both databases (mutations identified in some types of syndrome and cancer samples, or that have been mapped to terms describing similar phenotypes -“melanoma” and “skin carcinoma” for example-) and only 26 were found only in OMIM and GAD. All but four of these 26 mutations were associated to 3 or less phenotypes and in most cases they were associated to related terms. One example of this last case is mutation M196R in gene TNFRSF1B. This mutation was mapped once to “lupus erythematosus” (DOID:8857) whereas in another record from GAD it was mapped to “systemic lupus erythematosus” (DOID:9074). While these two terms have different identifiers, they refer to similar phenotypes.

As a summary, the mutations and their related information seem to have a normal distribution in our dataset, similar to that observed in studies annotating data from OMIM with DO[171]. Moreover, DO terms, genes and mutations with abnormally high number of associations have been manually reviewed and appear to be coherent in the framework they are present.

4.2. Identification of disease-specific mutation features

The normalization of all the annotations for the genes and mutations with terms from different ontologies provides an excellent framework to perform enrichment analysis and identify pairs of ontology terms that are over or underrepresented in disease-associated mutations. In order to do so we used the “elim” algorithm by Alexa *et. al.* as explained in the material and methods section and compared the distribution of annotations in the DO and 3 other controlled vocabularies: GO, PFAM and SO.

DO vs GO

Though the GO terms are not a feature of the mutation itself but of its gene, we thought that it would be a good control of our method because of the extensive use of GO in similar analysis[72].

We obtained 3199 pairs of DO and GO terms that were statistically significant after applying the aforementioned thresholds and the Bonferroni correction. Of these, 2352 were pairs representing an enrichment of mutations (odds ratio > 1) and 847 a depletion of mutations (odds ratio < 1). If we classify the results according to the 3 main branches of GO terms, 1898 involved a GO term of the “Biological process” branch, whereas 806 and 495 involved GO terms from the “Molecular function” and “Cellular component” branches respectively. Finally, by dividing the pairs according to general groups of diseases, we observe that 1609 pairs involved the DO term “Cancer” (DOID:162) or one of its children and 351 contained a term related to “Simple Genetic Diseases” (DOID:0050177).

We observed some pairs that make biological sense and that would be expected from the biology of the disease. For example, we see an enrichment of mutations causing different blood coagulation disorders in genes annotated with the GO term “Blood coagulation”. Another example would be the enrichment of mutations associated to “inborn errors of metabolism”, “inborn errors of carbohydrate metabolism” and “inborn errors of aminoacid metabolism” in genes annotated with the GO term “metabolic process”. This provides evidence that method is able to identify relevant disease-feature pairs.

Regarding cancer-related mutations, we observed an enrichment (table 7) in genes with GO terms related to transcription (“regulation of transcription”, “transcription initiation”, “transcription factor binding” or “transcription factor complex”), the DNA damage response pathway (“double-strand break repair”, “response to DNA damage stimulus”, “base-excision repair” or “DNA repair”) or several pathways previously associated with cancer such as Wnt (“Wnt receptor signaling pathway”), JNK (“JNK cascade”), MAP kinases (“MAPKKK cascade”), SMAD (“SMAD protein complex assembly”) or Ras (“small GTPase mediated signal transduction”). On the other hand, we also found some GO terms that were depleted in cancer mutations (table 3), such as “ion transport”, “glycolysis”, “protein glycosylation”, “fatty acid beta-oxidation” or “integral to membrane”.

Table 7.- Representative associations between DO and GO terms. Terms marked with zero, one, two or three asterisks (*) have corrected P values below 5E-2, 1E-2, 1E-3 and 1E-4 respectively.

GO term	Enriched DO terms	Depleted DO terms
Blood coagulation	Hemorrhagic disorder***, coagulation factor deficiency***	Cancer***
Metabolic process	Lung carcinoma*, glycogen storage disease***, inborn errors carbohydrate metabolism***	Pancreatic neoplasm***, breast carcinoma***, genetic skin disease*
Regulation of transcription	Cancer***, lymphoma***, bone carcinoma**	Disease of metabolism***, simple genetic disease***
Transcription factor binding	Cancer***, bone carcinoma***	-
Double-strand break repair	Lung carcinoma, colorectal cancer***, cancer*	Musculoskeletal system disease
Wnt receptor signaling pathway	Colorectal cancer***, stomach carcinoma***, bone carcinoma	Nervous system disease**
JNK cascade	Cancer***	-
MAPKKK cascade	Skin carcinoma***, endocrine gland cancer***	-
Small GTPase mediated signal transduction	Skin carcinoma***	-
Ion transport	Myopathy***, brugada syndrome***, long QT syndrome***	Cancer***, lung carcinoma***
Glycolysis	Mitochondrial disease***, glycogen storage disease***	Carcinoma
Integral to membrane	Long QT syndrome***, muscular dystrophy***, retinitis pigmentosa***	Colorectal cancer***, kidney neoplasm***, lymphoma***

Interestingly, we observed 89 GO terms that were enriched in mutations associated to certain types of cancer and depleted in others (some are exemplified in table 8). These include “signal transduction”, “metabolic process”, “kinase activity” or “transmembrane receptor activity”. It is particularly in these cases that our approach is more useful, since these differences between cancers cannot be identified when comparing the distribution of pathogenic mutations against a set of random mutations or nsSNP.

Table 8.- DO/GO pairs that distinguish between cancer subtypes Terms marked with zero, one, two or three asterisks (*) have corrected P values below 5E-2, 1E-2, 1E-3 and 1E-4 respectively.

GO term	Enriched DO terms	Depleted DO terms
Signal transduction	Skin carcinoma***, lymphoma***, sarcoma***, cancer**	Kidney neoplasm***
Kinase activity	Lymphoma***, lung carcinoma***, cancer***, sarcoma***	Breast carcinoma**
Transmembrane receptor activity	Lung carcinoma***, cancer	Breast carcinoma**
Angiogenesis	Cancer***, kidney neoplasm***	Breast carcinoma**, lung carcinoma***
Metabolic process	Brain neoplasm***, lung carcinoma*	Breast carcinoma***, pancreatic neoplasm***
Aging	Cancer*, central nervous system neoplasm***	Lung carcinoma*

DO vs PFAM

As observed in the association study between DO and GO, there are several associations between DO terms and PFAM domains that are consistent with the biology of the disease (table 9). For example, mutations associated to “Adenosine deaminase deficiency” tend to fall within the “Adenosine deaminase” domain. Another example is the enrichment of mutations causing “Collagen disease” in the “Collagen triple helix domain”.

Regarding cancer-related mutations, there are several PFAM domains that are consistently enriched in mutations associated to this disease, including “Miro-like”, “Beta-transducin”, “Ras”, “PIP 3 and 4 kinase” or “Protein kinase”. Moreover, there are also some PFAM domains *depleted* in cancer mutations, such as “ABC transporter transmembrane region” (which is consistent with the association found between the DO term “cancer” and the GO term “integral to membrane” or the SO term “transmembrane” which will be discussed in the next section), “Cytochrome P450”, “Sulfatase” or “Intermediate filament”.

In order to identify diseases showing similar mutational landscapes across the different PFAM domains, we generated a heatmap representing the logarithm of the OR between every pair of DO terms and PFAM domains using our mutation data. As explained above, the full heatmap containing all DO terms would be very hard to interpret, so we simplified the number of DO terms until only non-related terms were included in the heatmap and the coverage of mutations was above 90%. Rows and columns were then grouped using a hierarchical clustering algorithm (figure 20).

Table 9.- Examples of statistically significant pairs between DO terms and PFAM domains. Terms marked with zero, one, two or three asterisks (*) have corrected P values below 5E-2, 1E-2, 1E-3 and 1E-4 respectively.

PFAM	Enriched DO terms	Depleted DO terms
ABC transporter	Simple genetic disease***, endocrine system disease***	-
Beta transducin	Colorectal cancer**	-
Collagen triple helix	Genetic skin disease***, collagen disease***, Metabolic bone disease***	Cancer**,
Connexin	Genetic skin disease***	-
Cytochrome p450	-	Cancer***
Dual specificity phosphatase	Cancer***	-
EGF-like domain	Congenital disorder***	Cancer***
Intermediate filament	-	Cancer
Ion transport	Genetic disorder***, myopathy***	Carcinoma***
Ligand binding domain of nuclear hormone receptor	-	Carcinoma
MH2	Colorectal cancer***	-
Miro-like	Skin cancer, lymphoma*, endocrine gland cancer***, abdominal cancer***, bone marrow cancer*	-
Phosphatidylinositol 3- and 4- kinase	Cancer**	-
Kinase	Cancer***, lymphoma***, sarcoma**, skin cancer*, small intestine cancer, stomach cancer***	Breast cancer***
RAS family	Lymphoma**, skin cancer*, endocrine gland cancer***, abdominal cancer*	-
Sarcoglycan	Muscular dystrophy***	-
Short chain dehydrogenase	Disease of metabolism***	-

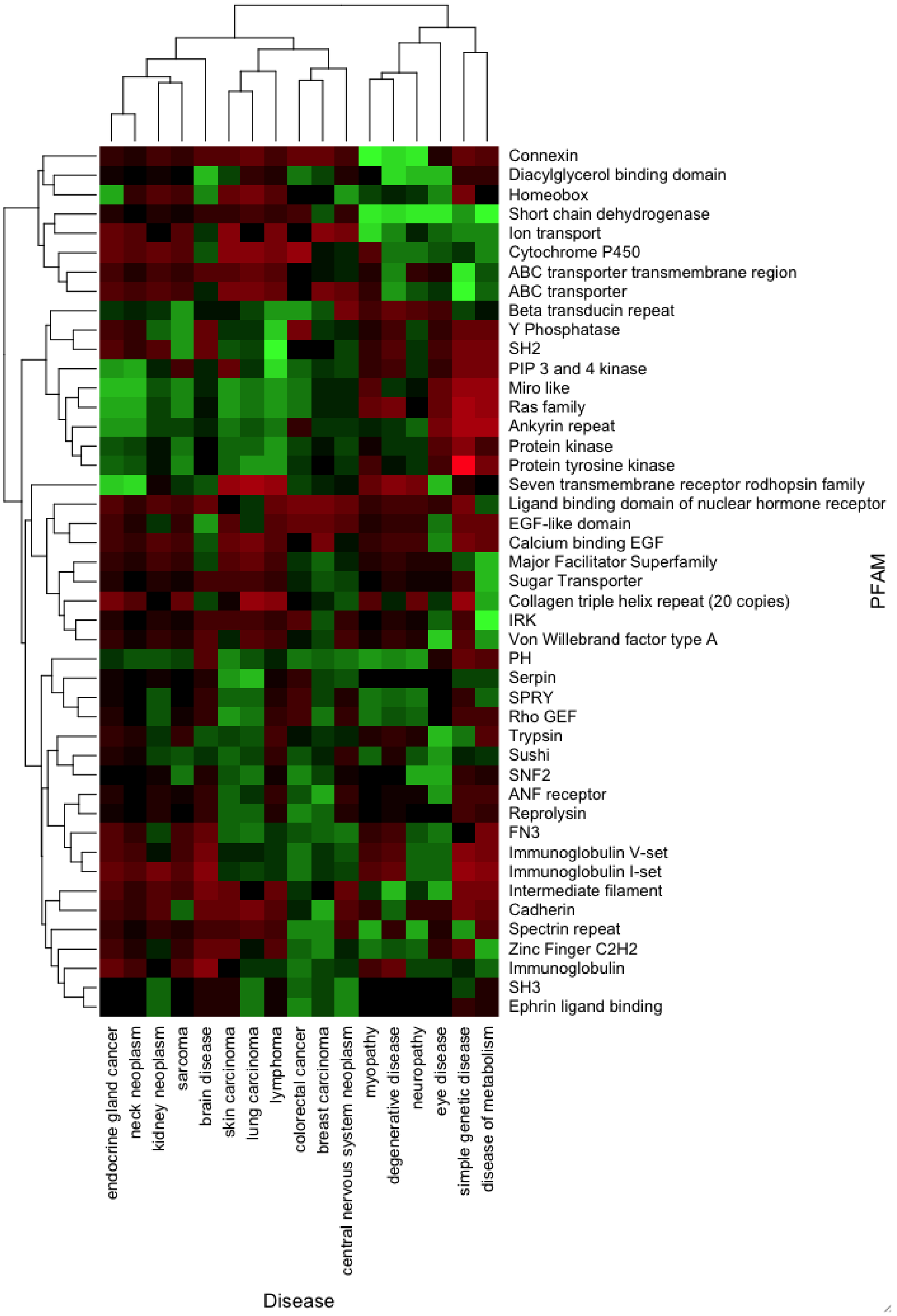


Figure 20.- Heatmap distribution of DO terms and PFAM domains according to the logarithm of the odds ratio. Red values indicate negative OR (depletion of mutations of the pair) while green values indicate positive OR (enrichment of mutations)

The first interesting observation from that heatmap is that all types of cancers cluster together in one group whereas other diseases, but brain disease, cluster in another different group. In terms of domain clustering, there is a group of domains that is clearly enriched in mutations associated to all types of cancers and depleted in most other diseases. This group includes PFAM domains involved in Ras signaling ("Miro-like"or "Ras family"), phosphorylation cascades, ("PIP 3 and 4 kinase", "Protein kinase", "SH2") or protein-protein interactions ("Ankyrin repeat", "Beta-transducin repeat").

There is a group of domains that includes some domains related to extracellular functions ("Serpin", "ANF receptor", "FN3"), others involved in transcription regulation ("ZNF C2H2", "SNF2") or cell adhesion ("Sushi", "Ig I-set" or "Ig V-set") and that separates between two markedly different groups of cancers. One group, formed by "neck neoplasm", "endocrine gland cancer", "kidney neoplasm" and "sarcoma", and the other group, formed by the rest of types of cancers enriched. This is a similar situation than the one observed with GO terms, where some GO terms are able to distinguish between cancer types (section "a" of this chapter).

Finally, there is a group of domains that is consistently depleted in mutations related to most cancers and enriched in some particular phenotypes. These include domains mainly related to membrane transport ("Major Facilitator Superfamily", "Sugar transporter", "Ion transport", "ABC transporter") or extracellular functions ("EGF-like domain", "Calcium binding EGF", "Collagen triple helix repeat").

Mutations in the protein kinase domain and breast cancer

As shown in table 9 and figure 20, though there is a strong association between the term "cancer" and the PFAM domain "protein kinase"[49], we found that there are less mutations associated to "breast carcinoma" (DOID:3459) in this domain than expected.

Interestingly, there are reports describing differences in the distribution of neutral and pathogenic mutations along the kinase domain[53]. To further explore this idea, we looked potential differences in the distribution of the mutations within the kinase domain depending on the cancer type. Aiming for that goal, we built a multiple alignment of all the 163 kinase domains in our database using MAFFT[172]. Then, we mapped all the mutations to the multiple alignment and separated the alignment in 12 canonical subdomains[173] representing the most conserved regions of the domain.

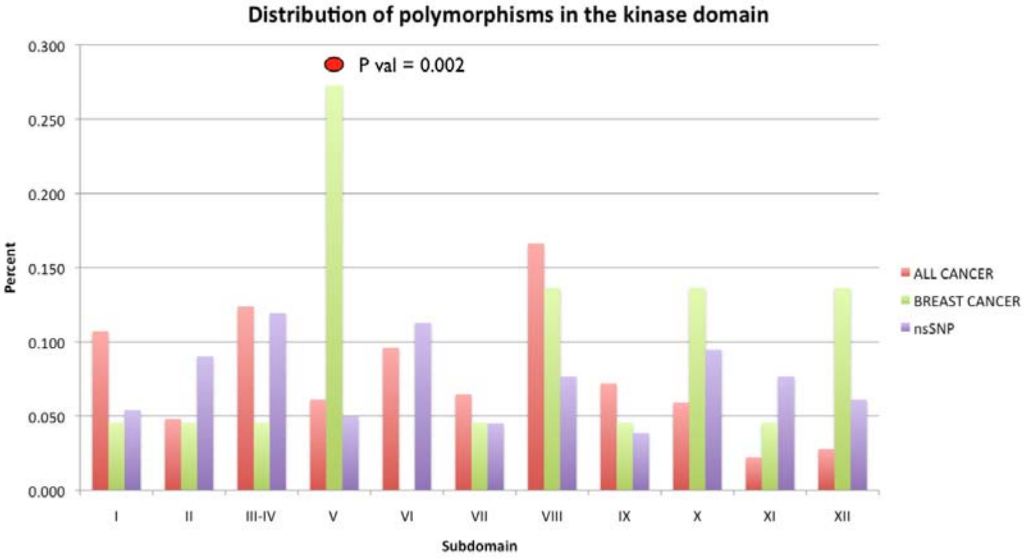


Figure 21.- Distribution of different types of polymorphisms along the subdomains of the kinase domain. Subdomain V shows a statistically significant difference between the fraction of mutations associated to breast cancer and those associated to other types of cancer.

In order to have a neutral model of the distribution of the mutations, we retrieved non-synonymous SNP (nsSNP) from dbSNP for all our disease-associated kinases and did the same analysis. We next calculated the ratio between mutations and nsSNP in each subdomain for different diseases. We found a higher proportion of cancer-related mutations in subdomains I and VIII and a lower proportion in subdomains IX to XI, in line with the results obtained by Torkamani *et al*[53]. However, we found a different distribution of the mutations associated to "breast carcinoma". In this case, we don't observe a higher proportion of mutations in subdomain I, but we do in subdomain, VIII like in the rest of cancers (3 mutations in 3 genes), and also in subdomain V (with 6 mutations in 6 genes), which was not associated to cancer before (figure 21).

DO vs SO

Of all the three controlled vocabularies that we have used to perform enrichment analysis, SO is probably the least explored in biomedicine research. We identified a total of 82 statistically significant DO/SO pairs (supplementary table 9.4), of these 33 referred to the DO term "cancer" or one of its children. These 33 cancer-related pairs included SO terms such as "alpha helix", "beta-strand", "coiled coil", or "polypeptide region", however the biological interpretation of most of them its confusing, since these are very general terms and, thus, they provide very little information, which makes it very difficult to extract biologically meaningful hypotheses.

Yet, after manually reviewing the associations, we selected 6 pairs for further analysis. These involved 3 SO terms that show a statistically significant underrepresentation in cancer: “disulfide bond” (SO:0001088), “peptide localization signal” (SO:0001527) and “transmembrane region” (SO:0001077). Conversely, “serine-rich region” (SO:28, which was added to the ontology by us), “compositionally-biased regions of peptide” (SO:0001066) and “intrinsically unstructured polypeptide regions” (SO:0100003) were overrepresented in cancer-associated mutations. Table 10 summarizes these results.

Table 10 .- SO terms associated to cancer mutations

OR	Feature	Observed number of mutations	Expected number of mutations	P Value
< 1	Transmembrane region	117	235	3,41 e-38
	Peptide localization signal	24	43	1,01 e-6
	Disulfide bond	12	42	3,7 e-14
> 1	Serine-rich region[1]	84	49	2,76 e-11
	Compositionally biased regions	216	162	2,54 e-14
	Intrinsically unfolded regions	1753	1602	1,01 e-13

Association to disease or association to genes

Since disease-associated genes have a bias in some of their properties when compared to the rest of the genome[174,175,62] we wondered whether the SO terms associated to cancer mutations were also associated to the genes. If that were the case, the associations found would not be specific for the mutations, but rather an intrinsic property of the set of cancer-causing genes.

In order answer this question we took for each feature associated to cancer all the genes in our database that had the feature, mutated or not. Then we divided the genes in two groups: genes associated to cancer and genes associated to other diseases, and repeated the two-tailed Fisher’s test. We did not perform this assay in the “intrinsically unstructured polypeptide region” (IUR) as this particular feature is usually described in qualitative terms (longer or shorter) instead quantitative (presence vs. absence), therefore it is difficult to analyze in binary terms as the rest of our comparisons.

[1] Association of SO term “Serine-rich region” is calculated using the DO term “Carcinoma” instead of “Cancer”. If “Cancer” is used instead, the P value is 0,02. All the other associations are calculated using the DO term “Cancer”

In the cases of “transmembrane region”, “peptide localization signal” and “disulfide bonds” we confirmed the associations previously observed: there are fewer genes than expected having the feature mutated when taking into account only those genes that already have the feature (table 11). We could not confirm the associations of those features overrepresented, “compositionally biased regions” and “Ser-rich regions”.

Table 11 .- Analysis of the associations using only genes with the features

OR	Feature	Observed genes mutated in cancer	Expected genes mutated in cancer	P Value
< 1	Transmembrane region	57	104	1,01 e-19
	Peptide localization signal	22	40	5,39 e-7
	Disulfide bond	9	27	1,87 e-9
> 1	Serine-rich region	13	11	0,11
	CBR	81	83	0,68

This analysis showed for the features under-represented in cancer mutations that, even when the feature is present in the protein, for some reason, cancer mutations affect the feature less times than expected when compared to other pathological mutations (the set of mutations associated to other diseases).

In the case of the two overrepresented features, “Serine-rich regions” and “compositionally biased regions”, we did not observe any statistically significant association. In these cases it could be that the association that we are observing for the mutations is caused by a bias in the properties of the genes related to the diseases. In other words, we might be observing more mutations associated to cancer and falling in compositionally biased regions because cancer-related genes tend to have more compositionally biased regions than genes associated to other diseases.

Confirmation of the SO associations with an independent dataset

To confirm the 6 associations found between cancer and SO terms we used an independent dataset of cancer-associated mutations from the Cancer Genome Atlas. This new dataset consisted on 18834 mutations from over 800 samples of 5 different types of tumors: acute myeloid leukemia, colorectal cancer, glioblastoma, ovarian cancer and rectum adenocarcinoma. We could retrieve at least one feature for 14283 (76%) of these mutations.

We then compared the proportions of the 6 SO terms associated to cancer ("disulfide bonds", "transmembrane regions", "peptide localization signals", "Ser-rich region", "compositionally biased regions" and intrinsically unstructured regions") in this new dataset of cancer mutations with that of the mutations associated to other diseases than cancer in our original dataset (table 12).

Table 12 .- Comparison of the independent dataset of cancer mutations from CGA with mutations associated to other diseases than cancer in OCG

OR	Feature	Observed number of mutations in CGA	Expected number of mutations in CGA	P Value
< 1	Transmembrane region	787	880	1,8 e-6
	Peptide localization signal	98	120	2,5 e-4
	Disulfide bond	82	116	2,8 e-9
> 1	Serine-rich region	103	110	0,06
	CBR	1602	1395	6,0 e-100
	IUR	6769	6381	4,2 e-150

This analysis showed that there are fewer mutations than expected in "disulfide bonds", "transmembrane regions" and "peptide localization signals" in the set of mutations from the CGA when compared to the mutations associated to other diseases than cancer (table 12). Notice that in this case there are fewer differences between groups of mutations (table 10). This could be caused by an increased number of passenger mutations in the CGA compared to the version of COSMIC that we used.

Moreover, we also observed more mutations than expected in "intrinsically unstructured regions" and "compositionally biased regions" in the set of CGA when compared to the set of mutations associated to other diseases (table 12). Finally, we found no difference in the proportion of mutations falling in "Serine-rich regions" in mutations from the CGA when compared to mutations associated to other diseases.

Internal associations between SO terms

Some biochemical features can overlap in a protein, thus, a mutation can map to more than 1 SO term. This is an important aspect to take into account in our analysis, because those SO terms that overlap more than expected might be providing the same information, and thus be less relevant for our results.

In order to assess the influence of this event in our dataset we performed an enrichment analysis looking for pairs of over or underrepresented SO terms. We successfully identified 87 pairs of SO terms at ratios higher or lower than expected (supplementary material 9.5), indicating that they are providing similar information in the first case and that they are mutually exclusive in the latter.

Interestingly, mutations falling in “intrinsically unstructured regions” tend to be located also in “serine-rich regions” (pvalue below 1E-15) and “compositionally biased regions” (pvalue below 0.01) more often than expected, which is consistent with previously reported data[176,177]. This implies that the overrepresentation of these 3 terms in cancer mutations is probably reflecting the same trend. Another example is mutations in “phosphorylation” sites that also happen more than expected in “intrinsically unstructured regions”[178] (pvalue below 1E-32).

Illustrative examples

Our data suggests that mutations in the same gene might be causing different phenotypes not only because they are occurring in different domains, as previously described[109], but also because they alter different biochemical features.

Interestingly we found in our dataset 4 genes exemplifying this phenomenon. These genes have two distinct groups of mutations, one group causing cancer-related disorders and the other some other type of disease (figure 22)**.**

For example, gene CACNB2 (a gene that modulates G-protein inhibition) is associated to colorectal cancer by a mutation located in a compositionally-biased region and to Brugada syndrome (a heart disease) by a mutation located outside of that type of region.

Something similar happens with SCN3B, another voltage channel modulator with a different domain composition, and the same diseases, but in this particular case, the mutation associated to Brugada syndrome is located at the peptide localization signal, whereas the two mutations associated to colorectal cancer are located outside this region.

In the case of FZD4 (a G-coupled receptor for Wnt proteins) the mutation causing retinopathy of prematurity falls within the transmembrane domain of the protein, while the mutation associated to colorectal cancer is outside of this region.

Finally, the mutation in CD40, another receptor, causing immunodeficiency with increased IgM is located in a cysteine involved in a disulfide bond and the mutation related to skin carcinoma outside of it.

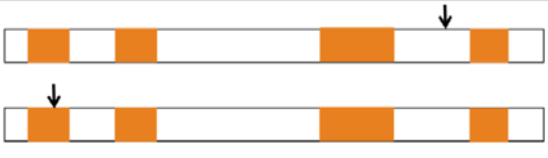
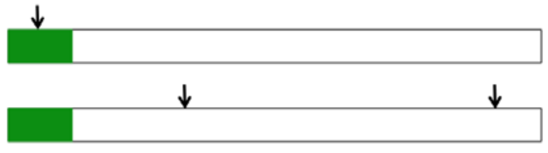
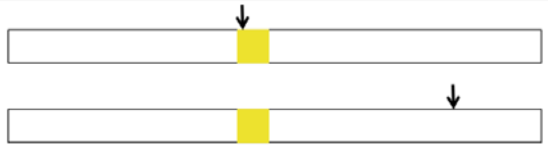
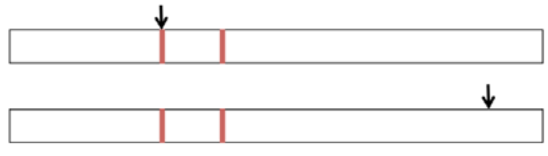
Feature	Gene	Phenotype	Distribution
Compositionally biased region	CACBN2	Brugada syndrome Colorectal cancer	
Peptide localization signal	SCN3B	Brugada syndrome Colorectal cancer	
Transmembrane region	FZD4	Retinopathy of prematurity Colorectal cancer	
Disulfide bond	CD40	Immunodeficiency with increased IgM Skin carcinoma	

Figure 22.- Illustrative examples of the associations. These four genes are associated to two phenotypes, one of them being cancer the other a different disease. Arrows indicate the mutations, and the colored region the feature.

4.3. Network based identification of disease genes

Proteins do not perform their functions alone, but in cooperation with other proteins. This information could be used to predict disease-related genes. To this purpose several groups have used ideas derived from graph theory, such as the use of bipartite graphs[133], or algorithms like the direct neighbor[119], diffusion kernels or random walks[179].

While different algorithms and networks have been used in the quest for disease genes, so far, to the best of our knowledge, no systematic study looking for differences in their performance depending on the disease has been attempted. Here we will aim to identify the optimal network/algorithm combination for different diseases.

Design and development of CCBG

This database contains information on disease-related genes and several types of relationships between them, including physical interactions, coexpression in healthy tissues, paralogy or belonging to the same biological pathway. As in OCG, all the genes are univocally identified using ENSEMBL gene identifiers and the terms describing the diseases have been mapped to the DO using the same parsers that were used to build OCG.

The final database contains 17696 genes. Of these, 3240 (18%) are associated to 666 different DO terms (1343 including their parents). It is important to notice that the number of gene-disease associations in CCBG is higher than in OCG because in this case, there is no need for a particular known mutation in the gene to be store the association.

There is a remarkable overlap of nodes between networks, with 9054 genes (51%) appearing in two or more networks and 578 (3%) appearing in all four (figure 23). Considering each network alone, the one derived from paralogy data has the most unique nodes (nodes that are only present in this network): 6543 out of 14580, which represents a 45% of the whole network. In the other networks this percentage goes down to 20% in the coexpression network, 8% in the metabolic and 5% in the physical interaction network.

Regarding interactions, there are a total of 912796 relationships among genes, but only 3245 (0.3%) appear in more than one network (figure 24) and only a single edge is present in all four networks, so overall, there is little overlap among networks. This is probably a reflection of the different biology that each type of relationship represents. The network with the most shared edges, in terms of percentage, is the one from physical interaction data, with 6%. For the pathway network, the paralogy network and the coexpression network this figure is 4%, 2% and 0.3% respectively.

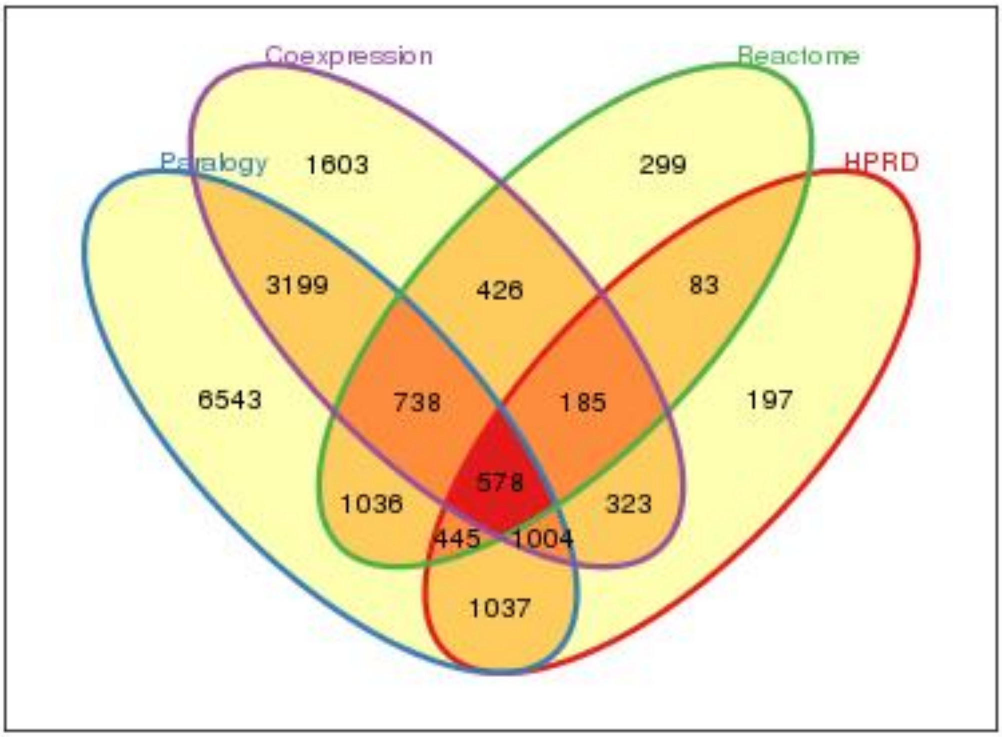


Figure 23.- Venn diagram representing the overlap between nodes (genes) in the different networks

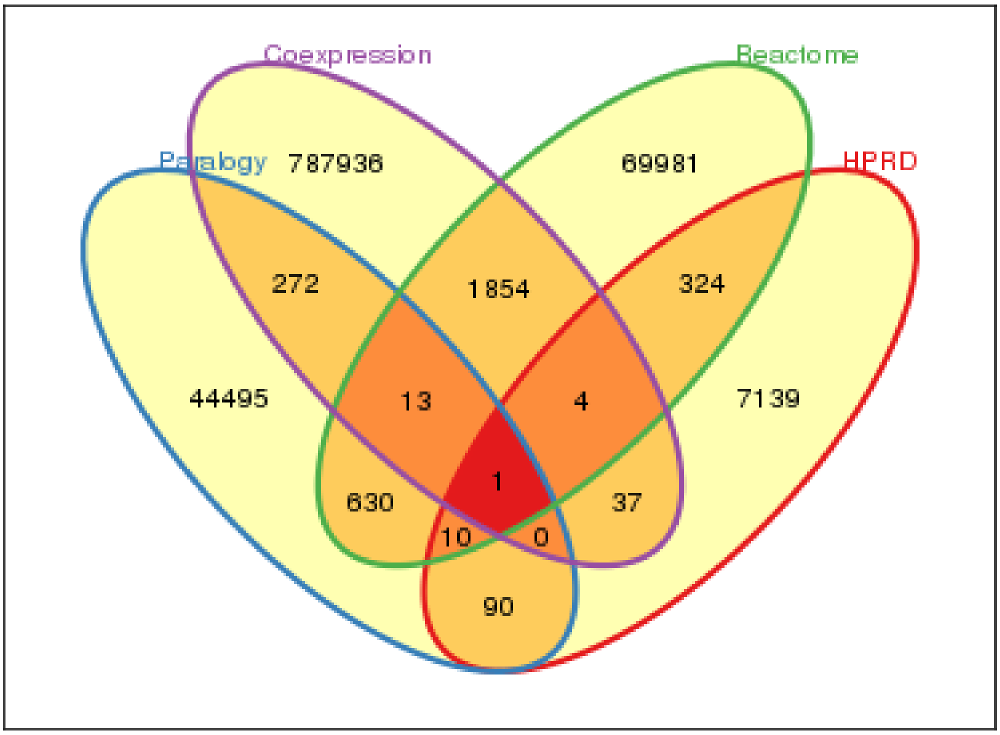


Figure 24.- Venn diagram representing the overlap between edges (connections) in the different networks

Topological characterization of the networks

We next assessed the topological properties of the networks to check their consistency with previous descriptions of biological networks. As shown in figure 25 the networks derived from HPRD and BioGPS follow a power-law distribution of their node's degrees.

Interestingly, the networks derived from paralogy and metabolic information show a different distribution but for different reasons. In the case of the paralogy network, its node's degree distribution is caused by the fact that the nodes are forming cliques (is a subset of nodes of a network that are connected all with each other) of gene families that are not connected between them. The degree distribution suffers a sudden drop at about 50, which means that there isn't any gene family with more than 50 members. This is clearly exemplified in figure 26 where we can see a representation of all networks with nodes colored and scaled according to their degree.

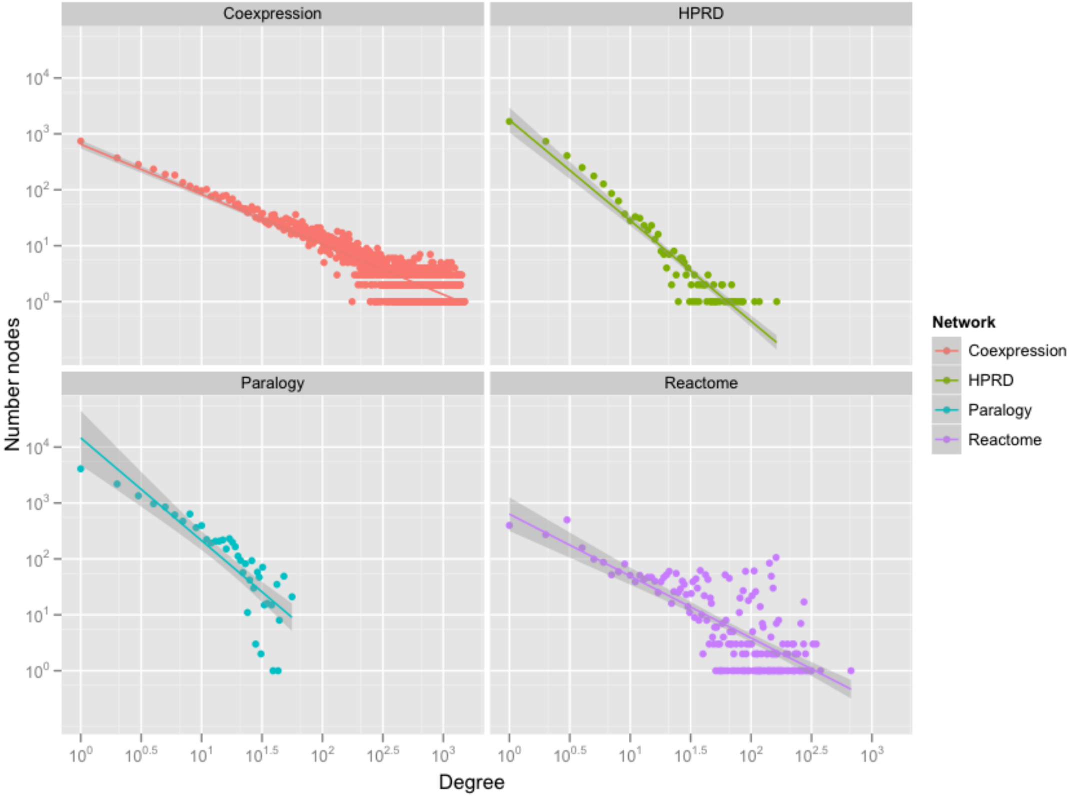


Figure 25.- Degree distributions of the 4 networks.

The network created from metabolic and pathway information from Reactome deviates also from the power-law distribution but in this case this is due to the definition of the network. In this network an edge connects two nodes when they belong to the same biological pathway, thus cliques are formed between genes sharing the same pathway. However, unlike the paralogy network, edges connecting cliques between them appear every time a gene belongs to more than a single pathway (figure 26).

Topological properties of the networks are summarized in table 13. The network with the biggest size is the paralogy network, whereas the smallest is the one derived from Reactome pathway data. The coexpression network is the one with highest density.

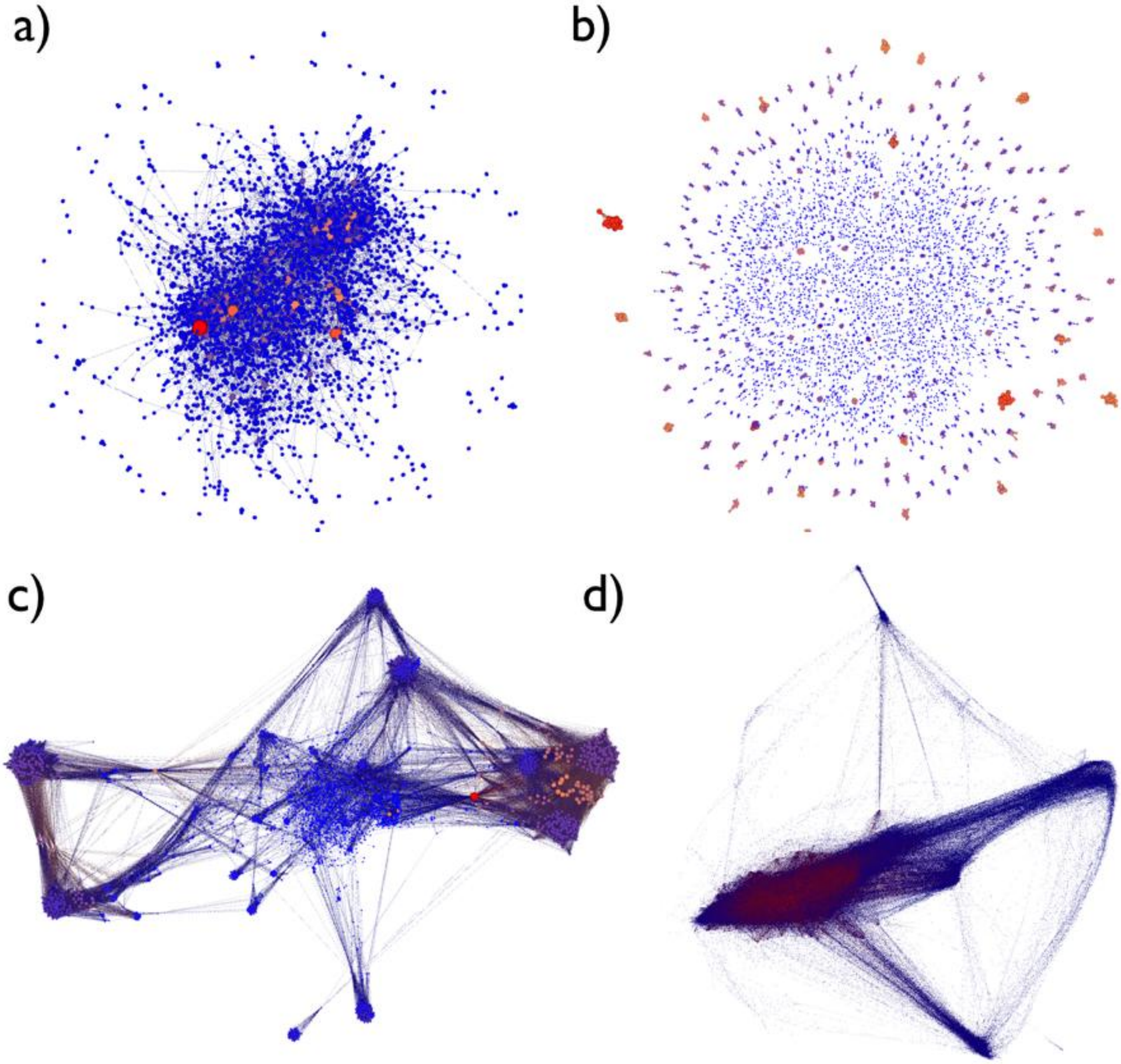


Figure 26.- Representation of the four different networks. Nodes are colored and sized according to their degree. Nodes with higher degrees are colored red and are bigger, while nodes with lower degrees are colored blue and are smaller. (a) Representation of the network derived from physical interactions. Most nodes are and are forming a single cluster. (b) Network derived from the paralogy data. Nodes are forming unconnected cliques that represent the different gene families. (c) Representation of the metabolic/pathway network. There is a high density of edges and cliques are connected through genes that belong to more than one pathway. (d) Coexpression network. Most hubs in this network (coloured red) are clustering together.

Table 13.- Topological properties of the different networks

Parameter	HPRD	Paralogy	Reactome	Coexpression
Number of nodes	3852	14580	3790	8056
Density	0.001	0.000	0.010	0.024
Av. degree	3.95	6.24	38.43	196.15
Av. path length	4.95	1.25	4.87	3.96
Diameter	13	2	20	14
Av. clustering	0.088	0.636	0.622	0.559

The paralogy network has the lowest diameter despite having the biggest size and the lowest density. This is because in this network there are a very large number of unconnected components. This implies that the biggest distance in this network has to be calculated inside the strongly interconnected families of genes (with an average path length of 1.25), thus, the maximum distance is only of 2.

Consistency of the disease-gene associations

In order to check the consistency of the disease-gene associations we calculated the number of genes per disease. As shown in figure 27, gene-disease associations in CCBG show the same logarithmic distribution that was observed in OCG (section 4.1). Over 90% of the disease terms (612 out of 676) are associated to 10 or less genes. The diseases with most genes are several types of cancers and diabetes (including types I and II).

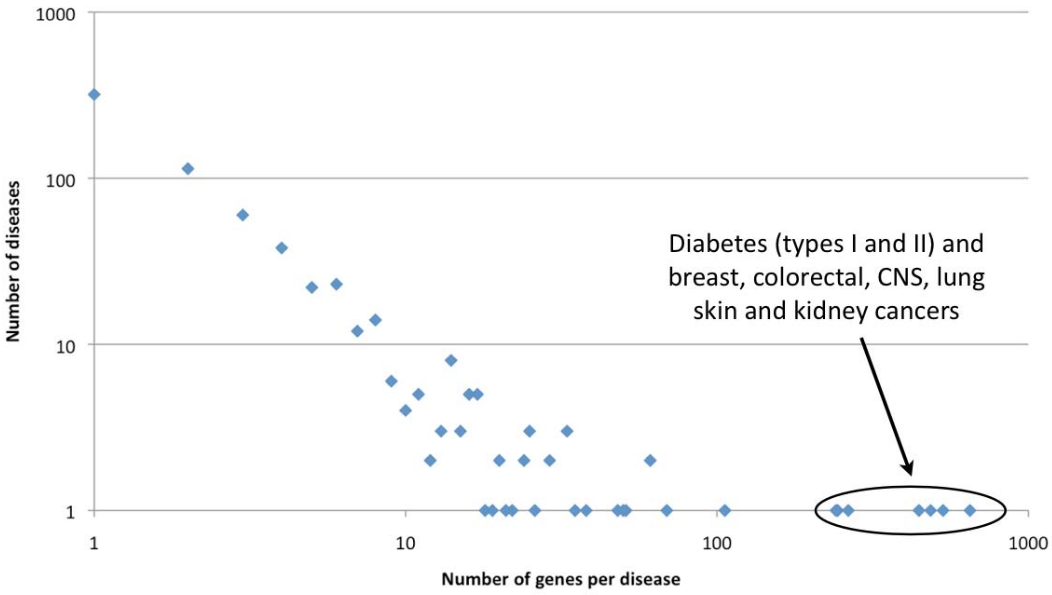


Figure 27.- Representation of the gene-disease associations. As in the case of OCG, associations in CCBG show a logarithmic distribution. Diseases with the higher number of genes associated are highlited.

Selection of the disease models

We next looked for 5 different DO terms to be used as models to evaluate the different algorithms and networks. In order to do so, we searched DO terms that are supposed to have differences in their underlying biology and that had at least 20 genes in each network, so that we would be able to perform cross-validation to estimate the performance of all the algorithm/network combinations (table 14).

After evaluating these two criteria for several diseases, we finally decided to use the DO terms “Simple genetic diseases”, “Neurodegenerative disorders”, “Myopathy”, “Cancer” and “Colorectal cancer” as disease models. Not only they have more than 20 genes in each network, but also from the analysis of the features associated to their associated mutations (section 4.2), we had evidence that these disorders are caused by alterations in different PFAM domains. This suggests that the network alterations, at least in the protein-protein interaction network, might be different [109].

Table 14.- Total number of genes associated to each disease in each network

Network	Cancer	Neurodegenerative disorders	Simple genetic diseases	Diabetes
Physical interactions	832	50	84	87
Paralogy	1784	118	198	181
Pathway	638	60	147	116
Coexpression	1102	80	151	126

As shown in table 15 and supplementary material 9.6 there is little overlap in the genes associated to each disease (with the obvious exception of “colorectal cancer” which is included in “cancer”, as it is a child of this term and, thus, all of its genes are also associated to this term), which also suggests that the underlying biology of each disease is likely to be different. We decided to include also colorectal cancer in our group of case diseases because it would be interesting to see if there are any differences in the performance of the different networks and algorithms between this term and its parent cancer.

Table 15.- Percent of unique genes for each disease in each network

Network	Cancer	Neurodegenerative disorders	Simple genetic diseases	Diabetes
Physical interactions	94%	62%	54%	55%
Paralogy	95%	68%	65%	66%
Pathway	90%	53%	61%	61%
Coexpression	95%	64%	66%	68%

Prioritization of disease-associated genes

Most network-based approaches to predict novel disease-associated genes use the same network and the same algorithm regardless of the disease. We wondered if it would be possible to find an optimal network/algorithm combination and whether this combination would be different for different diseases. We first looked for variations in the capability of 4 types of networks and 5 different of algorithms to predict the genes associated to the 5 distinct diseases, for a total of 100 combinations. In order to do so we calculated the area under receiver operating curve (ROC) for each combination in a 5-fold cross validation experiment.

With this data we next tried to answer 3 different questions:

- Which network works better?
- Which algorithm works better?
- Which disease is better predicted?

In order to answer these questions we generated three different boxplots, each one comparing a different variable and all of them representing the different 100 combinations.

Which network works better?

Figure 28 shows the changes in the AUCs when predicting genes associated to the same disease, using the same method but changing the network. The first remarkable observation is that there is no single network that stands out from the rest. The performance of each network seems to be strongly dependent on the algorithm used and the disease that is being predicted.

Analyzing the results per method some interesting trends in each network appear. For instance, when using the direct neighbor method with distance 1, the Paralogy and Reactome networks usually perform the better. For example, in cancer the average AUCs of the Paralogy and Reactome networks are 0.64 and 0.60 respectively, compared to 0.55 for the HPRD and 0.51 for the coexpression network. In the case of colorectal cancer the values are 0.60 and 0.59 for the Paralogy and Reactome networks and 0.54 and 0.53 for HPRD and coexpression.

However, when using the same method, but up to distances to 2 or 3, the differences change and the Reactome network performs much worse, being the worst network in diabetes (AUC of 0.57 for DN2 and 0.53 for DN3) and simple genetic diseases (AUC of 0.49 when using DN2 and 0.42 with DN3). Interestingly, when using the two diffusion methods, DK and RWR, the performance of the Reactome network in simple genetic diseases, cancer and in neurodegenerative disorders improves again. For example, in the case of simple genetic diseases this network goes from an AUC of 0.42, when using the DN3 method, to 0.64 if the method is DK and 0.63 if it is RWR. Thus, the Reactome network seems to benefit of using diffusion methods that exploit its whole topology to explore longer distances.

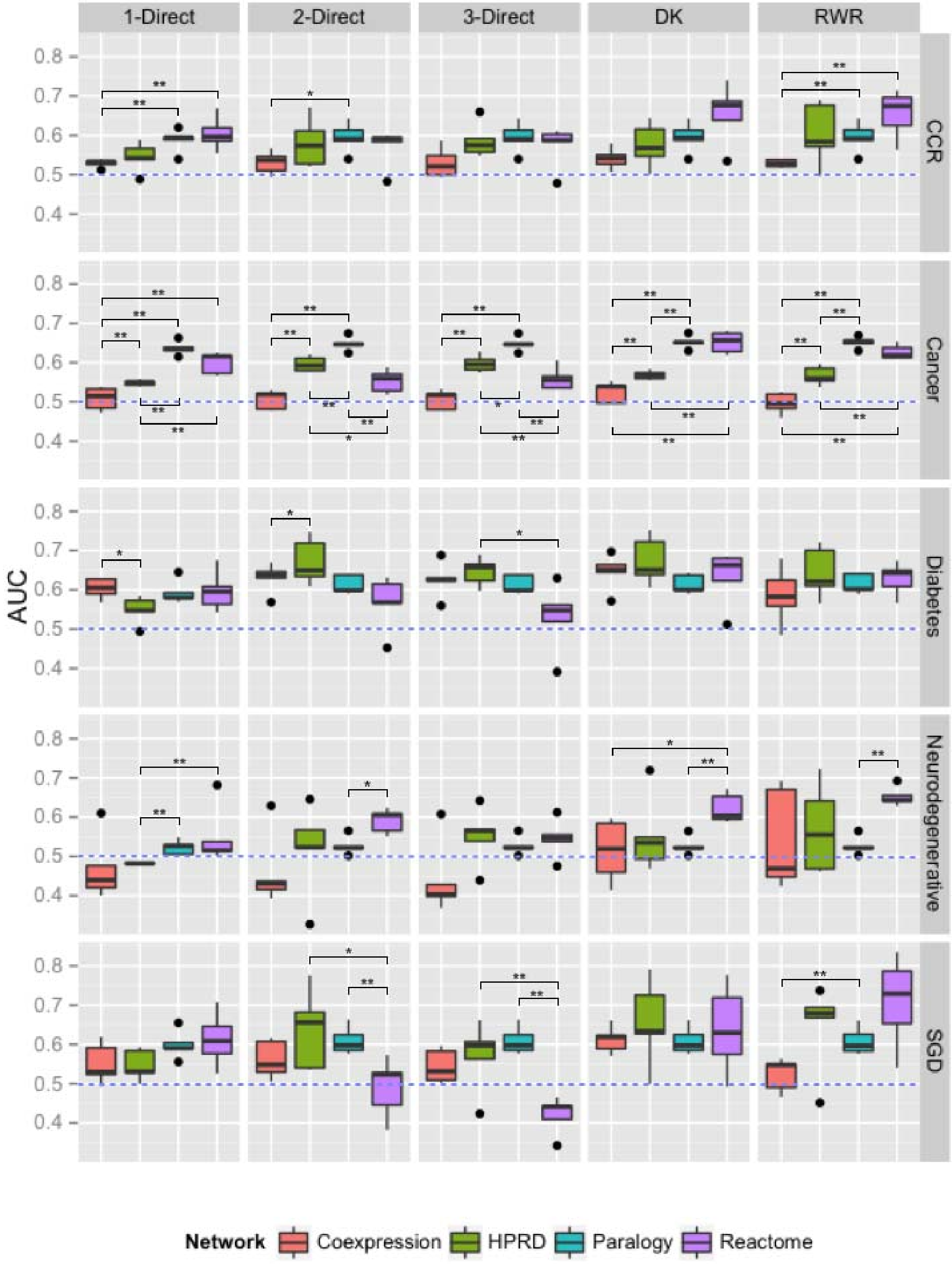


Figure 28.- Comparison of the results obtained in each network grouped according to the predicted disease and the method used. Pairs marked with one or two asterisks indicate p-values below 0.05 and 0.01 respectively. The dashed purple line marks the 0.5 AUC value threshold.

Analyzing the results grouping them per disease shows also some tendencies. When trying to predict genes associated to colorectal cancer, the coexpression network is the worst when using 3 of the 5 methods. In the case of cancer, the paralogy network is the top performer regardless of the method (average AUC of 0.65 in all methods), joined by the Reactome network when using the direct neighbor in distance 1, and the two diffusion methods.

In the case of diabetes, there are not many differences, except in the case of the direct neighbor with distance 2, where the HPRD network seems to perform better (average AUC 0.67), and the same method with distance 3, where the Reactome network performs the worst. In the case of neurodegenerative disorders, the Reactome network seems to be better than the others, whereas in simple genetic diseases, if one uses the direct neighbor with distances 2 or 3, this network is the worst, as explained above. Finally, in the case of simple genetic diseases, there are not many differences, but it seems that the Reactome and HPRD networks are the best when using the diffusion methods (AUCs of 0.64 and 0.66 for the Reactome and HPRD networks respectively using the DK and of 0.65 and 0.71 when using the RWR).

Wrapping up, it seems that networks derived from different types of biological information perform differently at predicting different diseases, thus it would be interesting to use networks derived from other information than protein-protein interaction data to predict genes associated to diseases.

Which algorithm works better?

We next compared the performance of the different algorithms at predicting the disease genes when using the same network. This is an interesting question, since diffusion based methods have been claimed to perform better than direct neighbor algorithms[115,180]. However, figure 29 shows that there are little differences in general and in few cases the difference between the two groups is significant.

The network that benefits the most from diffusion methods is the Reactome network. When using this network to predict disease-related genes in 4 of the 5 diseases the diffusion methods outperform the DN methods. In the only disease where the differences are not significant, colorectal cancer, a strong tendency in the same direction can also be appreciated (the AUCs for the DN methods are 0.60, 0.55 and 0.56 for distances 1, 2 and 3 respectively, whereas the AUCs for DK and RWR are 0.65 and 0.63 respectively).

Another important observation is that, in the case of the paralogy network, the choice of the algorithm is not important. In figure 29 we can see that, in all the diseases, the predictive power of this network doesn’t change with the algorithm. This is probably caused by the particular topology of this network, which is formed by several independent cliques (the average shortest path of the network is 1.25), but no connections at all between cliques. Thus, in this case, exploiting the full topology of the network does not provide any advantage compared to simply count the neighbors associated to the feature of our interest.

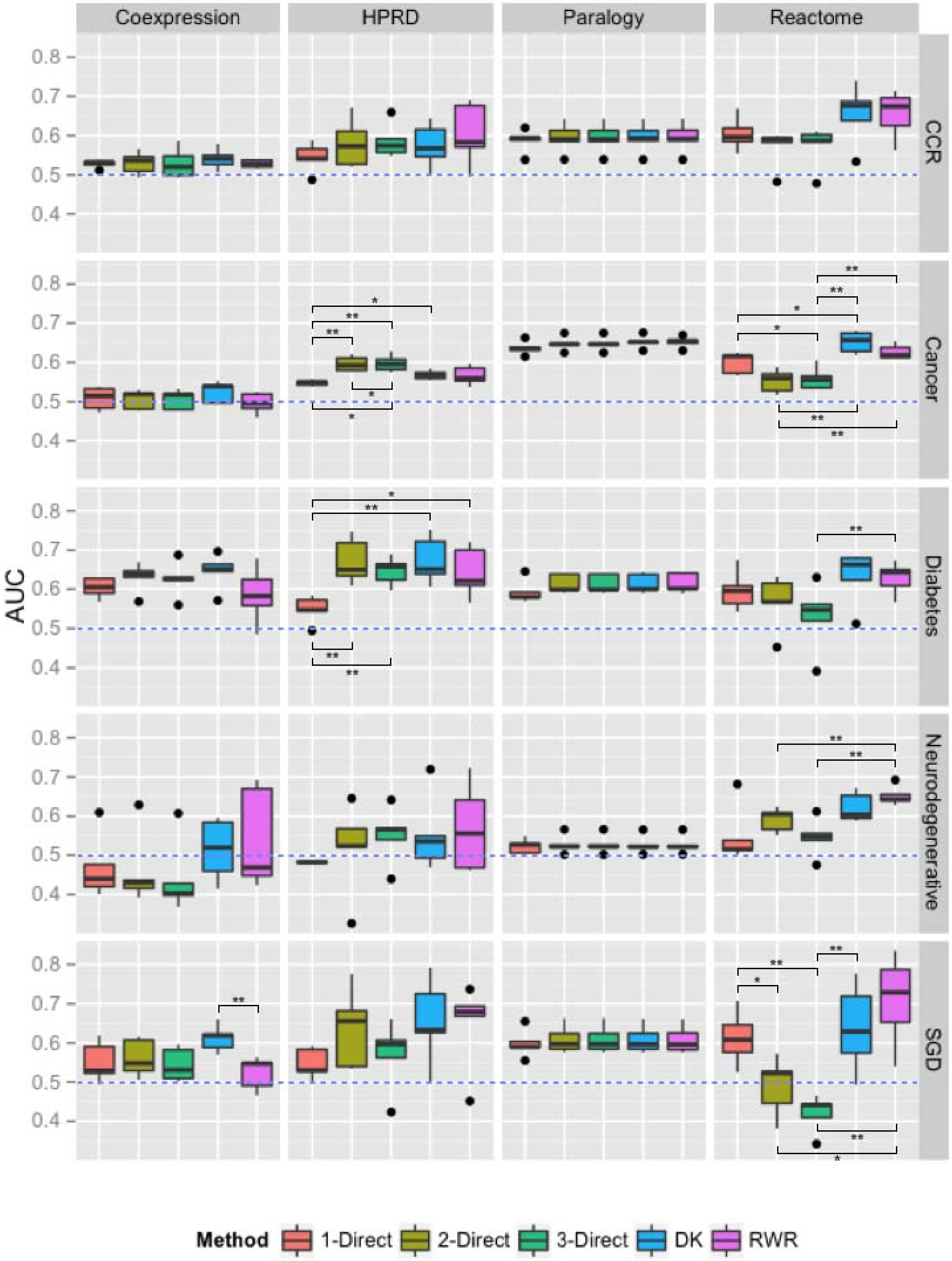


Figure 29.- Comparison of the results obtained in each method grouped according to the network used and the predicted disease. Pairs marked with one or two asterisks have P values below 0.05 and 0.01 respectively. The dashed purple line marks the 0.5 AUC threshold.

As a summary, our results point to the fact that the superior performance of diffusion methods is dependent on the network being used to make the predictions and that the performance of the different algorithms seems to be strongly influenced by the topology of the network.

Which disease is predicted better?

We finally evaluated which diseases were predicted better using each method and each network (figure 30). The coexpression network seems to work only to predict genes related to diabetes (average AUC between 0.59 and 0.65 depending on the method) since most values for the other diseases range around 0.50.

Regarding the HPRD network, it seems to be able to predict with similar performance all the disease-related genes with the exception of those associated to neurodegenerative disorders (average AUC between 0.48 and 0.56, whereas for all the other diseases the range is between 0.55 and 0.67). The disease that seems to be better predicted this network is diabetes, which is the top performer when using 4 of the 5 algorithms.

When using the paralogy network, cancer genes are predicted the most successfully (average AUC of 0.65 in all methods), whereas those associated to neurodegenerative disorders don't seem to benefit from this network (average AUC 0.53). Finally, though there aren't any clear trends, the Reactome network seems to poorly predict genes associated to simple genetic diseases when using the 2 and 3 direct neighbor methods (AUCs of 0.49 and 0.42 respectively), but apart of this, there is little variation between diseases regardless of the method.

The group of disease genes that seems to be predicted better is the one associated to simple genetic diseases. In eight out of the total 20 network/algorithm combinations this disease is among the best predicted, and the highest AUC value of all combinations belongs to this disease when predicted using the Reactome network and RWR (0.71). This is important to take into account as most methods are evaluated using only gene-disease associations from OMIM, which are the strongly represented in this group, and it seems that predicting genes associated to other phenotypes is more complicated, as their AUCs are lower.

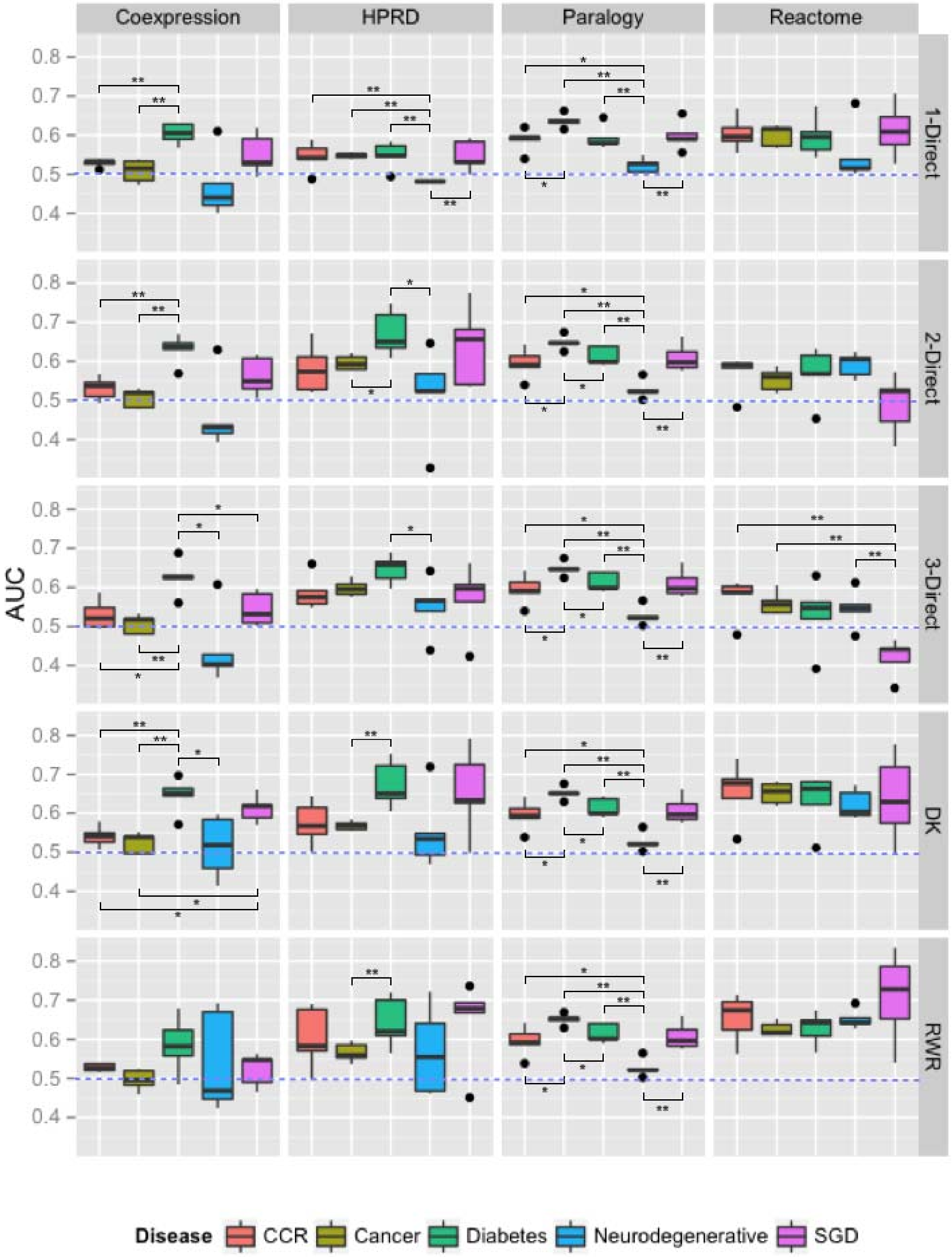


Figure 30.- Comparison of the results obtained in each disease grouped according to the network and the method used. Pairs marked with one or two asterisks have a P value below 0.05 and 0.01 respectively. The dashed purple line indicates the AUC 0.5 threshold.

Correlations between networks

We next wondered whether different networks were predicting different genes. If that were the case the integration of the networks would not be likely to increase the performance of the networks alone. However, as shown in figure 31, there is no correlation between the rank of cancer associated genes in one network with the ranking of the same genes in another using the direct neighbor 1 algorithm. The same is true for the other diseases and methods tested (supplementary material 9.7).

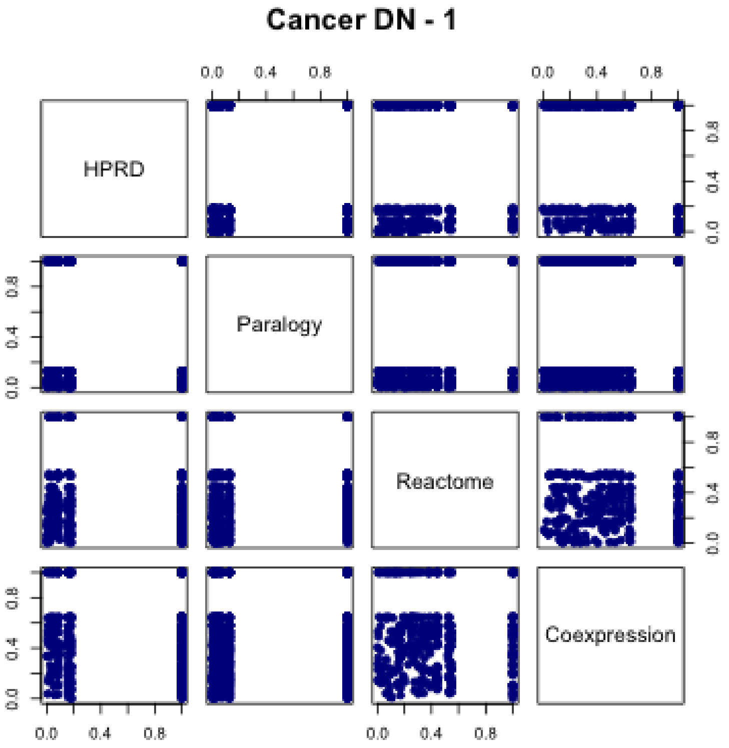


Figure 31.- Correlation between the rank for candidate cancer genes in the different networks using the direct neighbor algorithm with distance 1. There is no correlation between any network, indicating that each type of network is predicting different genes.

An interesting observation that arises from these correlation plots, is that the dispersion of the points, which is a reflection of the variability of the scores, correlates with the average degree of the network and how far the method explores the network. Thus, simpler methods like direct neighbor with distance 1, have a lower variability of scores than more complex methods like DK or RWR. The same happens between the HPRD network, which has a lower average degree and a lower distribution of score, than the coexpression or Reactome networks.

Influence of the different parameters in the performance of the RWR

After evaluating the different methods, not only in terms of their predictive power, but also of speed and ease of use, we decided to only use the RWR to evaluate the different network combination methods. The performance of this method depends on three different factors: the restart probability, the topology of the network and the initial probability vector. Given that this method would be the only one used for the next part of the project, we decided to investigate the influence of these three factors more in detail.

Restart probability

We compared the AUC obtained in each network when predicting each disease and varying the restart probability between 0.1 and 0.9 in 0.1 intervals. As shown in figure 32, though some tendencies may be observed, there are no statistically significant differences in any network/disease combination.

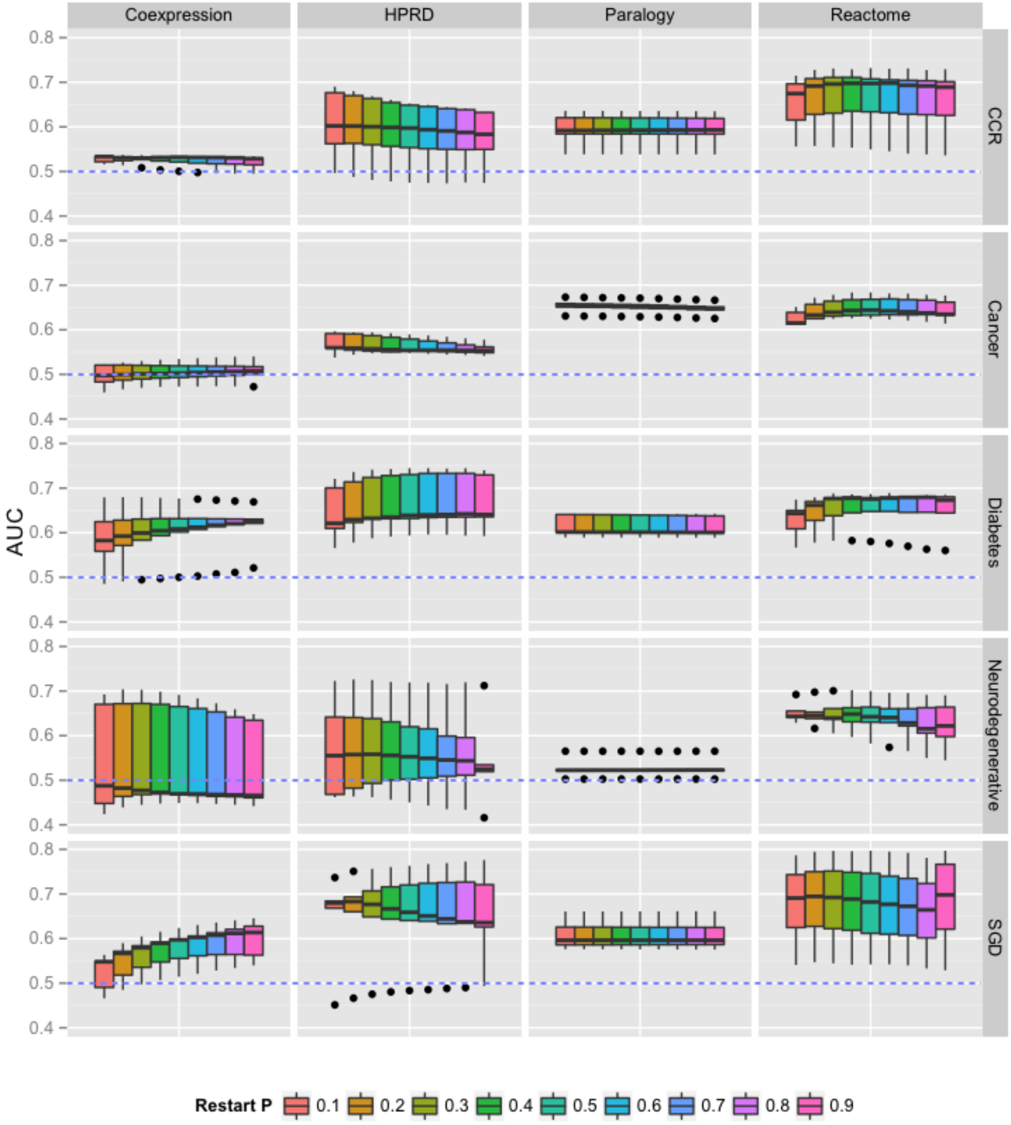


Figure 32.- Influence of the restart probability when predicting each disease depending on the network. There are no statistically significant differences in any case. The dashed purple line marks the AUC value of 0.5

After analyzing these results, we decided to keep the restart probability at 0.1 in order to explore longer distances in the network.

Initial probability vector

The main idea of the RWR is to calculate the “closeness” of all the genes in a given network to a set of “seed” or “initial” genes that are supposed to have some biological meaning. These genes are the ones that have non-null equal probabilities of being a starting point of the random walk, thus, define the initial probability vector.

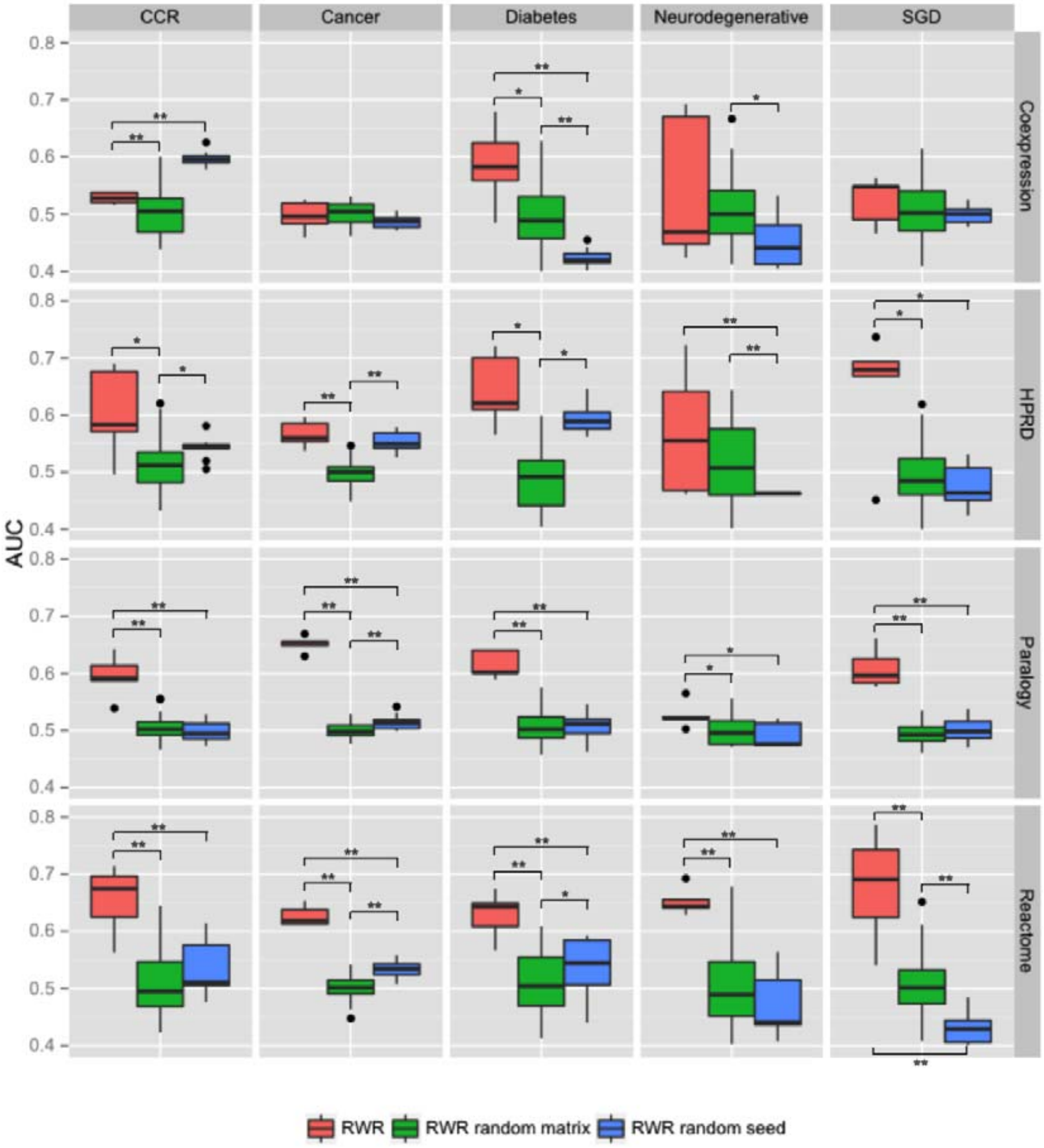


Figure 33.- Comparison of the performance of our standard RWR method, and the same method using either random initial probability vectors or randomized networks

In our case, we used as a seed genes associated to the different diseases of our interest. However, in order to evaluate the influence of these “seed” genes, we created for each disease 10 random vectors of the same size of the original disease-related initial probability vector, and calculated the predictive power of this vector.

Surprisingly, as shown in figure 33, some of these random vectors are indeed able to predict genes associated to the disease being studied, particularly when using the HPRD network (AUCs of 0.54, 0.55 and 0.60 when predicting colorectal cancer, cancer and diabetes respectively). In the case of colorectal cancer using the coexpression network, the results obtained when using the random seed are even better that when using the disease seed.

Topology of the network

In order to assess the effect of the particular topology of each of our 4 networks we generated, for each one of them, 10 randomized networks. These networks preserve the original topology in terms of number of nodes, number of edges and degree distribution, however the specific connections between genes have been randomized. As shown in figure 33, randomizing any of the four networks results in removing all their predictive power. This is reflected by an average AUC around 0.5 in all cases.

Summary

Our results indicate that networks derived from other biological types of information than protein-protein interactions can be successfully used to predict disease-related genes. The performance of each network, however, seems to be strongly dependent on the type of disease being studied and, to a lesser extend, to the algorithm being used. Something similar happens with the different network algorithms. Our data supports the idea that the performance of the different algorithms depends strongly on the disease being studied and the network being used. Thus, simpler algorithms, like direct-neighbor counting, can perform as good as more complex difussion-based method, like RWR or DK, depending on the type of network. In this line, the evaluation of the influence of the different parameters in the RWR highlighted that, at least for this method, the predictive power seems to come from both, the seed being used to train the method and the network topology.

4.4. Performance of the combined networks

We next assessed whether the combination of networks performed better than the networks alone when trying to predict genes related to disorders. We compared 4 different ways to integrate the networks: juxtaposition, addition, weighted addition and Bayesian inference.

In order to simplify the layout of the plots, in the following figures the initial letter of the network substitutes full network names: “HPRD” is “H”, “Paralogy” is represented as “P”, “Reactome” as “R” and the coexpression network as “C”. For example, network noted as “H-P” is the combination of “HPRD” and “Paralogy” networks. The p-values will not be represented, as it would further complicate the representation of the data.

Figure 34 shows the AUC values obtained in each disease using each combination of networks and combining them by either addition or juxtaposition. The main conceptual difference between the two is that, in the case of juxtaposition all edges have the same weight, whereas in simple addition, edges have a weight proportional to the number of networks where they appear (i.e., if they appear in a single network, they have a weight of 1, but if they appear in 2 networks they have a weight of 2).

While one can observe differences in the performance of different network combinations, those will be analyzed in the following figure. The important point of figure 34 is that there are no differences when combining the networks using the simple addition or the juxtaposition approaches. This is probably due to the fact that there are very few edges shared between networks; thus, these are likely to have very small effect to the overall prediction. In order to simplify further analyses, since there is no significant difference between addition and juxtaposition, only results for the addition method are shown.

We then compared the performance of the networks combined using the simple addition method, with that of the networks alone (figure 35). For most diseases we found at least one combination that outperformed the majority of the single networks.

In the case of colorectal cancer, combining the coexpression network (C) with any other network generated a metanetwork that was better than the coexpression network. For example, while the coexpression network has an AUC of 0.53, the metanetworks H-C, P-C and R-C have AUCs of 0.56, 0.62 and 0.57 respectively. Something similar happens with the paralogy network. This network has an AUC of 0.59 when predicting colorectal cancer genes, whereas the combination networks H-P, P-R, and P-C have AUCs of 0.64, 0.62 and 0.62 respectively. While no network combination outperformed the Reactome network alone (AUC of 0.65), the network derived from the addition of HPRD, Paralogy and Reactome, H-P-R, performed as good as it while increasing their coverage significantly.

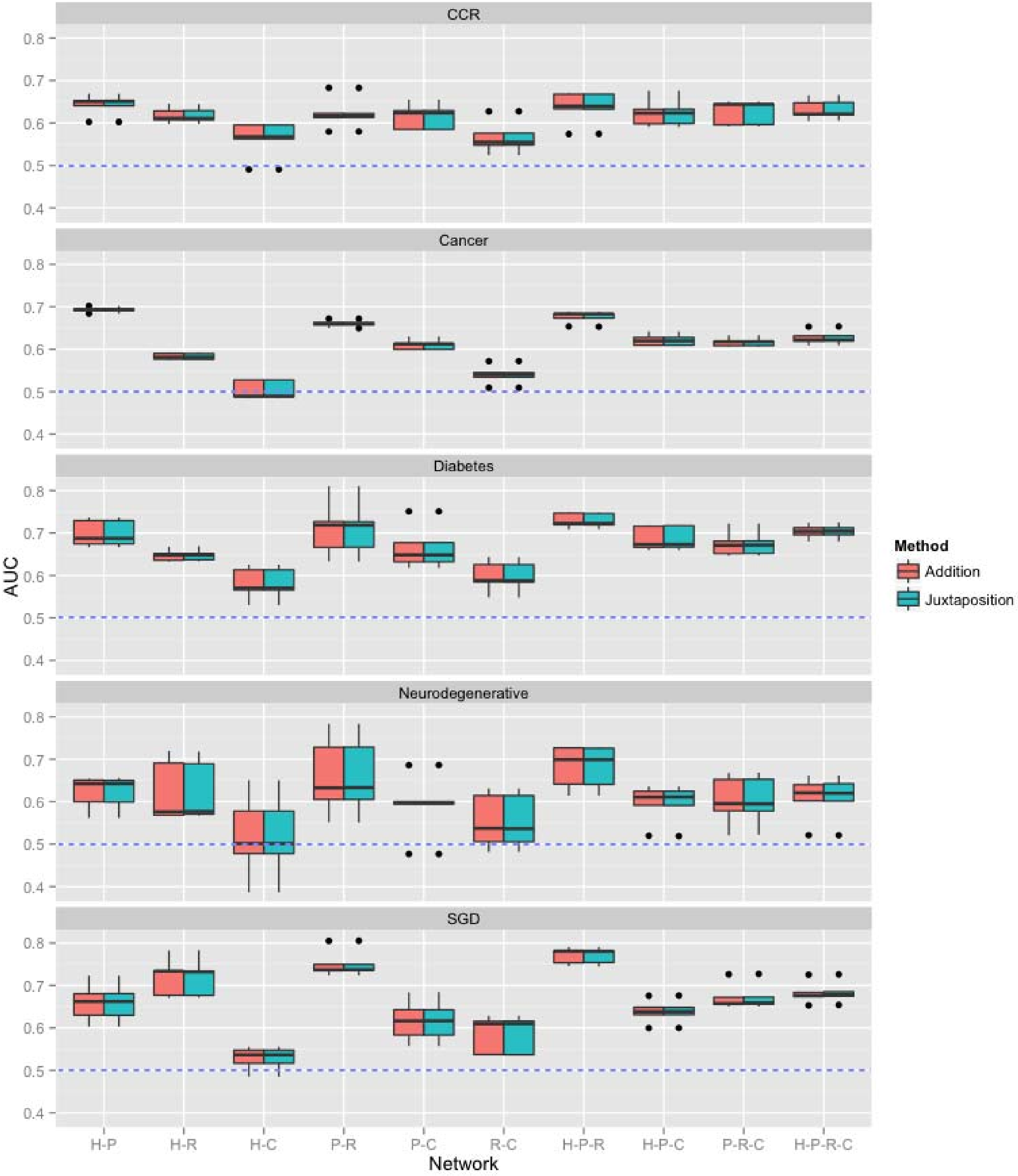


Figure 34.- Performance of the predictions obtained in each network combination using the simple addition and the juxtaposition algorithms. There is no difference between them, probably because there is little overlap between the edges of each network. The dashed line marks an AUC level of 0.5

In the case of cancer, there are similar trends. For example, any combination of the coexpression network with another one again performs better than the coexpression network alone. We have been able to find combinations of networks that outperform all the single networks, but the one from paralogy data (0.65). For example, network P-R (0.66) is better than the Reactome network alone (0.63), or networks H-P (though it has an AUC of 0.69, the difference is not statistically significant) or H-P-R (0.68, again not statistically significant compared to the Paralogy network) are better than the HPRD network alone (0.57).

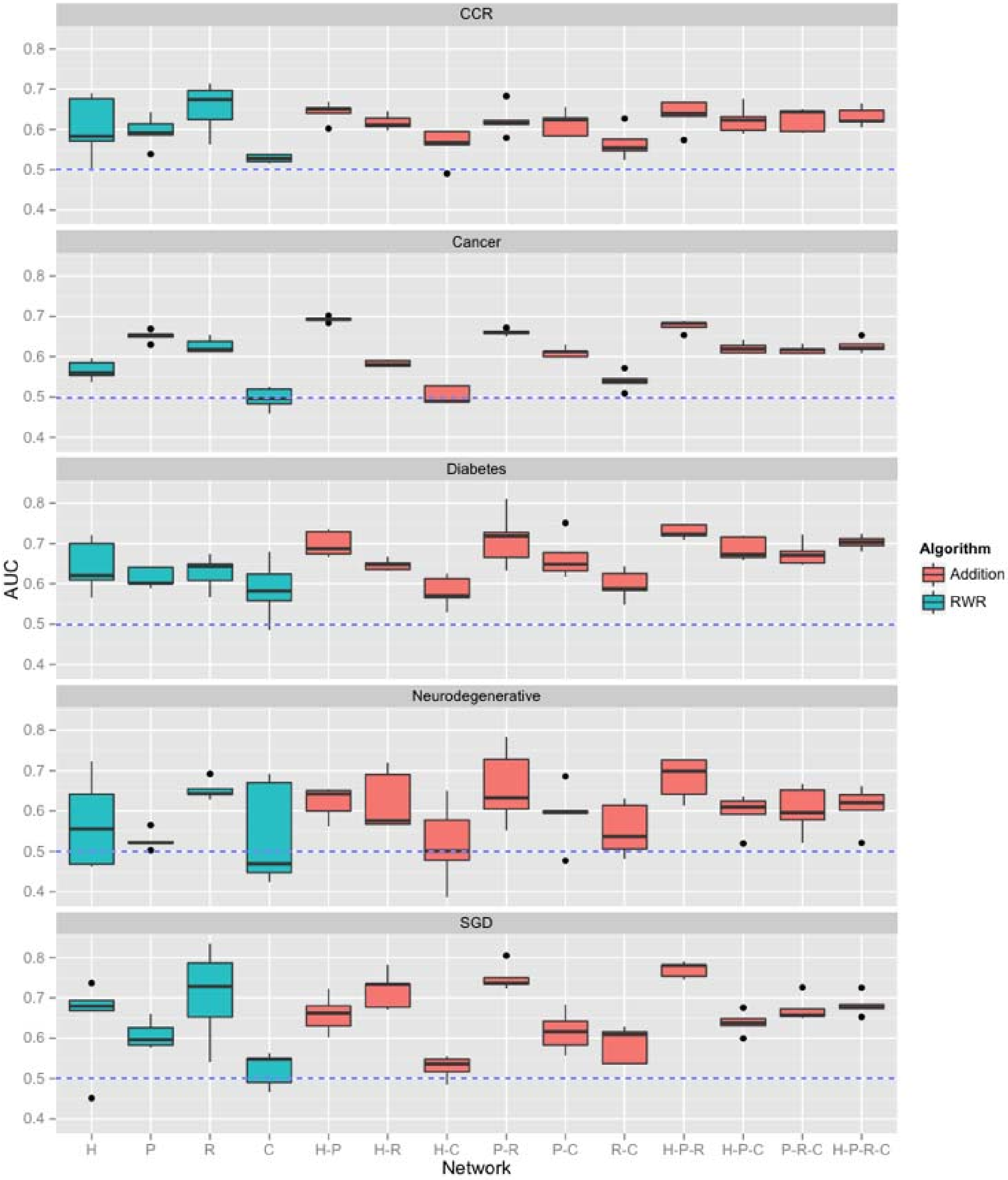


Figure 35.- Performance of the RWR in every disease using both, the networks alone (light blue) and those resulting from their addition (light red). Results with combined networks usually yielded similar AUC than networks alone, but are better in terms of coverage. Dashed light purple line marks the AUC level of 0.5

Diabetes is the disease that seems to benefit the most from the combination of networks, since it is only disease where we could find a combination of networks that outperforms all the networks alone with statistically significant differences H-P-R (AUC of 0.73, compared to 0.70 when using the HPRD network, pvalue 0.03). In the case of neurodegenerative disorders and simple genetic diseases, we could not find any combination with a better AUC than the Reactome network alone (AUC values of 0.65 and 0.71 in neurodegenerative disorders and simple genetic diseases respectively), but again, we could find some that performed as good as this network but with significantly larger coverage of the genome, such as H-P-R, which has AUC values of 0.68 and 0.77 respectively.

Overall, it seems that when combining two or more networks it is very difficult to increase the predictive power above that of the best network alone. What is possible though, is to extend the coverage of that network using other sources of information without significantly affecting its predictive power.

Interestingly, the addition of all 4 networks, labeled as H-P-R-C in figure 35, was only once among the best networks, in the case of CCR (AUC of 0.63). Thus, in most diseases, adding new information (edges) to the network doesn't imply an increase in the capability to predict disease-related genes, at least using the simple addition method. In order to check whether this was dependent on how we integrated the networks we compared the results obtained using the simple addition algorithm with those obtained by the Bayesian Inference and the weighted addition optimized by simulated annealing (figure 36).

In all the diseases the results derived from the Bayesian Inference where worse than those obtained by both, simple and weighted addition of the networks, with the only exception of neurodegenerative disorders. However, in this last case, the same tendency is observed, though the differences are not statistically significant. Moreover, the weighted addition of the networks was better than their simple addition in all diseases but colorectal cancer and neurodegenerative disorders. Moreover, the simulated annealing is able to increase the predictive power of the H-P-R-C network in 3 of the 5 diseases: Cancer (from 0.63 to 0.68), diabetes (from 0.70 to 0.73) and simple genetic diseases (from 0.68 to 0.74).

Another remarkable observation is that the weights obtained for each network by simulated annealing (figure 37) usually correlate with the performance of the network in the same disease. For example, in the case of cancer, the two most important networks are "Physical interactions" and "Paralogy", whereas in the case of neurodegenerative disorders they are the "Pathway" and "Physical interactions" networks. It is important to notice that there are some exceptions to this tendency. For example, in the case of cancer, the Pathway network was better than the Physical interaction network, however the weight of the latter is 4 times the one of the former.

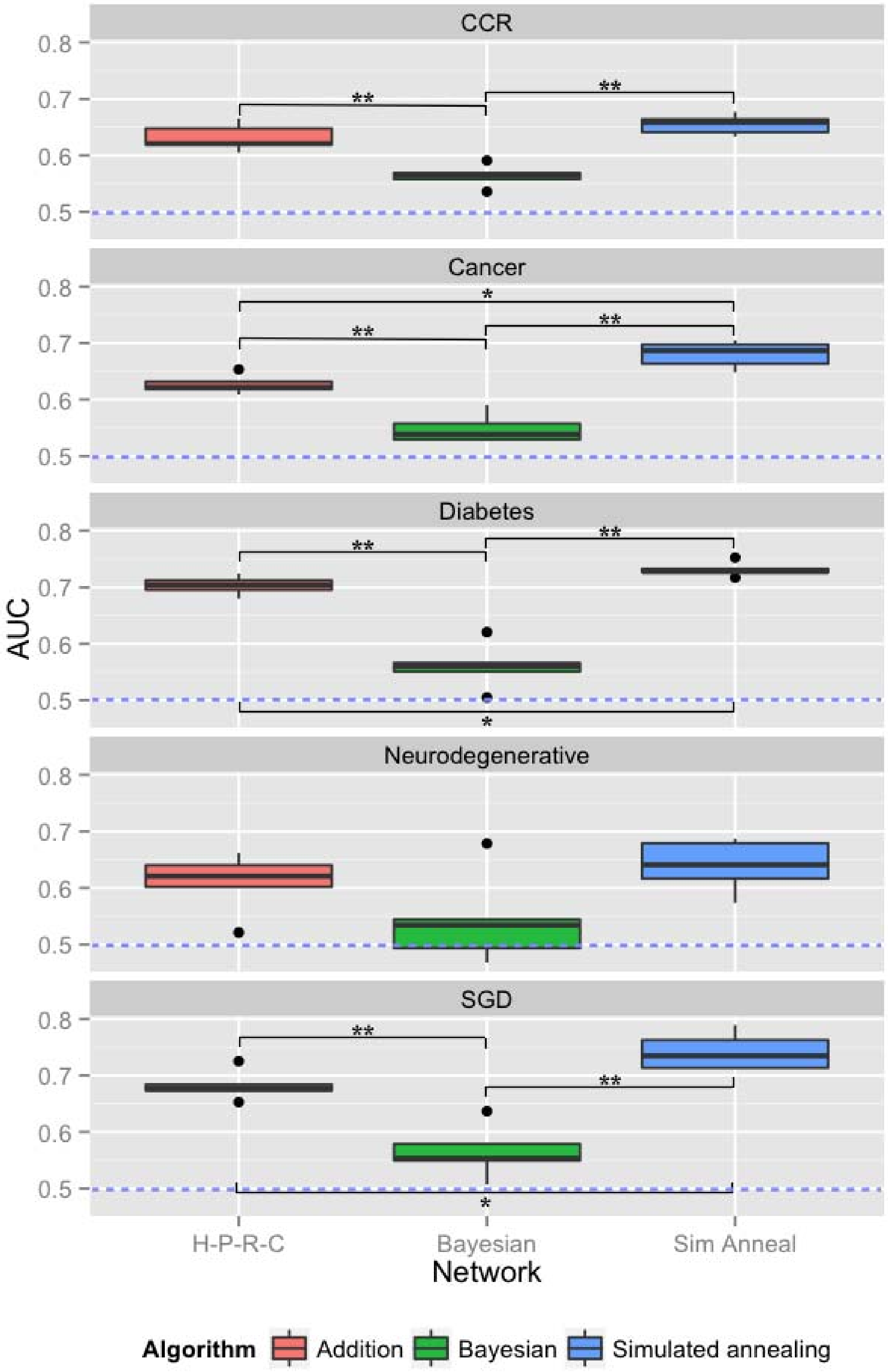


Figure 36.- Performance of the different algorithms used to combine the 4 networks when predicting genes associated to each disease. Bayesian Inference was the worst performer in 4 of the 5 diseases. Pairs marked with one or two asterisks have P values below 0.05 and 0.01 respectively. Dashed purple line marks the AUC level of 0.5

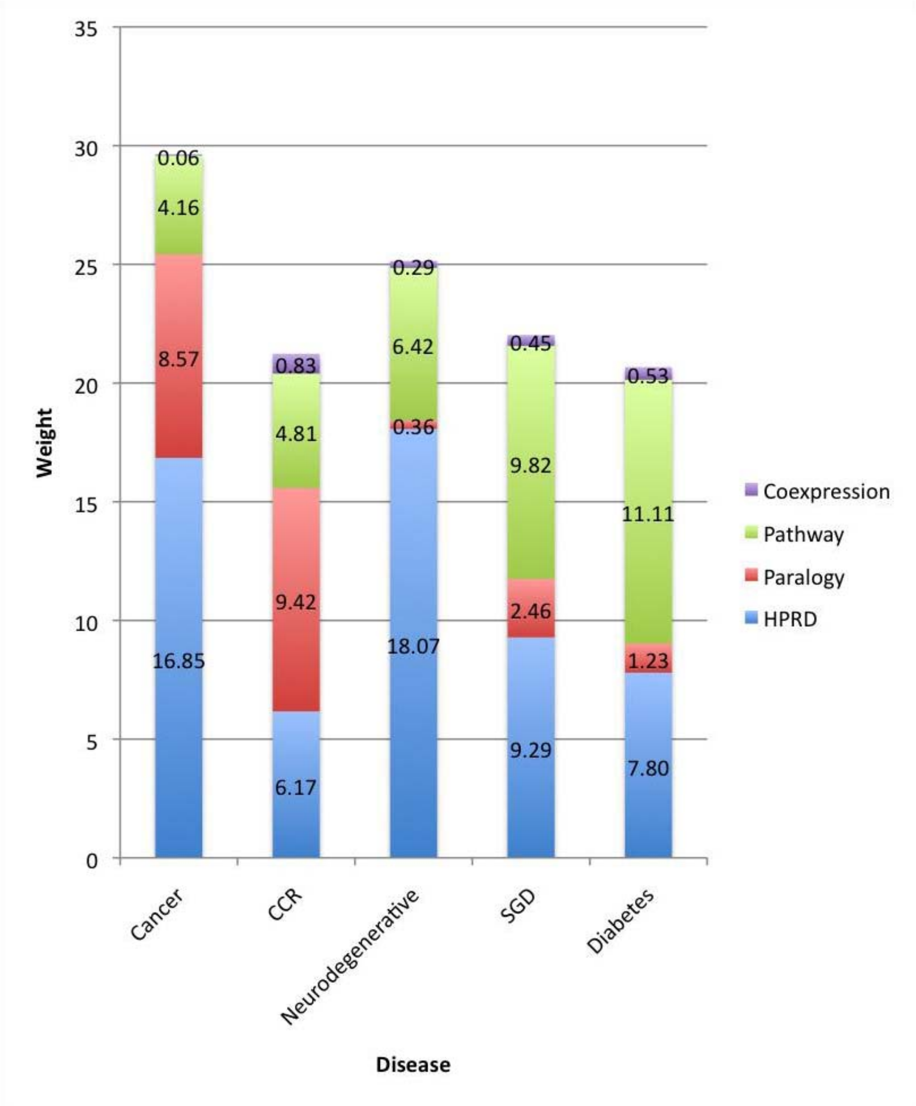


Figure 37.- Optimized network weights for each disease. Weights were obtained by simulated annealing

Recently, some groups have used another approach to integrate biological information in networks. It consists in using a Bayesian Classifier trained with several sources of information and a set of true positive and true negative interactions to score all the possible interactions of a protein network. The classifier uses all the biological evidence to score the interactions, which gives rise to a weighted functional network.

We downloaded one such functional network constructed by Marcotte *et. al.*[142]. It has been constructed using 21 different types of biological information. In order to evaluate this network we launched a RWR with the same sets of disease genes and compared its performance with that of our network-combination algorithms.

The functional network ranked in all cases among the top networks to predict disease genes (figure 38). However, the network resulting from simple addition of the 4 networks, H-P-R-C, performed as good as the functional network by Marcotte in all the diseases. Moreover, in the case of cancer, the weighted combination of the 4 networks obtained by simulated annealing was actually better with a statistically significant difference (AUCs of 0.68 and 0.63 for the simulated annealing and functional network respectively, with a p-val of 0.015). Also in the case of cancer, a combination of only three of our networks, H-P-R, outperformed the functional network with AUCs of 0.68 and 0.63 respectively (p-val < 0.01).

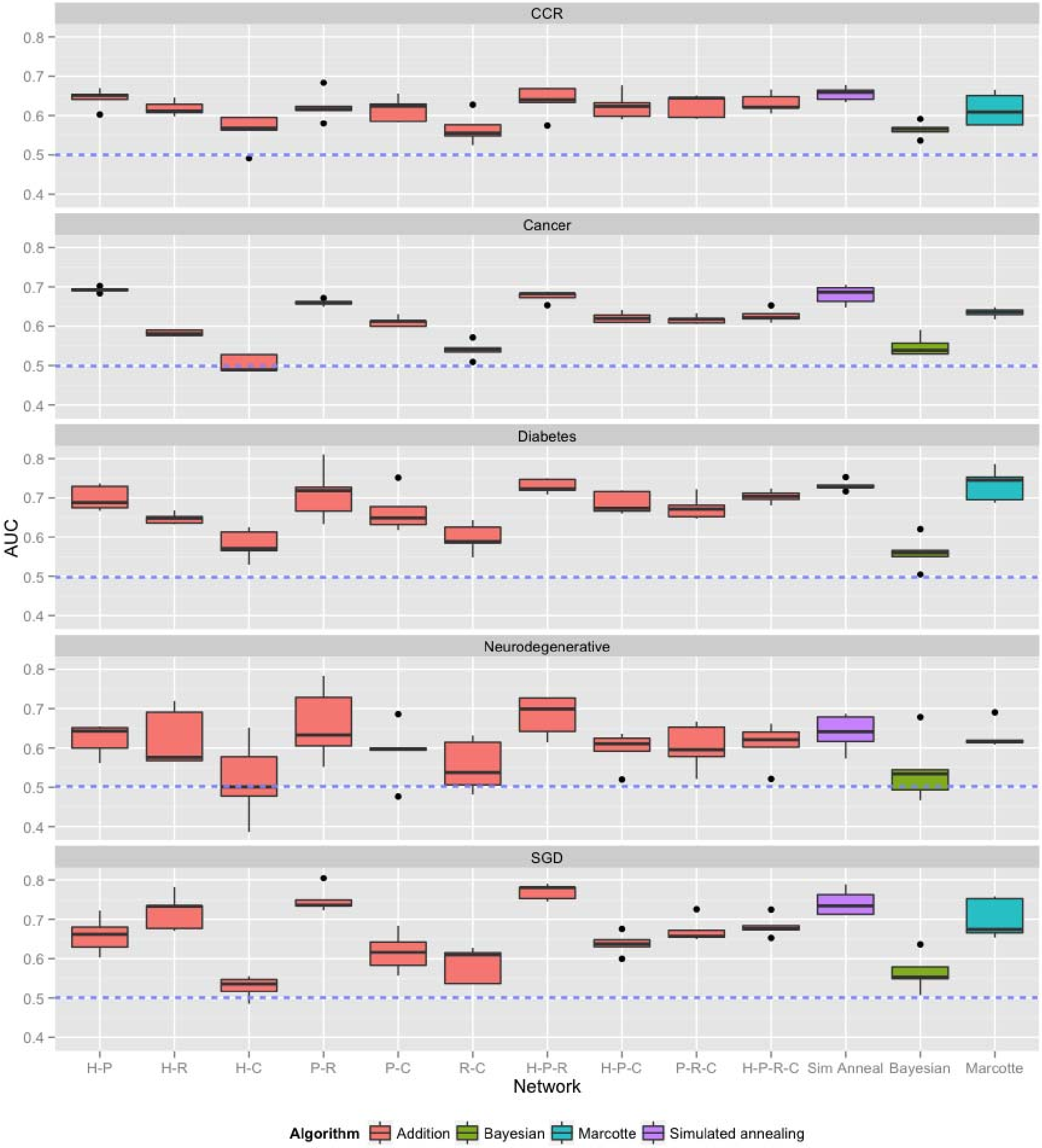


Figure 38.- AUCs obtained when predicting genes associated to the different diseases with each method that combines 2 or more networks. The dashed light purple line marks an AUC level of 0.5

4.5. A case study: prediction of driver genes in colorectal cancer

We then studied the performance of our method when predicting driver genes in colorectal cancer. In order to do so, we decided to use the network derived from the addition of HPRD, Paralogy and Reactome networks, H-P-R. We chose this network instead of the one optimized by simulating annealing because the calculations are faster, because the transition matrix is calculated only once, and the performance in both, cancer and CCR, are quite similar (p-values of 0.54 and 0.69 in colorectal cancer and cancer respectively). Moreover, the weight of the coexpression network obtained by simulated annealing is very close to 0, which means that the metanetwork is using very little information from this network to make the predictions.

We defined our list of true driver genes as all those that have been found mutated in 15 or more CCR samples according to COSMIC, yielding a total of 482 genes. Of these, we excluded those already associated to either CCR or cancer in general according to our database. After this filtering step, we ended with a total of 353 genes not associated to CCR, according to our database, and 252 new cancer associated genes not originally included in our database. Notice that the numbers are different because, in the case of CCR genes we are only excluding from the driver list genes associated to CCR in our database. In the case of cancer, we are excluding all the driver genes that are associated to *any* type of cancer, thus the number is lower.

Standard RWR

When we measure our ability to predict this new driver genes, the results obtained using the H-P-R network and the seed with all the cancer genes in our database, or only those associated to CCR, were very similar in terms of AUC (AUCs of 0,66 and 0,65 respectively). They were also quite similar when compared to the average AUC calculated in the crossvalidation experiment (0,64).

In a more realistic set up, we ranked all the new genes compared with the 9 and 24 closest genes in the genome (figure 39a and 39b respectively). There is a clear enrichment in both cases towards the first positions of the ranking. Interestingly, when we remove all the driver genes that have a score of 0 (which are 2 in the case of cancer and 5 in the case of CCR), this enrichment is even more clear (figures 39c and 39d).

Boosted RWR

Using initial probability vectors adjusted by mutation frequencies did not yield significantly better results compared to using our standard vectors. In terms of AUC, using genes only associated to CCR we obtained a value of 0,66 which is very similar to the preivous value of 0,65. When using all cancer genes, the AUC obtained was 0,65 that is even a little bit lower than the one obtained before (0,66).

The rank analysis, shown in figure 40 also led to similar observations than before: there is a strong enrichment of driver genes towards the first positions of the ranking using either CCR or all-cancer genes and removing genes with 0 score (which are 41 in the case of cancer and 52 in the case of CCR) led to an even stronger enrichment. However, including this type of information, at least in the initial seed, does not seem to improve the overall performance of the method.

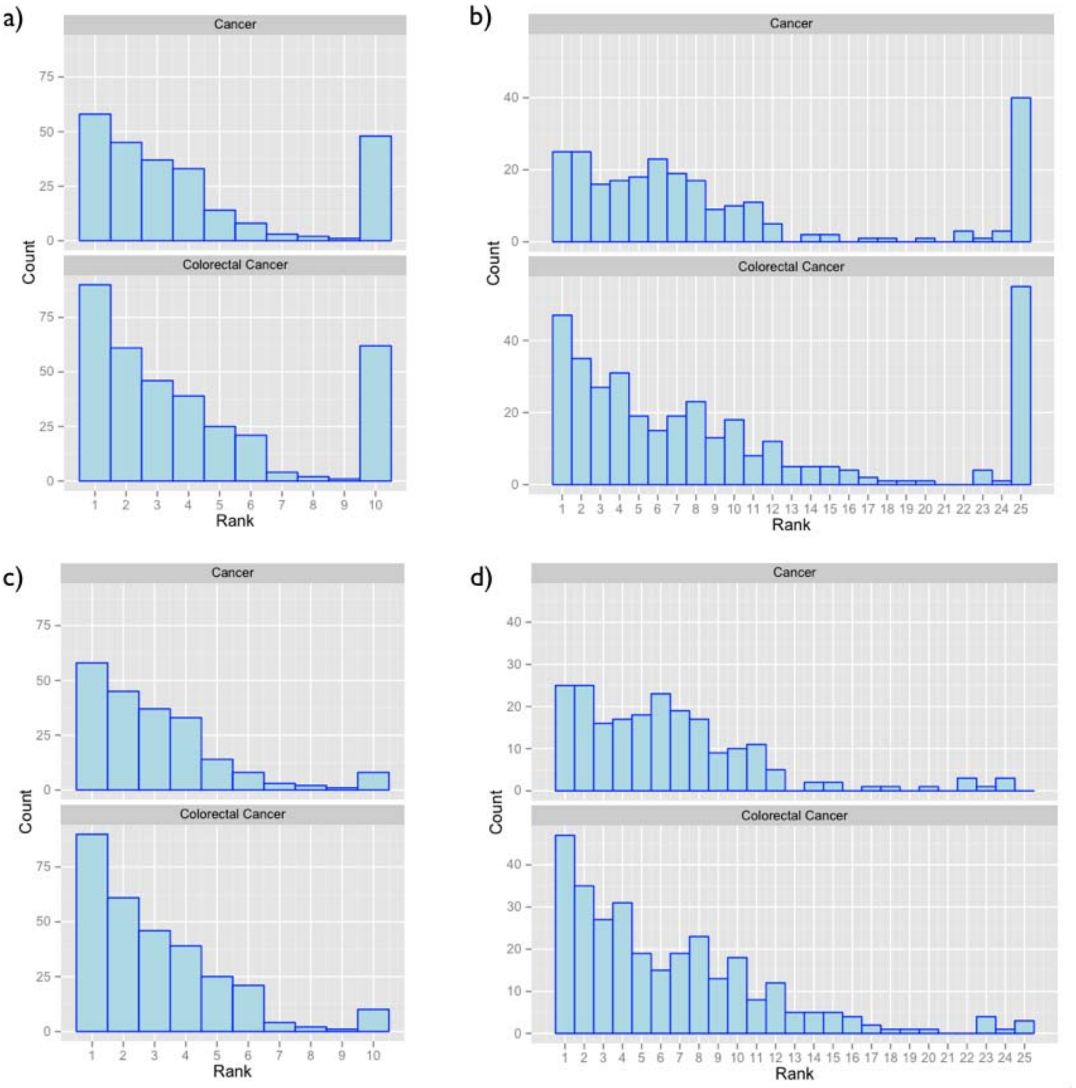


Figure 39.- Driver CCR genes are consistently ranked in the top positions of a group including the 9 (a,c) or 24 (b,d) nearest genes in the genome. This enrichment is consistent when using both, all cancer genes (top graphs) or only CCR-related genes (bottom graphs) and is even more evident when removing driver genes with 0 score (c,d)

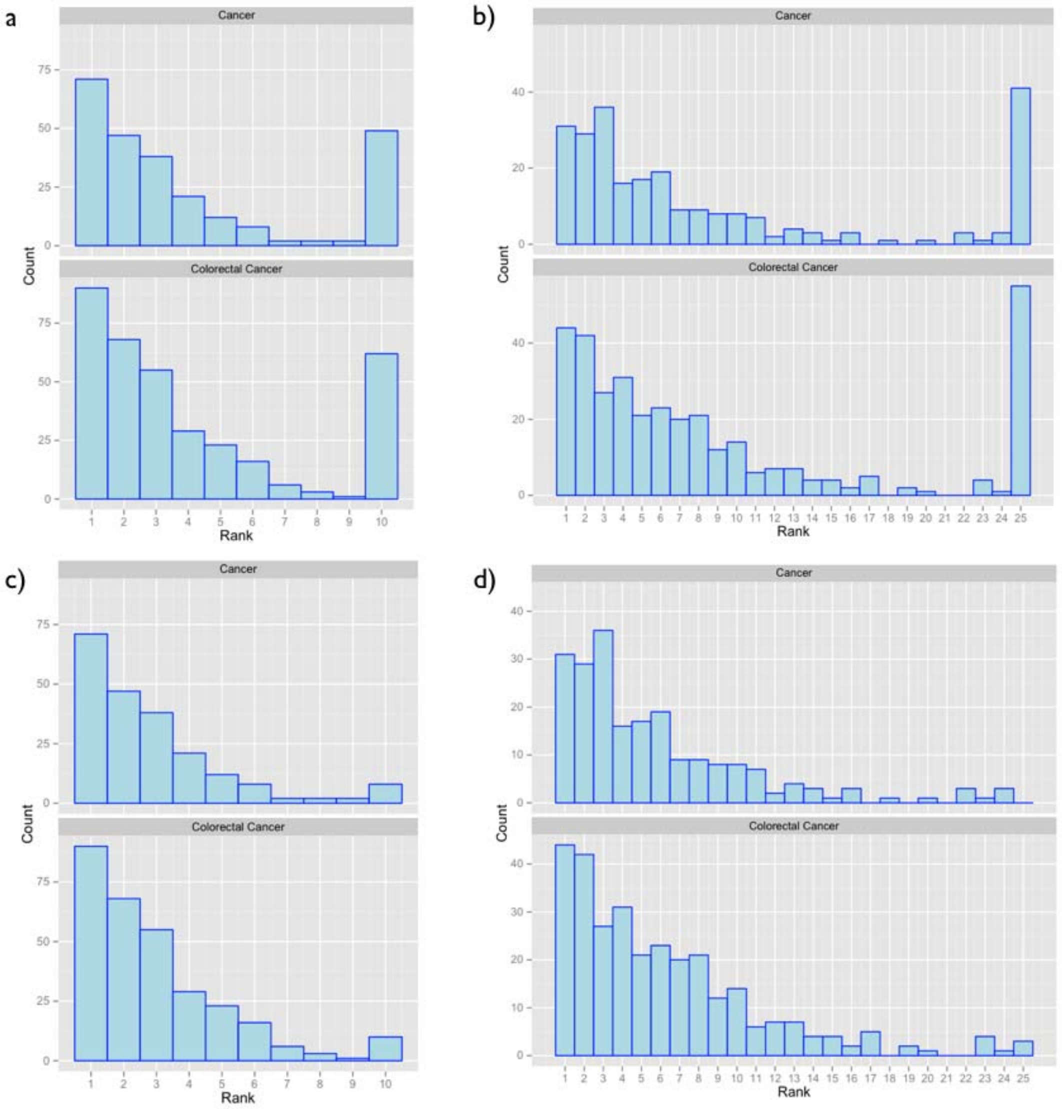


Figure 40.- Results of the rank enrichment experiment when using the RWR and initial probability vectors adjusted to mutation frequency. Again, driver CCR genes are consistently ranked in the top positions of a group including the 9 (a,c) or 24 (b,d) nearest genes in the genome. This enrichment is consistent when using both, all cancer genes (top graphs) or only CCR-related genes (bottom graphs) and is even more evident when removing driver genes with 0 score (c,d)

Summary

According to our data, it seems that, while it is possible to generate metanetworks of similar predictive power than the best network alone, however, it is difficult to go beyond that, since only in one disease, diabetes, we have been able to identify a network combination that outperformed all the networks alone. However, by adding different networks one can achieve similar performances while extending the disease coverage. Moreover, it is better to combine the information from different networks by directly adding the networks than combining the scores, at least when using a Bayesian classifier. Finally, there is not a direct correlation between the amount of information used to generate a metanetwork and its performance.

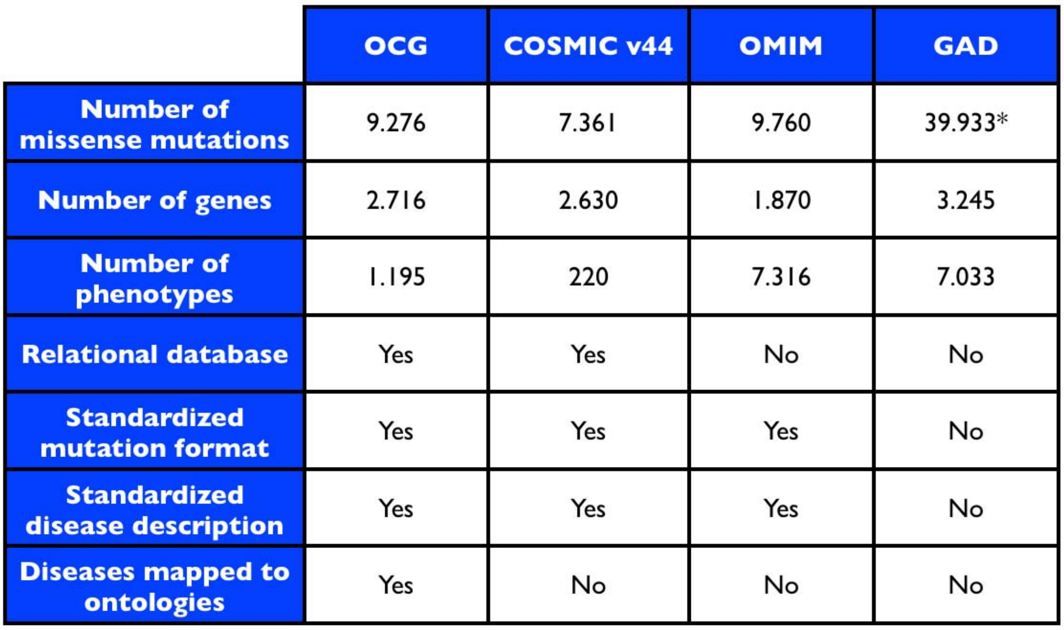
5. Discussion

With the increased use of genome-wide technologies that identify genomic regions and mutations susceptible to be associated to disease, it is imperative to develop bioinformatic tools able to deal with this information in order to (I) properly understand the biology of the diseases and (II) pinpoint those genes truly associated to the phenotype of interest from those that are not. In this thesis we explored two of the most used computational approaches in this direction: the use of enrichment analysis to identify biological features specifically mutated in certain diseases and the use of biological networks and graph theory algorithms to identify disease-associated genes.

5.1. Development of a database of disease-associated genes and mutations

To identify disease-specific mutated features using the enrichment analysis, we needed a proper framework that could not be provided by existing databases about disease-associated mutations. For example, while there are several resources and databases about different biomedical aspects of diseases[181,182], our analysis required a resource focused on disease-related mutations and protein features, where all the annotations are normalized using controlled vocabularies or ontologies. This reduces the number of available public resources and none of them fulfilled all our needs, so in order to overcome some of these limitations, we have developed a new database named OCG (table 16). This database is stored as a relational database instead of a simple text file, all its annotations have been normalized using controlled vocabularies or ontologies and all the mutations have been verified in their corresponding ENSEMBL proteins. Thus, it provides an adequate framework to retrieve the information needed to identify mutated features associated to certain diseases.

Table 16.- Summary of the properties of different databases



	OCG	COSMIC v44	OMIM	GAD
Number of missense mutations	9.276	7.361	9.760	39.933*
Number of genes	2.716	2.630	1.870	3.245
Number of phenotypes	1.195	220	7.316	7.033
Relational database	Yes	Yes	No	No
Standardized mutation format	Yes	Yes	Yes	No
Standardized disease description	Yes	Yes	Yes	No
Diseases mapped to ontologies	Yes	No	No	No

*The number of mutations in GAD is difficult to estimate, thus the number of entries is reflected in the table

The design of the OCG database has been inspired in that of COSMIC[183] in its 44th version. Its mutation coverage is wide as it contains 9.276 mutations. While this number may seem very low when compared with databases such as COSMIC, which contains over 200.000 mutations in its version 61, its quite similar to that of COSMIC itself in the version we used (7.361 mutations) or OMIM[6] (9.760 mutations). This last comparison is particularly relevant, since OMIM contains curated mutations that are very likely to be causal whereas databases with much larger numbers of mutations, like COSMIC v61, are more likely to include passenger mutations.

In terms of diseases our database has 1195 different DO terms, which is an intermediate value between the 220 types of cancer, present in COSMIC and the 7316 distinct phenotypes in OMIM. It is important to take into account that not all these 7316 phenotypes from OMIM are diseases. While non-disease phenotypes described in OMIM, such as gene expression or drug sensitivity, might be important for other studies, are not relevant for our study. Moreover, as we have shown in chapter 4 (figures 17, 18 and 19), the distribution of the mutations, genes and diseases is consistent with previous reports[171], following a logarithmic scale with most genes or diseases having low numbers of mutations and few genes or disorders associated with many mutations. Regarding the mapping of the phenotype descriptions to DO terms, we have achieved an accuracy and recall of 85% and >33% respectively using OMIM (some of the mutations for which we had identified a DO term could not be mapped to ENSEMBL proteins, so they were discarded). These are similar parameters to those obtained by other authors trying to map disease descriptions from OMIM to either the DO[171] or MeSH[184], another biomedical ontology.

Limitations of our dataset

While globally speaking the database schema and content are appropriate to pursue the subsequent objectives, our dataset has some limitations that have to be taken into account when analyzing the results. First of all, we have not been able to extract all the information from the original databases. For example, though there are 9760 missense mutations in the version of OMIM that we used, due to problems mapping the phenotype terms to the right DO term (either because the name is different in OMIM and DO or simply because the exact word for that phenotype does not exist in the DO), we could only keep 3222 (33%) of them. Another problem that we found is that, sometimes, OMIM and ENSEMBL coordinates do not match exactly, so the mutation cannot be mapped to the ENSEMBL protein and it is not stored in the database.

Another potential limitation of our study arises from the imbalance between the two main groups of mutations (those from COSMIC and those from OMIM/GAD). We have roughly 45% more mutations from COSMIC than from OMIM and GAD (6056 and 3307 respectively). However, according to our results, the set of COSMIC and OMIM/GAD mutations have little overlap (chapter 4, figure 16). Moreover there is a low number of mutations associated to complex diseases, which are mainly those coming from GAD. It would be interesting to have a higher number of complex mutations in our analysis in order to have a more complete dataset and extend the scope of our analysis, however, the lack of standardization of GAD makes this very complicated to automatize. As a summary, we think that the dataset is diverse enough to make a first approach in that direction and, at least, analyze the differences between mutations associated to cancer and those related to simple genetic disorders.

5.2. Disease-specific mutated features

Enrichment analyses have been widely used in biomedical studies, as they are particularly well suited to analyze genome-wide data, such as that derived from microarray or ChIP experiments. However, though enrichment analyses can be performed using any type of controlled vocabulary or ontology[79] and there are over 100 ontologies in the OBO foundry[78], according to our knowledge, enrichment analyses have been extensively performed in the framework of GO terms, and few attempts to extend them to other ontologies have been performed.

We performed an enrichment analysis trying to identify pairs of terms that showed statistically significant associations using 4 different controlled vocabularies: the DO, GO, SO and PFAM. Most studies trying to identify disease-associated features have focused on comparing the properties of mutations associated to either diseases in general[81], or some more detailed phenotypes[185,72], with those of neutral polymorphisms (such as nsSNPs) or simulated random mutations[49]. Our approach is different in that we are always comparing disease-related missense mutations. By doing this, we can identify subtle differences in the pathogenic mechanisms of mutations associated to different phenotypes.

While each controlled vocabulary is meant to describe different aspects of protein biology, there are some areas in which they overlap. As would be expected, associations involving terms in those overlapping areas of knowledge are consistent along the different controlled vocabularies. For example, there are less mutations than expected associated to cancer in genes “integral to membrane” (GO:0016021), “transmembrane regions” (SO:0001077) or the “ABC transporter transmembrane region” (PF00664).

Identification of known associations

Associations found between DO and GO terms provide a first good control of the performance of the method. We found 3199 such pairs, most of which could be expected from the biology of the disease. For example, DO terms related to inherited blood disorders are associated to genes whose function is related to blood coagulation. One of the factors contributing to these associations comes from the fact that some GO annotations are inferred from phenotypes observed in individuals where the gene is mutated. Of the total 44.914 annotations of human genes with GO terms, 2725 are inferred from mutant phenotypes (evidence code “IMP”).

Another group of DO/GO associations that provided us good evidence about the performance of the method are those that have been previously described. For example, we have several GO terms that are strongly associated to the DO term “Cancer” (DOID:162) such as “Transcription”, “DNA repair” or “MAPKKK cascade”, which have been previously related to cancer[175,59,77].

Interestingly, we found also some of the statistically significant pairs between DO and GO that had been identified in one of the few enrichment analysis previously performed using these two ontologies[139]. In that work, LePendu *et. al.* showed that genes annotated with the GO term "aging" (GO:0007568) are enriched in DO terms such as "Cancer" or "Alzheimer's". In agreement with these studies, using our analysis we identified DO terms such as "Brain's disease" or "Cancer" as associated to this same GO term.

Regarding associations between disease and PFAM domains, it is important to remark that others have recently analyzed disease-related mutations in that context. For example, Zhong *et. al.* described a series of proteins that when mutated in different domains are associated to different phenotypes[109]. This observation has been recently confirmed an attributed to mutations altering interfaces of domain-domain interactions[186]. In another recent publication, Nehrt *et. al.*[49] used the PFAM domains to group somatic mutations from samples of either breast or colorectal tumors. When the mutational landscape of somatic mutations is analyzed by grouping the mutations according to the gene containing them[59,47,187] two main patterns emerge: strong peaks, which are genes systematically mutated in cancer and, thus, very likely to be drivers, and weak "hills" of genes that are mutated less frequently but that may contain some genes that contribute to cancer in specific contexts. In order to identify these context-specific drivers, they decided to group the mutations according not to the gene where they are happening, but to the domain were they are located. They observed a series of domains, each of which can happen in one or more proteins that are enriched in somatic mutations when compared to a random distribution of the same mutations. These domains include, among others, "P53", "MH2", "APC", "PI3K p85 binding domain" or "Miro" in the case of colorectal cancer and "P53", "PI3ka" or "IL8" in the case of breast cancer. Interestingly, we also found a strong enrichment in the "Miro" domain in both types of cancers and "MH2" in colorectal cancer. The Miro domain is a GTPase signaling domain, found in proteins involved in mitochondrial motility. This could indicate that mutations in this particular pathway are important for these particular diseases. Proteins containing this domain have different subfunctionalization codes that differentiate them from the rest of GTPases. Therefore, they could be used as a nice framework for further therapeutic studies. However, we did not find the other associations, probably because, as reported in the paper, these domains are mutated in one or two different proteins and we need a minimum of 3 to take the association into account.

Breast cancer and the kinase domain

The GO term "Kinase activity" (GO:0016301) has been previously associated to cancer[77,188], and mutations associated to the DO term "Cancer" and some of its children are, indeed, strongly enriched in this GO term (corrected p value 2.27 E-26). However, mutations associated to "Breast carcinoma" are *depleted* in this GO term (corrected p value 0.0001).

A similar association has been found using the PFAM domain "Protein kinase domain". This domain has been previously associated to cancer[110,54] and, indeed shows a strong enrichment in mutations associated to several types of cancer according to our data, but it also shows a strong depletion of mutations associated to breast carcinoma. One of the possible interpretations of these results is that only the mutation in a reduced set of kinases can cause cancer in this tissue (19 genes), in contrast to other related diseases, such as skin cancer, where the number of kinases that can be mutated to originate the disease seems to be larger (27 genes).

It is also established that there are differences in the distribution of the mutations along the kinase domain depending on whether the mutation is associated to cancer or not[54]. This feature has even been used to differentiate driver from passenger mutations[189]. Subdomain I of protein kinases is a preferential cancer hotspot[54] because it is implicated in the active-inactive transition of the domain, thus mutations in that subdomain may cause a permanent activation or inactivation of the protein. A similar effect has been hypothesized for mutations in subdomain VIII, since it contains the DFG flexible loop and a phosphorylation site involved in regulation. On the other side, subdomain V, which is enriched in breast cancer mutations according to our data (6 mutations in 6 genes), has not been previously associated to cancer or any other disease and is thought to play basic structural roles. While the underlying mechanisms of some of the mutations in this subdomain that are associated to breast cancer have been previously described (for example, mutation L184S in MAPKKK4 inactivates the gene, which is a tumor suppressor[190]), more work on the role of subdomain V is needed in order to understand its relationship with breast carcinoma.

DO and SO associations

Grouping mutations according to their presence in PFAM domains has proven to be successful when interpreting the properties of disease-related mutations, as aforementioned. One of the factors contributing to this success is that it provides a lower level of granularity than the GO. In this line, the use of SO to analyze disease-related mutations becomes the next straightforward step and an even provides a higher resolution level of analysis.

The SO has been previously used to annotate variations and features in the human genome[191] and is gaining attention in recent years, to the point that SO annotations are included in ENSEMBL since 2012[192]. This has led to the development of some tools to annotate personal genomes and provide simple statistics[193], retrieve known mutation consequences[194] or retrieve genome SO annotations by using DAS servers[195].

We analyzed the potential associations among the 1195 DO terms describing diseases and 88 SO terms describing sequence features by analyzing the number of shared mutations between every pair of DO and SO terms. We found 82 statistically significant associations after Bonferroni correction. Of these, 76 associations involved very general DO terms or were difficult to interpret, thus we did not perform any further analysis on those. The remaining 6 associations were selected for further analysis. All of them involved a DO term related to cancer and a sequence feature that was informative enough to make some biological hypothesis: “disulfide bonds", "transmembrane region", "protein localization signal", "compositionally bised region of peptide", "serine-rich region" and "intrinsically unfolded region". The first three SO terms were underpresented in cancer mutations, whereas the other three showed an enrichment of mutations associated to that phenotype.

We could confirm 3 of these associations using only genes that have the considered feature, which suggests that the associations are not caused by a bias in the properties of the genes. Moreover, we could also confirm 5 of these associations using a different set of cancer mutations. Considering that cancer is no longer viewed as a single treat, but a compendium of multiple diseases, our data indicate that biochemical properties are indeed informative to identify common mutated protein features involved in the development of the disease.

The influence of the unfolded protein response in cancer evolution

Mutations in disulfide bonds are often pathogenic and associated to several diseases in our database such as dysplasia (MATN C304S), Fabry disease (GLA C56G and GLA C202W), Marfan syndrome (several mutations in FBN1) or retinitis pigmentosa (CRB1 C1181R) among others. Since many cancer mutations are expected to be pathogenic by disrupting the protein's structure[81] it is perplexing to find that mutation of disulfide bonds is underrepresented in cancer mutations (12 mutations in 9 genes).

One plausible hypothesis is that the disruption of disulfide bonds may create unfolded protein products that are retained longer in the endoplasmic reticulum[196]. This retention could ultimately increase the stress in that organelle and activate the unfolded protein response (UPR)[197]. Although the UPR has been shown to protect cancer cells from apoptosis in some cases[198,199,200], over a certain threshold, its maintained activation leads to exactly the opposite and induces the apoptosis of these cells[201]. Given that cancer cells, particularly those in solid tumors, are already under endoplasmic reticulum stress mainly due to hypoxia conditions and glucose starvation[201], additional stress may be indeed disadvantageous. Therefore, unless the disrupted protein is giving a clear biological advantage, mutations increasing the overall stress in the cell (like disulphide bonds) may be under strong removal pressure. In this line, Geiler-Samerotte *et. al.*[202] have recently quantified the costs of protein misfolding in yeast and they found a positive correlation between the amount of misfolded proteins and the fitness decrease of the cell. UPR-induced apoptosis has been proposed to be involved in other diseases, such as type II diabetes. In this case islet beta-cells in the pancreas would die due to excessive accumulation of insulin in the ER that would lead to prolonged UPR activation and cell apoptosis[203].

In a similar line, there are several ways for a mutation in a transmembrane region to be pathogenic such as impeding interactions that occur through transmembrane domains[204,205] or affecting the correct conformational settings for the regulation of the receptor[206]. Mutations may also affect the proper folding of the transmembrane region of the protein impeding the insertion of the polypeptide in the membrane[207]. Supporting this ideas, it has been described that pathogenic missense mutations in transmembrane domains tend to occur in polar residues which are likely involved in salt-bridges[208], thus disrupting important interactions for protein stability. In this last case, the misfolded protein will remain in the ER increasing the organelle's stress and ultimately leading to UPR activation[209], which agrees with the hypothesis explained above for disulfide bonds and could help in understanding why mutations in these regions are underrepresented in cancer mutations (117 mutations in 60 genes). Finally, mutations in peptide localization signals, which are also underrepresented in cancer-associated mutations (24 mutations in 22 genes) are likely to disrupt the correct protein trafficking and cause the protein product to remain stalled in the ER[209] increase ER stress and activate the UPR, ultimately also having a pro-apoptotic effect in cancer cells.

Taking all into account, the three under-represented features in cancer mutations seem to affect tumor evolution. Cancer cells are under ER stress and mutations in any of those features are likely to be under strong selective pressure. Therefore, mutations in these regions may be unsustainable for the cell's viability and most likely purified during the evolution of the tumor population.

Notably, ER stress and UPR have been proposed to be a key mechanism of several drugs in preclinical and even clinical settings for cancer treatment[201,210]. It has been recently shown that Bortezomib (also known as Velcade), a cancer-approved drug that inhibits the proteasome, acts, at least partially, by inducing ER stress in multiple myelomas[211] or human pancreatic cancer cells[212]. Increase of ER stress through inhibition of secretion and vesicle trafficking by chemical agents is also being tested as an anti-cancer therapy in chronic lymphocytic leukemia[213].

Intrinsically unfolded regions and cancer

We also found 3 sequence features overrepresented in cancer mutations: "compositionally biased regions" (216 mutations in 72 genes), "intrinsically unstructured regions" -IURs- (1753 mutations in 928 genes) and "Ser-rich regions" (84 mutations in 12 genes), though we could not confirm the latter association in the set of mutations from the TCGA.

Considering that compositionally biased regions and serine-rich regions are highly unstructured[177,214] and are strongly related (section 4.2), we analyze all the three features together to avoid potential biases caused by this effect and to simplify the analyses.

IURs are regions of proteins that do not need a regular fold in order to be biologically functional[215,216]. These are highly flexible and dynamic regions involved in several cellular functions, particularly those that require protein-protein or protein-DNA interactions, such as signal transduction or transcription regulation[217]. IURs have a higher flexibility and a larger interaction surface than regularly folded regions that allow them to interact with high specificity and low affinity and with multiple partners, which explains their preferential positions in hubs of the interactome[218,219]. Moreover, proteins with high content of IUR are tightly regulated[220,221], particularly by post-translational modifications such as phosphorylations[178] or ubiquitinations[220].

Different types of alterations, such as copy number variations[61] or chromosomal rearrangements[222], involving proteins containing IURs have been previously associated to different diseases[223,224], particularly neurodegenerative disorders[225,226] and cancer[223,62]. In our dataset missense mutations in IURs are associated, among other diseases, to long QT syndrome (mutation S1103Y in gene SCN5A), congenital heart defects (N21H in gene CFC21), frontotemporal dementia (G272V in gene MAPT) or several types of cancer such as colorectal cancer (A189V in TP53).

In a recent paper[227] Vacic *et. al.* propose that missense mutations in IURs associated to disease should be analyzed differently than mutations in folded regions. They argue that since the properties that confer the biological function of the folded and intrinsically unfolded regions are different, the pathogenic mechanisms of the missense mutations in these regions should also be different. For example, since IURs tend to be less conserved[228], one of the most popular programs to predict whether a mutation will be pathogenic, SIFT[38], since it is precisely based on aminoacid conservation through evolution, tends to incorrectly predict the outcome of mutations in IUR[81]. This is an important point to consider when predicting driver mutations by using conservation-based derived scores. If one tries to identify driver cancer mutations by using scores derived from conservation, these mutations are probably going to be misinterpreted.

Missense mutations in IURs are expected to be pathogenic by different mechanisms, among others disrupting protein-protein or protein-DNA interactions, that could lead to "edgetic perturbations"[109]. In this line, alterations in IURs have been recently associated to rewiring interactomes through evolution[229]. Another possible pathogenic mechanism is that these mutations alter post-translational modification sites involved in the regulation of the protein, which could cause accumulation of dosage sensitive proteins[61]. These results agree with our observation that mutations occurring in post-translational modification sites are also located within intrinsically unstructured regions more than expected (section 4.2). For example, mutations Y591C in gene FLT3 (associated to lymphoma in COSMIC) and Y336F in PTEN (associated to brain cancer in COSMIC), are located at intrinsically unstructured regions and both mutations disrupt phosphorylation sites that have been described to be important for the proper regulation of the proteins' activity[230] and degradation rate[231] respectively. These pathogenic effects could also be caused indirectly if a missense mutation induces disorder/order transition of the region (D -> O)[232,227] that would impede the unfolded state of the region and block the accessibility of a post-translational modification site or disrupting an interaction surface. This would be the case for mutation R306C in gene MeCP2 that is associated to Rett syndrome in OMIM and was predicted to induce an order to disorder transition inside an IUR[227].

Discriminative associations

To the best of our knowledge, few analyses capable of discriminating between disease subtypes have been performed using controlled vocabularies. One of the few examples would be IntOGen[233], which is focused on cancer and uses the International Classification for Oncology. IntOGen is an excellent framework to analyze cancer genomic data and it includes several types of data, besides mutations, that we have not used in our analysis, such as expression from thousands of microarray experiments. However it does not include information on other disease types. The interesting thing of our method, besides those novel associations that help us highlight important aspects of disease biology (such as the role of UPR and IURs in cancer) is that it provides associations that show different trends in distinct phenotypes, even between DO terms that are closely related such as subtypes of cancers. It is in these cases that the design of our experiment becomes the most relevant, because these types of associations would be difficult to identify when comparing disease-related mutations to random sets of mutations or benign variations.

We have found several examples of GO terms, particularly in cancer, that show an association in one direction with a given DO term and another association in the other direction with another closely related DO term. For example, "Metabolic process" is enriched in mutations associated to "Lung carcinoma" and depleted in mutations associated to "Pancreatic neoplasm" and "Breast carcinoma". Another interesting example is "Angiogenesis", enriched in mutations related to "Cancer" and "Kidney neoplasm" and depleted in mutations associated to "Breast carcinoma" and "Lung carcinoma".

While we observed a total of 89 GO terms that follow this trend, this idea is better represented in figure 20. This figure shows a clustered heatmap representing whether a given DO/PFAM pair shows an enrichment or depletion of mutations. The distinct subtypes of cancer cluster together forming two distinct subgroups and DO terms describing other diseases cluster on the other side of the heatmap. In this figure it is possible not only to identify domains strongly enriched in cancer mutations, such as Miro, Ras or Ankyrin, but also to appreciate a series of PFAM domains, such as SNF2, Sushi FN3, or Protein Kinase that show a depletion of mutations in several types of cancers and an enrichment in other subtypes of tumors.

One possible use of all these associations that we have not explored in this thesis is the training of a classifier able to identify/prioritize pathogenic mutations. Similar approaches using features associated to disease to identify driver/pathogenic mutations have already been described[234]. For example, Nehrt *et. al.* have used domain enrichment analysis to identify PFAM domains enriched in driver mutations[49], but their background model, like that of Yue *et. al.*[235], were estimations of mutation rates instead of disease-associated mutations. In another recent publication Peterson *et. al.* compared the distribution of mutations associated to mendelian disorders along the different PFAM domains with that of cancer-related mutations[236]. While their framework is more similar to the one that we have used, their design is unable to identify differences between cancer subtypes, as all the mutations have been mapped directly to cancer. Protein annotations with GO terms have also been used to identify pathogenic nsSNPs either alone[237] or in combination with other protein features[169].

Summary: Disease-specific mutated features

Our findings indicate that mutations in the same gene might be causing different phenotypes, not only because affect different functions or interactions of the same protein[109], but also because they occur at regions with different biochemical properties. In conclusion our results provide evidence that using ontologies to identify non-obvious associations between mutated biochemical features and the diseases they are associated to is a valuable tool to this purpose. We have exemplified this approach by comparing cancer-associated mutations versus mutations associated to other diseases and provided examples of significant enrichment and depletions. The associations identified in this work are pointing to underestimated aspects of the biology of cancer and may help us in the future to discriminate between passenger and driver mutations.

5.3. Prediction of disease-related genes

Besides enrichment analysis, another class of methods that have been gaining attention in recent years to prioritize lists of putative disease-related genes are those base in biological networks. These have their origin after the description of the first interactomes in some model organisms[97,114,238], when scientists started using algorithms and concepts from network theory to systematically analyze the properties of such networks[239,91,94]. The observation that proteins closer in the interactome tended to perform similar functions[114], led some groups to try to exploit this idea in order to predict disease-related genes by using protein-protein interaction networks. This strategy has provided some successful results, such as the identification of the role of gene KIF1A in a neurological disorder[240], and it has been exploited using several types of algorithms and networks. We have performed a systematic study to evaluate which types of networks and algorithms perform better at predicting different types of diseases.

Performance of the different methods using individual networks

As mentioned before, most existing methods base their predictions in measuring the closeness of the candidate genes to a set of genes known to be associated to the phenotype of interest. We assessed the performance of 5 such algorithms in 4 different networks when predicting genes associated to 5 different diseases. By doing this we have been able to address whether there are any networks, algorithms or diseases that show superior performance over others.

Regarding the algorithms, although previous reports[115,180] claim a better performance of the diffusion algorithms, such as the Random Walk with Restart (RWR)[179] or Diffusion Kernels (DK)[179] over simpler methods, like direct neighbor counting (DN)[119]. However, we did not observe major differences in the performance of both classes of methods. There are, however, some cases in which the diffusion methods outperform the DN algorithms. For instance, when predicting genes related to simple genetic diseases or neurodegenerative disorders using the Reactome network. This seems to be strongly dependent on the type of network used. For example, in the case of the paralogy network, due to its particular topology (it is basically formed by highly interconnected cliques), we have not found significant differences between all the algorithms. In contrast, the Reactome network, showing a high internal connectivity (see section 4.3) similar to the paralogy network, is better analyzed by diffusion methods probably due to the connections between cliques. This is exemplified in the case of simple genetic diseases where the use of RWR produces higher outcomes than other methods, by means of AUC values, being 0.71 compared to 0.49 obtained by DN2.

Regardless of the algorithm used, most groups only use networks derived from protein-protein interaction data to make their predictions[179,119,241,242,243] and few attempts to use other types of networks, such as those derived from metabolic[244] or regulatory[102] data, to predict disease genes have been performed. However, as shown in this work, we have demonstrated that the use of these networks is comparable to that of PPIs and even in some cases, outperform them. In fact, out of 25, only in 3 of the different combinations analyzed here PPIs rank among the top networks in terms of AUC. For example, the network derived from phylogenetic data, the Paralogy network, performs significantly better than the PPI network from HPRD[159] at predicting genes related to cancer (pvalue below 0.01), regardless of the method used to prioritize the genes. This suggests that, though cancer-related genes have been claimed to be interactome hubs[245,111,246] and several approaches have taken advantage of topological features in protein-protein interaction networks to identify cancer genes[110,247], paralogy information can be better at predicting disease-related genes. One possible explanation could be that, as it has been previously shown, cancer is usually caused by alterations of certain cellular functions or pathways rather than specific genes. One consequence of this phenomenon is that alterations in the same pathway or cellular function tend to be mutually exclusive[248]. Thus if two paralogs are performing similar functions, mutations altering one or the other might lead to cancer, which would explain this better performance of the Paralogy network over the HPRD. One factor contributing to this hypothesis is that cancer genes tend to have less paralogs[174], thus, it could be more likely that a single mutation altering one of the members of the family is enough to cause cancer. Another idea that comes from these results is that not only paralogs tend to be associated to disease by mutations in equivalent positions (as described by Yandell *et. al.*[249]) but that they are also associated to similar phenotypes.

We also have observed that diseases are predicted differently, in terms of AUC, when using distinct networks. For example, as explained before (figure 28), cancer is predicted better when using the paralogy network (AUC of 0.65) than when using HPRD (AUC of 0.57) or the coexpression network (AUC of 0.52). In this last case, it could be that cancer cells have seriously altered expression patterns (a feature that has been even used to discriminate between cancer types[250]), thus the network derived from healthy expression data is not informative enough to make any prediction in this disease. In fact, the coexpression network seems to be good only at predicting genes related to diabetes (AUC of 0.63). In the cases of HPRD and Reactome networks, there are little differences between diseases.

The main idea that emerges from these results is that networks derived from different sources of information perform differently when predicting distinct diseases. Moreover, as we have shown, different networks tend to predict different genes (reflected in figure 31, which shows that there is no correlation between the ranking of one gene in one network and the ranking of the same gene in another network). These results support the notion that networks that combine several types of data could outperform networks derived from single sources, an idea that has been already tested using various approximations and that will be discussed in the following section.

Finally, the evaluation of the influence of the different parameters in the performance of the RWR algorithm provided some interesting results. For example, the restart parameter, which in principle should affect that maximum length of the random walk and thus, could potentially alter the results, has no influence in the performance of the network (figure 32), which is something that has been previously reported[137]. Another interesting observation is that the random seed sets have some predictive power, particularly in the HPRD network (figure 33). This network had AUCs of 0.54, 0.55 and 0.60 when predicting colorectal cancer, cancer and diabetes respectively when using random seeds. Our interpretation of these results is that genes associated to diabetes, cancer or colorectal cancer tend to occupy central locations in HPRD, a phenomenon that has been described for cancer genes[174]. If that were the case, regardless of the starting point, the probability of ending in a disease gene would be greater than that of ending in a non-disease gene. Interestingly, the random seeds have no predictive power for simple genetic diseases (AUC of 0.50), which could be related to the fact that these genes, unlike those related to cancer, do not have higher connectivity than average genes in the genome[113].

Performance of the combined networks

Several methods to combine different types of biological information in form of network have been tested, but they can be classified into two categories: those that combine the information *a priori* and those that combine it *a posteriori*. Among the first group we found methods that rely on generating metanetworks using several sources of information to later run the prediction algorithm. For example, one of such methods consists in using a Bayesian classifier to infer which of all the possible interactions in a given proteome are true. The classifier is trained with a set of true-positive interactions and another set of true-negative interactions, and several sources of information are used to predict which of all the possible interactions in a set of proteins are more likely to be true. These methods create functional networks that are weighted according to the probability of the interaction to be true[143,141,140]. Another class of methods that belongs to the *a priori* group, consists in creating bi-partite graphs in which each type of biological information is represented as independent networks, connected with each other in specific points shared between the different networks[131,251,252,132].

The *a posteriori* group consists of those methods that combine the results obtained from several sources of information, after running the algorithms to produce a meta-score or a meta-ranking. This can be done by combining either the raw scores[136] or the ranking of the candidate genes[253,254,137].

In this thesis we have explored the two groups of methods by (a) integrating the scores obtained by a RWR in each single network using a Bayesian classifier, which would fall in the *a posteriori* classification, and (b) creating a meta-network by adding the different networks, which would be included among the *a priori* methods.

Combining the networks seems to perform better than combining the scores

The first interesting result is that the Bayesian classifier trained with the RWR scores always performed worse than the addition of the 4 networks, with the exception of neurodegenerative disorders, though a strong tendency in that direction was also observed in that case. In figure 36 one can observe that, for example, in the case of cancer the AUC of the H-P-R-C network is around 0.62 whereas the Bayesian classifier has an AUC of 0.65 (pvalue below 0.01). These results suggest that when predicting genes related to disease it is better to combine the information from the different networks by directly adding them to create a meta-network than to combine their distinct scores, at least using a Bayesian classifier. However, it is important to keep in mind that there are several other approaches that fall in the *a posteriori* group that we could have tried, thus we cannot generalize this affirmation beyond the scope of our results. One approach that might improve the results of combining the scores is the one described by Chen *et. al.*[136]. In that work they combined the scores derived from running the DK algorithm in several networks by, among other things, selecting only the best network for each gene (the one that give the gene the highest disease score). By doing this they claimed an AUC of 0.80 at predicting diseases from OMIM. Another possibility would be to use ranking statistics instead of the raw scores. This idea, as we have explained before, has also been applied before and several variations have been tried, being probably ENDEAVOUR it's most well-known example[138].

Combined vs single networks

There are several ways to combine two or more networks. Here we have explored 3 of them: juxtaposition, addition and weighted addition. The first method consists in the union of "N" networks giving all the edges the same weight regardless of the number of networks where the edge appears. The addition method, again, adds "N" networks but gives the edges a weight proportional to the number of networks where it appears. Finally, the weighted addition generates a metanetwork where the edges are weighted according to the sum of weights of their original networks (if an edge appears in two networks, one of which has a weight of 1 and the other a weight of 0.7, the edge in the metanetwork will have a weight of 1.7).

We first compared the results obtained when juxtaposing or adding the different networks. As we have shown in figure 34 there are no differences at all between the two. This is likely to be caused by the fact that the number of edges shared between networks is very low (figure 24). Thus, the influence of their weight in the overall performance is also quite small.

We then compared the performance of the individual networks with that of their different combinations. The first important result that we obtained is that the combination of two or more networks does not always outperform the networks alone. In fact, we could not obtain any combination of networks that performed better than the best network alone in 4 of the 5 diseases. The only exception was diabetes, where the network resulting from the addition of H-P-R outperformed the best single network in that disease, HPRD (AUCs of 0.73 and 0.70 respectively, pvalue 0.03). However, for the other diseases the resulting meta-network could perform only as good as the best individual network, in terms of AUC. Nevertheless, it is worth to note that the coverage of the meta-network is larger than that of the networks alone. Thus, globally speaking, if a meta-network has the same AUC than a simple network, since it includes more disease genes, one could say that it is better to predict disease-related genes.

These results agree with those described in a recent publication by Gonçalves *et. al.*[180]. In that publication they compared the performance of the whole STRING functional network[255] with that of the networks derived from each source of information at predicting disease-associated genes. They observed that the network derived from the integration of the individual networks was not better than the best individual source of information for that network, which was text mining.

Influence of the amount of information

Given that the different networks represented very different types of relationships between genes (as reflected by the little amount of shared edges between networks, figure 24) and that they tend to predict different genes (as shown in figure 31, there is no correlation between the ranks obtained by the same genes in different networks), one might think that metanetworks including more sources of information would perform better than those with less types of information. Moreover, some publications describing methods to prioritize lists of putative disease-related genes disclaim the results obtained when using different combinations of information to perform their predictions and, usually, the combination that performs the best is the one that uses all types of information[136,138]. However, our results seem to indicate that there is no correlation between the amount of information used to generate the metanetwork and its performance.

The first evidence in that direction comes from the fact that, as we have explained, in 4 of the 5 diseases we have not found any combination of networks that outperform the best single network (figure 35). Something similar happens when comparing the performance of metanetworks resulting from the addition of 2 or 3 networks. For example, in the case of cancer the H-P (0.69) or P-R (0.66) networks performed as good as the H-P-R (0.68) and better than the P-R-C (0.61, pvalues below 0.01 in both cases) and H-R-C networks (0.54, pvalues below 0.01 in both cases). When analyzing genes associated to other diseases the conclusions are similar. When predicting genes related to diabetes, for instance, the P-R network (0.70) performed as good as the P-R-C network (0.67). Moreover, recent publications suggest that tissue-specific networks perform better than global networks at predicting genes involved in phenotypes[256]. However, when we tried to include this information by adding the coexpression network to the other networks adding this type of information decreased the performance of the other networks. Two facts support this idea: (I) the combination of the HPRD, Paralogy and Reactome networks was among the best combinations in all the diseases, whereas the combination of all networks was rarely among the best performers; (II) the optimized weight of the coexpression network calculated by simulated annealing was the lowest (below 0.5) in all the diseases. More evidence supporting this lack of correlation between amount of information and performance comes form the observation the functional network by Lee *et. al.*[143] that integrates up to 21 types of biological information did not perform better in any disease than the combination of only 3 networks: HPRD, Reactome and Paralogy (figure 38). Finally, the weighted metanetwork H-P-R-D optimized by simulated annealing, while performed better than the network resulting from the simple addition of these four networks, did not outperform in any case the H-P-R network.

Prediction of CCR driver genes

In a recent publication Bömingen *et. al.* evaluated 8 different gene prioritization tools[257], including some network-based such as *Gene Wanderer* by Köhler *et. al.*[127]. They assessed the performance of each gene prioritization tool using a series of gene-disease associations that were not included in the training set of the methods. By doing this, they could test the performance of the methods in a setting more resembling of a real-world scenario. They observed that, while all methods could predict the new disease-gene associations to a certain degree, most of them performed worse than what was claimed in the original paper. They attributed this result to an overestimation of the performance by cross-validation based benchmarks.

In order to assess whether we were properly estimating the performance of our methods and to test them with external data, we tried to predict cancer driver genes that were not associated to cancer in our database. Aiming for that goal, we downloaded the latest available summary of mutations associated to colorectal cancer in COSMIC (v61). We defined our set of driver as those who had been found mutated in at least 15 tumor samples. This rendered a list of 482 genes. Of these, 129 and 230 were already associated to CCR and cancer respectively in our database. We tried to predict the remaining 353 and 252 genes using as seed sets all the CCR and all the cancer genes respectively, using the network derived from the addition of HPRD, Paralogy and Reactome and a RWR.

There were no differences in terms of AUC when trying to predict the driver colorectal cancer genes using the cancer seed (0.66) or the CCR seed (0.65). This suggests that genes associated to related phenotypes in the DO can be used to predict genes related to the disease of interest when using our method. This is an idea that has been explored before, but calculating the similarity between diseases using text-mining approaches instead of ontologies[106,105]. Moreover, these values are also quite close to those obtained in the cross-validation benchmarking, 0.64 for colorectal cancer and 0.68 for cancer in general. This suggests that our cross-validation settings give us a good and objective idea of the performance of the different methods.

Methods trying to predict cancer driver genes or mutations that make use mutation frequency data are becoming very popular. For example in a recent publication Gonzalez-Perez *et. al.* described a method that integrates pathogenic scores from several predictors with mutation frequency data[258]. Moreover, some other groups added, for instance, the p values of GWAS studies to the candidate genes before running the network algorithm[242,142]. Therefore, we also tried to improve the results of our method by adding the gene mutation frequency data to the seed used to start the RWR. This adjusted seed, however, did not improve the AUC values of the method (0.65 and 0.66 for cancer and CCR respectively). One hypothesis explaining this lack of improvement is that in the H-P-R network colorectal cancer genes seem to be central, as random seeds have some predictive power in this network (AUC of 0.56), whereas randomized networks do not (AUC of 0.50).

We also simulated GWAS data by creating, for each colorectal cancer gene, a group of genes made of the 9 or 24 closest genes in the genome. We next ranked the group according to the scores obtained in the H-P-R network using the RWR and observed the ranking of the true driver gene. As we have shown in figure s 39 and 40, there is indeed an enrichment of driver genes towards the first positions of the ranking. This supports the idea that our method is able to identify known-driver genes.

Summary

Despite previous reports suggesting otherwise, our results indicate that difussion-based methods do not necessarily perform better than simple methods like direct neighbor counting, and whether they do or not seems to be influenced by the topology of the network being used. Similarly, though most people solely use protein-protein interaction networks our data suggest that networks derived from other sources of information can perform as good as this network, or even better, depending on the disease being studied.

Adding various types of biological information in form of networks can improve the predictive power of the resulting meta-network over the individual networks, in terms of coverage of disease genes, while maintaining the AUC values of the best network. However, this is not straightforward as depending on the type of information one is combining, the resulting combined network might perform worse. Another layer of complexity comes from the fact that, depending on the disease of interest, the optimal network combination might differ, though some general trends might be observed (for example, it seems that adding the coexpression network to any other network decreases its predictive power). Moreover, according to our results, it seems that it is better to combine the information from various networks by directly adding the networks than by combining the scores of the different genes in the obtained using the different networks alone. Finally, we have been able to use the combination of the HPRD, Paralogy and Reactome networks to successfully identify known driver genes in colorectal cancer.

Overall it seems that there is no rule of thumb on which networks and algorithms will perform better at predicting each disease-type, which is something that should be taken into account when designing automatic pipelines to prioritize list of disease-related genes. This is an aspect that is going to be important in the near future as whole genome/exome sequencing becomes more widespread or international projects, like the Cancer Genome Atlas[58] or the International Cancer Genome Consortium[259], make their data available to the public releasing hundreds of thousands of putative cancer driver mutations.

5.4. Implications of this work and future perspective

Besides the possible application of the associations found between the different ontologies in the first part of this project to either build a classifier of pathogenic mutations or to generate hypotheses on their biological origin (such as the role of the UPR in tumor evolution), our results highlight the importance of normalizing biomedical data using ontologies in the OMICs era. Complete normalization of biological data would help to automate the data analysis and increase our capability of hypothesis generation, a must when the data to analyze comes from thousands of experiments performed with samples of thousands of patients. Some successful efforts in this line are already happening. The clearest example would be the normalization of gene function annotations using GO and their widespread use to analyze genome-wide experiments. Another example of biological data that is currently being normalized would be the annotation of variations in the genome with the SO that we have commented. However, the type of biomedical data that probably would benefit scientists the most of its normalization is disease description. Though there have been some efforts in this direction, such as the example of IntOGen that we have commented, or the use of ICD10, another disease ontology, to describe tumor samples in the TCGA project, more work needs to be done in order to fully exploit this type of data. One great example of the potential use of normalized disease description is the recent work by Roque *et. al.* using data from electronic records of Danish patients done by Denmark's government, which has been normalized using ICD10.

Regarding the second part of the project, we have shown that it is important to make no *a priori* assumptions on which networks or algorithms will be the best at prioritizing genes associated to the disease of our interest. As we have commented, this is important when designing automated pipelines that analyze putative disease-related genes or mutations. Another group of methods that might be worth exploring are those that rely on centrality measures instead of closeness to other disease genes. As we have seen, while are able to predict disease-related genes, our AUC values range between 0.60 and 0.80, thus, there is still room for improvement. It is possible that, just like the Reactome network benefits of diffusion-based methods, other networks representing other types of data might benefit from methods based on centrality measures.

Another further development of the results described in this work could be the integration of the disease-specific mutated features with the metanetworks to create a mutation classifier that uses, not only network data, but also information coming from the mutation. While we have tried to include some of this information by weighting the seed genes according to their mutation propensity, this approach is not extensible to other diseases than cancer and, as we have seen, our method seems to not benefit from this idea. Thus, a framework integrating the disease-specific features with the metanetwork and the network algorithm could perform better than the metanetwork alone. If that were the case, unlike the mutation-frequency approach, this idea would be extensible to other diseases than cancer.

6. Conclusions

1. Mutations related to different diseases show a bias in the features that they are affecting, not only when compared to nsSNPs, but also when compared to mutations related to other diseases

2. Mutations in the same gene can be associated to different diseases, not only by disrupting different protein domains, but also by altering regions showing distinct biochemical features

3. Simple methods like direct neighbor counting can perform as good as more complex methods, such as those based on diffusion

4. Networks derived from other types of biological information show performances comparable those derived from protein-protein interaction data

5. Combining networks that exploit different kinds of information (*a priori* approaches) outperforms the combination of individual scores (*a posteriori* approaches) when predicting disease-related genes

6. Increasing the number of sources of information used to generate a meta-network does not imply an increase in the performance of the network at predicting disease-related genes

7. Bibliography

1. Garrod, a E. The incidence of alkaptonuria: a study in chemical individuality. 1902. *Molecular medicine (Cambridge, Mass.)* **2**, 274–82 (1996).
2. Avery, O., MacLeod, C. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *Journal of Experimental Medicine* **79**, 137–157 (1943).
3. Vischer, E. & Chargaff, E. The separation and quantiative estimation of purines and pyrimidines in minute amounts. *Journal of Biological Chemistry* **176**, 703–714 (1948).
4. Watson, J. & Crick, F. Molecular structure of nucleic acids. *Nature* **171**, 737–738 (1953).
5. Baglioni, C. The fusion of two peptide chains in hemoglobin Lepore and its interpretation as a genetic deletion. *Proceedings of the National Academy of Sciences of the United States of America* **48**, 1880–1886 (1962).
6. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research* **33**, D514 (2005).
7. Ross, C. a & Tabrizi, S. J. Huntington's disease: from molecular pathogenesis to clinical treatment. *Lancet neurology* **10**, 83–98 (2011).
8. Yamada, M., Tsuji, S. & Takahashi, H. Genotype-phenotype correlation in CAG-repeat diseases. *Neuropathology* **22**, 317–322 (2002).
9. Pagani, F., Raponi, M. & Baralle, F. E. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 6368–72 (2005).
10. Komar, a a, Lesnik, T. & Reiss, C. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS letters* **462**, 387–91 (1999).
11. Bartoszewski, R. a *et al.* A synonymous single nucleotide polymorphism in DeltaF508 CFTR alters the secondary structure of the mRNA and the expression of the mutant protein. *The Journal of biological chemistry* **285**, 28741–8 (2010).
12. Mégarbané, A. *et al.* The 50th anniversary of the discovery of trisomy 21: the past, present, and future of research and treatment of Down syndrome. *Genetics in medicine : official journal of the American College of Medical Genetics* **11**, 611–6 (2009).
13. Davoli, T. & De Lange, T. The causes and consequences of polyploidy in normal development and cancer. *Annual review of cell and developmental biology* **27**, 585–610 (2011).
14. Otter, M., Schrander-Stumpel, C. T. R. M. & Curfs, L. M. G. Triple X syndrome: a review of the literature. *European journal of human genetics : EJHG* **18**, 265–71 (2010).
15. Lettice, L. a *et al.* Enhancer-adoption as a mechanism of human developmental disease. *Human mutation* **32**, 1492–9 (2011).

16. Zollino, M. *et al.* Mapping the Wolf-Hirschhorn syndrome phenotype outside the currently accepted WHS critical region and defining a new critical region, WHSCR-2. *American journal of human genetics* **72**, 590–7 (2003).
17. Tunnacliffe, A. *et al.* Localization of Jacobsen Syndrome Breakpoints on a 40-Mb Physical Map of Distal Chromosome 11q Localization of Jacobsen Syndrome Breakpoints on a 40-Mb Physical Map of Distal Chromosome 11q. 44–52 (1999).doi:10.1101/gr.9.1.44
18. Nowell, P. A minute chromosome in chronic granulocytic leukemia. *Science* **132**, 14–16 (1960).
19. Walker, F. O. Huntington's Disease. *The Lancet* **369**, 218–28 (2007).
20. Myerowitz, R. Tay-Sachs disease-causing mutations and neutral polymorphisms in the Hex A gene. *Human mutation* **9**, 195–208 (1997).
21. Tosi, I., Toledo-Leiva, P., Neuwirth, C., Naoumova, R. P. & Soutar, A. K. Genetic defects causing familial hypercholesterolaemia: identification of deletions and duplications in the LDL-receptor gene and summary of all mutations found in patients attending the Hammersmith Hospital Lipid Clinic. *Atherosclerosis* **194**, 102–11 (2007).
22. Macpherson, A., Bjarnason, I. & Forgacs, I. Discovery of the gene for familial adenomatous polyposis. *BMJ: British Medical …* **304**, (1992).
23. Ellis, P. E., Dawson, M. & Dixon, M. J. Mutation testing in Treacher Collins Syndrome. *Journal of orthodontics* **29**, 293–7; discussion 278 (2002).
24. Chang, J. & Kan, Y. beta 0 thalassemia, a nonsense mutation in man. *Proceedings of the National Academy of Sciences of the United States of America* **76**, 2886–2889 (1979).
25. Prat, J., Ribé, A. & Gallardo, A. Hereditary ovarian cancer. *Human pathology* **36**, 861–70 (2005).
26. Ingram, V. Gene Mutations in Human Hemoglobin: The Chemical Difference between Normal and Sickle Hemoglobin. *Nature* (1957).
27. Tartaglia, M. *et al.* Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. *Nature genetics* **29**, 465–8 (2001).
28. Africa, W. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–73 (2010).
29. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
30. Venter, J. C. *et al.* The sequence of the human genome. *Science (New York, N.Y.)* **291**, 1304–51 (2001).
31. Mardis, E. R. Cancer genomics identifies determinants of tumor biology. *Genome biology* **11**, 211 (2010).
32. Endesfelder, D. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *The New England Journal of Medicine* **366**, 883–892 (2012).
33. Barrett, J. C. *et al.* Genome-wide association defines more than thirty distinct susceptibility loci for Crohn's disease. *Nature genetics* **40**, 955–962 (2008).
34. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–52 (2009).

35. Pandey, J. P. Genomewide association studies and assessment of risk of disease. *The New England journal of medicine* **363**, 2076–7; author reply 2077 (2010).

36. Dixit, A., Torkamani, A., Schork, N. J. & Verkhivker, G. Computational modeling of structurally conserved cancer mutations in the RET and MET kinases: the impact on protein structure, dynamics, and stability. *Biophysical journal* **96**, 858–74 (2009).

37. Herrgard, S. *et al.* Prediction of deleterious functional effects of amino acid mutations using a library of structure-based function descriptors. *Proteins* **53**, 806–16 (2003).

38. Ng, P. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome research* 863–874 (2001).doi:10.1101/gr.176601.1

39. Cheng, J., Randall, A. & Baldi, P. Prediction of protein stability changes for singlesite mutations using support vector machines. *Proteins: Structure, Function, and …* **1132**, 1125–1132 (2006).

40. Zhang, K. X. & Ouellette, B. F. F. CAERUS: predicting CAncER oUtcomeS using relationship between protein structural information, protein networks, gene expression data, and mutation data. *PLoS computational biology* **7**, e1001114 (2011).

41. Bromberg, Y., Yachdav, G. & Rost, B. SNAP predicts effect of mutations on protein function. *Bioinformatics* **24**, 2397 (2008).

42. Ku, C.-S., Naidoo, N. & Pawitan, Y. Revisiting Mendelian disorders through exome sequencing. *Human genetics* **129**, 351–70 (2011).

43. Bonnefond, A. *et al.* Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome. *PloS one* **5**, e13630 (2010).

44. Do, R., Kathiresan, S. & Abecasis, G. R. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Human molecular genetics* **21**, R1–9 (2012).

45. Baudot, A., Real, F. X., Izarzugaza, J. M. G. & Valencia, A. From cancer genomes to cancer models: bridging the gaps. *EMBO reports* **10**, 359–66 (2009).

46. Kumar, A. *et al.* Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 17087–92 (2011).

47. Sjöblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science (New York, N.Y.)* **314**, 268–74 (2006).

48. Greenman, C., Stephens, P. & Smith, R. Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).

49. Nehrt, N. L., Peterson, T. a, Park, D. & Kann, M. G. Domain landscapes of somatic mutations in cancer. *BMC genomics* **13 Suppl 4**, S9 (2012).

50. Shan, Y. *et al.* Oncogenic mutations counteract intrinsic disorder in the EGFR kinase and promote receptor dimerization. *Cell* **149**, 860–70 (2012).

51. Fröhling, S. *et al.* Identification of Driver and Passenger Mutations of FLT3 by High-Throughput DNA Sequence Analysis and Functional Assessment of Candidate Alleles. *Cancer Cell* 501–513 (2007).doi:10.1016/j.ccr.2007.11.005

52. Ferrer-costa, C. *et al.* Structural bioinformatics PMUT : a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics* **21**, 3176–3178 (2005).

53. Torkamani, A., Kannan, N., Taylor, S. S. & Schork, N. J. Congenital disease SNPs target lineage specific structural elements in protein kinases. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 9011–6 (2008).

54. Dixit, A. *et al.* Sequence and structure signatures of cancer mutation hotspots in protein kinases. *PloS one* **4**, e7485 (2009).

55. Pasqualucci, L. *et al.* Analysis of the coding genome of diffuse large B-cell lymphoma. *Nature genetics* **43**, 830–7 (2011).

56. Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150**, 251–63 (2012).

57. Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473–7 (2010).

58. Cancer, T. & Atlas, G. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–15 (2011).

59. Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science (New York, N.Y.)* **321**, 1801–6 (2008).

60. Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science (New York, N.Y.)* **321**, 1807–12 (2008).

61. Vavouri, T., Semple, J. I., Garcia-Verdugo, R. & Lehner, B. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell* **138**, 198–208 (2009).

62. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradović, Z. & Dunker, A. K. Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins. *Journal of Molecular Biology* **323**, 573–584 (2002).

63. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research* **40**, D109–14 (2012).

64. Stevens, R., Goble, C. a & Bechhofer, S. Ontology-based knowledge representation for bioinformatics. *Briefings in bioinformatics* **1**, 398–414 (2000).

65. Huang, D. W., Sherman, B. T. & Lempicki, R. a Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* **37**, 1–13 (2009).

66. The Gene Ontology Consortium The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* **32**, 258–261 (2004).

67. Subramanian, A. Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the …* **102**, 15545–15550 (2005).

68. Kim, S.-Y. & Volsky, D. J. PAGE: parametric analysis of gene set enrichment. *BMC bioinformatics* **6**, 144 (2005).

69. Tu, K., Yu, H. & Zhu, M. MEGO : gene functional module expression based on gene ontology. *Biotechniques* **38**, 277–283 (2005).

70. Huang, D. W. *et al.* The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome biology* **8**, R183 (2007).

71. Fröhlich, H., Speer, N., Poustka, A. & Beissbarth, T. GOSim-an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC bioinformatics* **8**, 166 (2007).

72. López-Bigas, N., Blencowe, B. J. & Ouzounis, C. a Highly consistent patterns for inherited human diseases at the molecular level. *Bioinformatics (Oxford, England)* **22**, 269–77 (2006).

73. Tabibiazar, R. *et al.* Signature patterns of gene expression in mouse atherosclerosis and their correlation to human coronary disease. *Physiological genomics* **22**, 213–26 (2005).

74. Costello, C. M. *et al.* Dissection of the inflammatory bowel disease transcriptome using genome-wide cDNA microarrays. *PLoS medicine* **2**, e199 (2005).

75. Jiang, W. *et al.* Constructing disease-specific gene networks using pair-wise relevance metric: application to colon cancer identifies interleukin 8, desmin and enolase 1 as the central elements. *BMC systems biology* **2**, 72 (2008).

76. Shi, Z., Derow, C. K. & Zhang, B. Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. *BMC systems biology* **4**, 74 (2010).

77. Hu, P., Bader, G., Wigle, D. a & Emili, A. Computational prediction of cancer-gene function. *Nature reviews. Cancer* **7**, 23–34 (2007).

78. Smith, B. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology* **25**, 1251–5 (2007).

79. Tirrell, R. *et al.* An ontology-neutral framework for enrichment analysis. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* **2010**, 797–801 (2010).

80. LePendu, P., Musen, M. a & Shah, N. H. Enabling enrichment analysis with the Human Disease Ontology. *Journal of biomedical informatics* (2011).doi:10.1016/j.jbi.2011.04.007

81. Mort, M. *et al.* In Silico functional profiling of human disease-associated and polymorphic amino acid substitutions. *Human Mutation* (2010).doi:10.1002/humu.21192

82. Roque, F. S. *et al.* Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS computational biology* **7**, e1002141 (2011).

83. Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–68 (2005).

84. Poirel, C. L., Owens, C. C. & Murali, T. M. Network-based functional enrichment. *BMC Bioinformatics* **12**, S14 (2011).

85. Geistlinger, L., Csaba, G., Küffner, R., Mulder, N. & Zimmer, R. From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics (Oxford, England)* **27**, i366–73 (2011).

86. Glaab, E., Baudot, A., Krasnogor, N., Schneider, R. & Valencia, A. EnrichNet: network-based gene set enrichment analysis. *Bioinformatics (Oxford, England)* **28**, i451–i457 (2012).

87. Aral, S. & Walker, D. Identifying influential and susceptible members of social networks. *Science (New York, N.Y.)* **337**, 337–41 (2012).
88. Eagle, N., Macy, M. & Claxton, R. Network diversity and economic development. *Science (New York, N.Y.)* **328**, 1029–31 (2010).
89. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, a L. The large-scale organization of metabolic networks. *Nature* **407**, 651–4 (2000).
90. Erdös, P. & Rényi, A. On random graphs. *Publ. Math. Debrecen* 290–297 (1959).
91. Watts, D. J. & Strogatz, S. H. Collective dynamics of "small-world" networks. *Nature* **393**, 440–2 (1998).
92. Barabási, A. & Albert, R. Emergence of Scaling in Random Networks. *science* **286**, 509–512 (1999).
93. Jeong, H., Mason, S. P., Barabási, a L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–2 (2001).
94. Albert, R., Jeong, H. & Barabasi, A. Error and attack tolerance of complex networks. *Nature* **406**, 378–82 (2000).
95. Dunne, J. a, Williams, R. J. & Martinez, N. D. Food-web structure and network theory: The role of connectance and size. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 12917–22 (2002).
96. Chen, B. L., Hall, D. H. & Chklovskii, D. B. Wiring optimization can relate neuronal structure and function. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 4723–8 (2006).
97. Li, S. *et al.* A map of the interactome network of the metazoan C. elegans. *Science (New York, N.Y.)* **303**, 540–3 (2004).
98. Gandhi, T., Zhong, J. & Mathivanan, S. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature genetics* **38**, 285–293 (2006).
99. Kuchaiev, O. & Przulj, N. Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics (Oxford, England)* **27**, 1390–6 (2011).
100. Duarte, N. C. *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 1777–82 (2007).
101. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science (New York, N.Y.)* **309**, 1559–63 (2005).
102. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).
103. Lee, T. I. *et al.* Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science (New York, N.Y.)* **298**, 799–804 (2002).
104. Albert, R. Scale-free networks in cell biology. *Journal of cell science* **118**, 4947–57 (2005).
105. Goh, K., Cusick, M., Valle, D. & Childs, B. The human disease network. *Proceedings of the National Academy of Sciences of the United States of America* (2007).
106. Van Driel, M. a, Bruggeman, J., Vriend, G., Brunner, H. G. & Leunissen, J. a M. A text-mining analysis of the human phenome. *European journal of human genetics : EJHG* **14**, 535–42 (2006).

107. Beadle, G. & Tatum, E. Genetic control of biochemical reactions in neurospora. *Proceedings of the National Academy of …* **27**, (1941).

108. Vidal, M., Cusick, M. E. & Barabási, A.-L. Interactome networks and human disease. *Cell* **144**, 986–98 (2011).

109. Zhong, Q. *et al.* Edgetic perturbation models of human inherited disorders. *Molecular systems biology* **5**, 321 (2009).

110. Jonsson, P. F. & Bates, P. a Global topological features of cancer proteins in the human interactome. *Bioinformatics (Oxford, England)* **22**, 2291–7 (2006).

111. Xia, J., Sun, J., Jia, P. & Zhao, Z. Do cancer proteins really interact strongly in the human protein-protein interaction network? *Computational biology and chemistry* **35**, 121–125 (2011).

112. Xu, J. & Li, Y. Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics (Oxford, England)* **22**, 2800–5 (2006).

113. Cai, J. J., Borenstein, E. & Petrov, D. a Broker genes in human disease. *Genome biology and evolution* **2**, 815–25 (2010).

114. Schwikowski, B., Uetz, P. & Fields, S. A network of protein-protein interactions in yeast. *Nature biotechnology* **18**, 1257–61 (2000).

115. Navlakha, S. & Kingsford, C. The power of protein interaction networks for associating genes with diseases. *Bioinformatics (Oxford, England)* **26**, 1057–63 (2010).

116. Wang, J., Chen, G., Li, M. & Pan, Y. Integration of breast cancer gene signatures based on graph centrality. *BMC systems biology* **5 Suppl 3**, S10 (2011).

117. Milenkovic, T., Memisevic, V., Ganesan, A. K. & Przulj, N. Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data. *Journal of the Royal Society, Interface / the Royal Society* **7**, 423–37 (2010).

118. Gillis, J. & Pavlidis, P. "Guilt by association" is the exception rather than the rule in gene networks. *PLoS computational biology* **8**, e1002444 (2012).

119. Oti, M., Snel, B., Huynen, M. a & Brunner, H. G. Predicting disease genes using protein-protein interactions. *Journal of medical genetics* **43**, 691–8 (2006).

120. Nabieva, E., Jim, K., Agarwal, A., Chazelle, B. & Singh, M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics (Oxford, England)* **21 Suppl 1**, i302–10 (2005).

121. Chua, H. N., Sung, W.-K. & Wong, L. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics (Oxford, England)* **22**, 1623–30 (2006).

122. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, a W. From molecular to modular cell biology. *Nature* **402**, C47–52 (1999).

123. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 7821–6 (2002).

124. Nibbe, R. K., Markowitz, S., Myeroff, L., Ewing, R. & Chance, M. R. Discovery and scoring of protein interaction subnetworks discriminative of late stage human colon cancer. *Molecular & cellular proteomics : MCP* **8**, 827–45 (2009).

125. Narayanan, T., Gersten, M., Subramaniam, S. & Grama, A. Modularity detection in protein-protein interaction networks. *BMC research notes* **4**, 569 (2011).

126. Ahn, Y.-Y., Bagrow, J. P. & Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **466**, 761–4 (2010).

127. Köhler, S., Bauer, S., Horn, D. & Robinson, P. Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human …* 949–958 (2008).doi:10.1016/j.ajhg.2008.02.013.

128. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. & Sharan, R. Associating genes and protein complexes with disease via network propagation. *PLoS computational biology* **6**, e1000641 (2010).

129. Milenković, T. & Przulj, N. Uncovering biological network function via graphlet degree signatures. *Cancer informatics* **6**, 257–73 (2008).

130. Przulj, N. Biological network comparison using graphlet degree distribution. *Bioinformatics (Oxford, England)* **23**, e177–83 (2007).

131. Yang, P., Li, X., Wu, M., Kwoh, C.-K. & Ng, S.-K. Inferring gene-phenotype associations via global protein complex network propagation. *PloS one* **6**, e21502 (2011).

132. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology* **25**, 309–16 (2007).

133. Radivojac, P. *et al.* An integrated approach to inferring gene-disease associations in humans. *Proteins* **72**, 1030–7 (2008).

134. Guo, X. *et al.* A computational method based on the integration of heterogeneous networks for predicting disease-gene associations. *PloS one* **6**, e24171 (2011).

135. Wu, X., Liu, Q. & Jiang, R. Align human interactome with phenome to identify causative genes and networks underlying disease families. *Bioinformatics (Oxford, England)* **25**, 98–104 (2009).

136. Chen, Y. *et al.* In silico gene prioritization by integrating multiple data sources. *PloS one* **6**, e21137 (2011).

137. Li, Y. & Patra, J. C. Integration of multiple data sources to prioritize candidate genes using discounted rating system. *BMC bioinformatics* **11 Suppl 1**, S20 (2010).

138. Aerts, S. *et al.* Gene prioritization through genomic data fusion. *Nature biotechnology* **24**, 537–44 (2006).

139. Li, L. *et al.* Discovering cancer genes by integrating network and functional properties. *BMC medical genomics* **2**, 61 (2009).

140. Wu, G., Feng, X. & Stein, L. A human functional protein interaction network and its application to cancer data analysis. *Genome biology* **11**, R53 (2010).

141. Franke, L. *et al.* Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *American journal of human genetics* **78**, 1011–25 (2006).

142. Lee, I., Blom, U. M., Wang, P. I., Shim, J. E. & Marcotte, E. M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome research* **21**, 1109–21 (2011).

143. Lee, I. *et al.* A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. *Nature genetics* **40**, 181–8 (2008).

144. Jones, P., Martens, L., Apweiler, R., Hermjakob, H. & Co, R. G. The Ontology Lookup Service : more data and better tools for controlled vocabulary queries. *Exchange Organizational Behavior Teaching Journal* **36**, 372–376 (2008).

145. Eilbeck, K. *et al.* The Sequence Ontology: a tool for the unification of genome annotations. *Genome biology* **6**, R44 (2005).

146. Becker, K. G., Barnes, K. C., Bright, T. J. & Wang, S. A. The Genetic Association Database. *Nature genetics* **36**, 431–432 (2004).

147. Flicek, P. *et al.* Ensembl 2011. *Nucleic acids research* **39**, D800–6 (2011).

148. Forbes, S. *et al.* The catalogue of somatic mutations in cancer (COSMIC). *Current protocols in human genetics* (2008).doi:10.1002/0471142905.hg1011s57.The

149. Brais, B. *et al.* Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. *Nature Genetics* **18**, 164–167 (1998).

150. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology* **292**, 195–202 (1999).

151. Ren, J., Wen, L. & Gao, X. CSS-Palm 2.0: an updated software for palmitoylation sites prediction. *Protein Engineerin, Design & Selection* **21**, 639–644 (2008).

152. Xue, Y., Zhou, F., Fu, C., Xu, Y. & Yao, X. SUMOsp: a web server for sumoylation site prediction. *Nucleic acids research* **34**, W254–7 (2006).

153. Shatkay, H. *et al.* SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. *Bioinformatics* **23**, 1410 (2007).

154. Blom, N., Gammeltoft, S. & Brunak, S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of molecular biology* **294**, 1351–62 (1999).

155. Julenius, K., Mølgaard, A., Gupta, R. & Brunak, S. Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology* **15**, 153–64 (2005).

156. Prilusky, J. *et al.* Structural bioinformatics FoldIndex © : a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* **21**, 3435–3438 (2005).

157. Alexa, A., Rahnenführer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics (Oxford, England)* **22**, 1600–7 (2006).

158. Cancer, T. & Atlas, G. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–8 (2008).

159. Keshava Prasad, T. S. *et al.* Human Protein Reference Database--2009 update. *Nucleic acids research* **37**, D767–72 (2009).

160. Wu, C. *et al.* BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome biology* **10**, R130 (2009).

161. Wu, Z., Irizarry, R., Gentleman, R., Murillo, F. & Spencer, F. A Model Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the american statistical association* **99**, 909–9147 (2004).

162. Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research* **39**, D691–7 (2011).
163. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics (Oxford, England)* **21**, 3940–1 (2005).
164. Lee, I. *et al.* Predicting genetic modifier loci using functional gene networks. *Genome research* **20**, 1143–53 (2010).
165. Bamford, S. *et al.* The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British journal of cancer* **91**, 355–8 (2004).
166. Reumers, J., Schymkowitz, J. & Rousseau, F. Using structural bioinformatics to investigate the impact of non synonymous SNPs and disease mutations: scope and limitations. *BMC bioinformatics* **10 Suppl 8**, S9 (2009).
167. Jiang, R. *et al.* Sequence-based prioritization of nonsynonymous single-nucleotide polymorphisms for the study of disease mutations. *American journal of human genetics* **81**, 346–60 (2007).
168. Thusberg, J. & Vihinen, M. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. *Human mutation* **30**, 703–14 (2009).
169. Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L. & Casadio, R. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Human Mutation* **30**, 1237–1244 (2009).
170. Wachi, S., Yoneda, K. & Wu, R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics (Oxford, England)* **21**, 4205–8 (2005).
171. Osborne, J. D. *et al.* Annotating the human genome with Disease Ontology. *BMC genomics* **10 Suppl 1**, S6 (2009).
172. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* **30**, 3059–66 (2002).
173. Mushegian, A. R., Bassett, D. E., Boguski, M. S., Bork, P. & Koonin, E. V Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 5831–6 (1997).
174. Rambaldi, D., Giorgi, F. M., Capuani, F., Ciliberto, A. & Ciccarelli, F. D. Low duplicability and network fragility of cancer genes. *Trends in genetics* **24**, 427–430 (2008).
175. Furney, S. J., Higgins, D. G., Ouzounis, C. a & López-Bigas, N. Structural and functional properties of genes involved in human cancer. *BMC genomics* **7**, 3 (2006).
176. Haynes, C. & Iakoucheva, L. M. Serine/arginine-rich splicing factors belong to a class of intrinsically disordered proteins. *Nucleic acids research* **34**, 305–12 (2006).
177. Harrison, P. M. Exhaustive assignment of compositional bias reveals universally prevalent biased regions: analysis of functional associations in human and Drosophila. *BMC bioinformatics* **7**, 441 (2006).
178. Iakoucheva, L. M. *et al.* The importance of intrinsic disorder for protein phosphorylation. *Nucleic acids research* **32**, 1037 (2004).

179. Bauer, S., Horn, D. & Robinson, P. N. Walking the Interactome for Prioritization of Candidate Disease Genes. 949–958 (2008).doi:10.1016/j.ajhg.2008.02.013.

180. Gonçalves, J. P., Francisco, A. P., Moreau, Y. & Madeira, S. C. Interactogeneous: Disease Gene Prioritization Using Heterogeneous Networks and Full Topology Scores. *PLoS ONE* **7**, e49634 (2012).

181. Haider, S. *et al.* BioMart Central Portal--unified access to biological data. *Nucleic acids research* **37**, W23–7 (2009).

182. Kapushesky, M. *et al.* Gene expression atlas at the European bioinformatics institute. *Nucleic acids research* **38**, D690–8 (2010).

183. Forbes, S. A. *et al.* COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Research* 1–6 (2009).doi:10.1093/nar/gkp995

184. Mottaz, A., Yip, Y. L., Ruch, P. & Veuthey, A.-L. Mapping proteins to disease terminologies: from UniProt to MeSH. *BMC bioinformatics* **9 Suppl 5**, S3 (2008).

185. Mooney, S. D. & Klein, T. E. The functional importance of disease-associated mutation. *BMC Bioinformatics* **5**, 1–5 (2002).

186. Wang, X. *et al.* Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature biotechnology* **30**, 159–64 (2012).

187. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science (New York, N.Y.)* **318**, 1108–13 (2007).

188. Lahiry, P., Torkamani, A., Schork, N. J. & Hegele, R. a Kinase mutations in human disease: interpreting genotype-phenotype relationships. *Nature reviews. Genetics* **11**, 60–74 (2010).

189. Torkamani, A. & Schork, N. J. Prediction of cancer driver mutations in protein kinases. *Cancer research* **68**, 1675–82 (2008).

190. Teng, D., III, W. P., Hogan, J. & Baumgard, M. Human Mitogen-activated Protein Kinase Kinase 4 as a Candidate Tumor Suppressor. *Cancer research* 4177–4182 (1997).

191. Reese, M. G. *et al.* A standard variation file format for human genome sequences. *Genome biology* **11**, R88 (2010).

192. Flicek, P. *et al.* Ensembl 2012. *Nucleic acids research* **40**, D84–90 (2012).

193. Moore, B., Fan, G. & Eilbeck, K. SOBA: sequence ontology bioinformatics analysis. *Nucleic acids research* **38**, W161–4 (2010).

194. Medina, I. *et al.* VARIANT: Command Line, Web service and Web interface for fast and accurate functional characterization of variants found by Next-Generation Sequencing. *Nucleic acids research* **40**, W54–8 (2012).

195. Gel Moreno, B., Jenkinson, A. M., Jimenez, R. C., Messeguer Peypoch, X. & Hermjakob, H. easyDAS: automatic creation of DAS servers. *BMC bioinformatics* **12**, 23 (2011).

196. Lodish, H. F. & Kong, N. The Secretory Pathway Is Normal in Dithiothreitol-treated Cells , But Disulfide-bonded Proteins Are Reduced and Reversibly Retained in the Endoplasmic Reticulum. *The Journal of biological chemistry* **268**, 20598–20605 (1993).

197. Walter, P. & Ron, D. The unfolded protein response: from stress pathway to homeostatic regulation. *Science (New York, N.Y.)* **334**, 1081–6 (2011).
198. Fernandez, P. M. *et al.* Overexpression of the glucose-regulated stress gene GRP78 in malignant but not benign human breast lesions. *Breast cancer research and treatment* **59**, 15–26 (2000).
199. Shuda, M. *et al.* Activation of the ATF6 , XBP1 and grp78 genes in human hepatocellular carcinoma : a possible involvement of the ER stress pathway in hepatocarcinogenesis. *Journal of Hepatology* **38**, 605–614 (2003).
200. Song, M. S. *et al.* Induction of Glucose-regulated Protein 78 by Chronic Hypoxia in Human Gastric Tumor Cells through a Protein Kinase C- ε / ERK / AP-1 Signaling Cascade Induction of Glucose-regulated Protein 78 by Chronic Hypoxia in Human Gastric. *Cancer Research* 8322–8330 (2001).
201. Boelens, J., Lust, S., Offner, F., Bracke, M. E. & Vanhoecke, B. W. The Endoplasmic Reticulum : A Target for New Anticancer Drugs. *In vivo* **21**, 215–226 (2007).
202. Geiler-Samerotte, K. & Dion, M. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proceedings of the National Academy of Sciences of the United States of America* (2011).doi:10.1073/pnas.1017570108/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1017570108
203. Fonseca, S. G., Gromada, J. & Urano, F. Endoplasmic reticulum stress and pancreatic β-cell death. *Trends in endocrinology and metabolism: TEM* **22**, 266–74 (2011).
204. Ng, D. P. & Deber, C. M. Modulation of the Oligomerization of Myelin Proteolipid Protein by Transmembrane Helix Interaction Motifs. *Biochemistry* **49**, 6896–6902 (2010).
205. He, L., Shobnam, N. & Hristova, K. Specific inhibition of a pathogenic receptor tyrosine kinase by its transmembrane domain. *Biochimica et biophysica acta* **1808**, 253–9 (2011).
206. Vidal, G. a, Clark, D. E., Marrero, L. & Jones, F. E. A constitutively active ERBB4/HER4 allele with enhanced transcriptional coactivation and cell-killing activities. *Oncogene* **26**, 462–6 (2007).
207. Pagant, S., Halliday, J. J., Kougentakis, C. & Miller, E. A. Intragenic Suppressing Mutations Correct the Folding and Intracellular Traffic of Misfolded Mutants of Yor1p , a Eukaryotic Drug Transporter. *Journal of Biological Chemistry* **285**, 36304–36314 (2010).
208. Partridge, A. W., Therien, A. G. & Deber, C. M. Missense mutations in transmembrane domains of proteins: phenotypic propensity of polar residues for human disease. *Proteins* **54**, 648–56 (2004).
209. Wang, X. *et al.* Familial CJD associated PrP mutants within transmembrane region induced Ctm-PrP retention in ER and triggered apoptosis by ER stress in SH-SY5Y cells. *PloS one* **6**, e14602 (2011).
210. Liu, Y. & Ye, Y. Proteostasis regulation at the endoplasmic reticulum: a new perturbation site for targeted cancer therapy. *Cell research* **21**, 867–83 (2011).
211. Lee, A.-H., Iwakoshi, N. N., Anderson, K. C. & Glimcher, L. H. Proteasome inhibitors disrupt the unfolded protein response in myeloma cells.

Proceedings of the National Academy of Sciences of the United States of America **100**, 9946–51 (2003).

212. Nawrocki, S. T. *et al.* Bortezomib inhibits PKR-like endoplasmic reticulum (ER) kinase and induces apoptosis via ER stress in human pancreatic cancer cells. *Cancer research* **65**, 11510–9 (2005).
213. Carew, J. S. *et al.* Targeting endoplasmic reticulum protein transport: a novel strategy to kill malignant B cells and overcome fludarabine resistance in CLL. *Blood* **107**, 222–31 (2006).
214. Simon, M. & Hancock, J. M. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome biology* **10**, R59 (2009).
215. Uversky, V. N. Intrinsically disordered proteins from A to Z. *The international journal of biochemistry & cell biology* **43**, 1090–103 (2011).
216. Uversky, V. N. The mysterious unfoldome: structureless, underappreciated, yet vital part of any given proteome. *Journal of biomedicine & biotechnology* **2010**, 568068 (2010).
217. Lobley, A., Swindells, M. B., Orengo, C. A. & Jones, D. T. Inferring Function Using Patterns of Native Disorder in Proteins. *Proteins* **3**, (2007).
218. Dunker, a K., Cortese, M. S., Romero, P., Iakoucheva, L. M. & Uversky, V. N. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *The FEBS journal* **272**, 5129–48 (2005).
219. Haynes, C. *et al.* Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS computational biology* **2**, e100 (2006).
220. Edwards, Y. J. K., Lobley, A. E., Pentony, M. M. & Jones, D. T. Insights into the regulation of intrinsically disordered proteins in the human proteome by analyzing sequence and gene expression data. *Genome biology* **10**, R50 (2009).
221. Gsponer, J., Futschik, M., Teichmann, S. & Babu, M. M. Tight Regulation of Unstructured Proteins: From Transcript Synthesis to Protein Degradation. *Science* **322**, 1365–1368 (2008).
222. Hegyi, H., Buday, L. & Tompa, P. Intrinsic structural disorder confers cellular viability on oncogenic fusion proteins. *PLoS computational biology* **5**, e1000552 (2009).
223. Uversky, V. N. *et al.* Unfoldomics of human diseases : linking protein intrinsic disorder with diseases. *BMC Genomics* **17**, 1–17 (2009).
224. Babu, M. M., Van der Lee, R., De Groot, N. S. & Gsponer, J. Intrinsically disordered proteins: regulation and disease. *Current opinion in structural biology* **21**, 432–40 (2011).
225. Uversky, V. N. Amyloidogenesis of Natively Unfolded Proteins. *Current Alzheimer Research* **5**, 260–287 (2008).
226. Tompa, P. Structural disorder in amyloid fibrils: its implication in dynamic interactions of proteins. *The FEBS journal* **276**, 5406–15 (2009).
227. Vacic, V. & Iakoucheva, L. M. Disease mutations in disordered regions - exception to the rule? *Molecular Biosystems* **8**, 27–32 (2012).
228. Brown, C. J., Johnson, A. K. & Daughdrill, G. W. Comparing models of evolution for ordered and disordered proteins. *Molecular biology and evolution* **27**, 609–21 (2010).

229. Mosca, R., Pache, R. a & Aloy, P. The Role of Structural Disorder in the Rewiring of Protein Interactions through Evolution. *Molecular & cellular proteomics : MCP* **11**, M111.014969 (2012).

230. Razumovskaya, E., Masson, K., Khan, R., Bengtsson, S. & Rönnstrand, L. Oncogenic Flt3 receptors display different specificity and kinetics of autophosphorylation. *Experimental hematology* **37**, 979–89 (2009).

231. Yim, E.-K. *et al.* Rak functions as a tumor suppressor by regulating PTEN protein stability and function. *Cancer cell* **15**, 304–14 (2009).

232. Hu, Y., Liu, Y., Jung, J., Dunker, A. K. & Wang, Y. Changes in predicted protein disorder tendency may contribute to disease risk. *BMC Genomics* **12**, S2 (2011).

233. Gundem, G. *et al.* IntOGen : integration and data mining of multidimensional oncogenomic data IntOGen is a framework that addresses a. *Nature Methods* **7**, 92–93 (2010).

234. Furney, S. J., Calvo, B., Larrañaga, P., Lozano, J. a & Lopez-Bigas, N. Prioritization of candidate cancer genes--an aid to oncogenomic studies. *Nucleic acids research* **36**, e115 (2008).

235. Yue, P. *et al.* Inferring the functional effects of mutation through clusters of mutations in homologous proteins. *Human mutation* **31**, 264–71 (2010).

236. Peterson, T. a, Nehrt, N. L., Park, D. & Kann, M. G. Incorporating molecular and functional context into the analysis and prioritization of human variants associated with cancer. *Journal of the American Medical Informatics Association : JAMIA* **19**, 275–83 (2012).

237. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS genetics* **5**, e1000534 (2009).

238. Giot, L. *et al.* A protein interaction map of Drosophila melanogaster. *Science (New York, N.Y.)* **302**, 1727–36 (2003).

239. Han, J., Bertin, N., Hao, T. & Goldberg, D. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **430**, (2004).

240. Erlich, Y. *et al.* Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis. *Genome research* **21**, 658–64 (2011).

241. Akula, N. *et al.* A network-based approach to prioritize results from genome-wide association studies. *PloS one* **6**, e24220 (2011).

242. Jia, P., Zheng, S., Long, J., Zheng, W. & Zhao, Z. dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics (Oxford, England)* **27**, 95–102 (2011).

243. García-Alonso, L. *et al.* Discovering the hidden sub-network component in a ranked list of genes or proteins derived from genomic experiments. *Nucleic acids research* 1–13 (2012).doi:10.1093/nar/gks699

244. Ma, H. *et al.* The Edinburgh human metabolic network reconstruction and its functional analysis. *Molecular systems biology* **3**, 135 (2007).

245. Sun, J. & Zhao, Z. A comparative study of cancer proteins in the human protein-protein interaction network. *BMC genomics* **11 Suppl 3**, S5 (2010).

246. Hernández, P. *et al.* Evidence for systems-level molecular mechanisms of tumorigenesis. *BMC genomics* **8**, 185 (2007).

247. Yao, C. *et al.* Multi-level reproducibility of signature hubs in human interactome for breast cancer metastasis. *BMC systems biology* **4**, 151 (2010).

248. Fleming, N. *et al. SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer. Cancer research* (2012).doi:10.1158/0008-5472.CAN-12-2706

249. Yandell, M. *et al.* Genome-wide analysis of human disease alleles reveals that their locations are correlated in paralogous proteins. *PLoS computational biology* **4**, e1000218 (2008).

250. Qiu, P., Gentles, A. J. & Plevritis, S. K. Discovering biological progression underlying microarray samples. *PLoS computational biology* **7**, e1001123 (2011).

251. Chen, Y., Jiang, T. & Jiang, R. Uncover disease genes by maximizing information flow in the phenome-interactome network. *Bioinformatics (Oxford, England)* **27**, i167–76 (2011).

252. Wu, X., Liu, Q. & Jiang, R. Align human interactome with phenome to identify causative genes and networks underlying disease families. *Bioinformatics* **25**, 98–104 (2009).

253. Winter, C. *et al.* Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS computational biology* **8**, e1002511 (2012).

254. Hwang, T., Zhang, W., Xie, M., Liu, J. & Kuang, R. Inferring disease and gene set associations with rank coherence in networks. *Bioinformatics (Oxford, England)* **27**, 2692–9 (2011).

255. Franceschini, a. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research* 1–8 (2012).doi:10.1093/nar/gks1094

256. Guan, Y. *et al.* Tissue-specific functional networks for prioritizing phenotype and disease genes. *PLoS computational biology* **8**, e1002694 (2012).

257. Börnigen, D. *et al.* An unbiased evaluation of gene prioritization tools. *Bioinformatics (Oxford, England)* **28**, 3081–3088 (2012).

258. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic acids research* 1–10 (2012).doi:10.1093/nar/gks743

259. Hudson, T., Anderson, W., Aretz, A. & Barker, A. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).

8. Resum en català

8.1. Introducció

L'augment de l'ús de les tecnologies d'abast genòmic, com ara la seqüenciació massiva, microarrays o les diverses tècniques de proteòmica, ha provocat una explosió de dades biomèdiques durant els darrers anys. En aquest context l'ús d'aproximacions computacionals per emmagatzemar, integrar i analitzar aquesta informació biològica és una prioritat.

Els mètodes bioinformàtics s'han aplicat amb èxit a nombrosos problemes derivats d'aquest augment exponencial de la informació, com per exemple la integració de dades experimentals provinents de diferents fonts d'informació o la priorització de llistes de gens candidats. Dins aquest últim camp hi ha dues tècniques que destaquen pel seu extensiu ús dins la comunitat científica: l'anàlisi d'enriquiment i la teoria de xarxes.

L'anàlisi d'enriquiment consisteix en comparar les propietats d'un grup determinat de gens o proteïnes amb les d'un grup control que generalment és el genoma sencer. Utilitzant eines estadístiques és possible identificar propietats que estiguin enriquides o empobrides en el grup d'interès respecte el grup control, el que pot permetre identificar aspectes interessants de la biologia d'aquest grup de gens. Per tal de poder explotar al màxim aquesta tècnica és necessari que les anotacions estiguin normalitzades, com a mínim, amb un vocabulari controlat, ja que sino no és possible fer una correcta comparació de la distribució d'aquestes anotacions entre els dos grups (podríem perdre anotacions degut a la seva incorrecta anotació). No obstant, és fins i tot més recomanable l'ús d'ontologies per analitzar les anotacions de les propietats enlloc del de vocabularis controlats, ja que els algorismes desenvolupats recentment permeten explotar les relacions entre els termes que s'inclouen en aquest. Això generalment permet augmentar el poder estadístic, el que és rellevant quan es fan anàlisi a escala genòmica.

La teoria de xarxes ve de les matemàtiques i tracta de descriure les propietats d'uns objectes matemàtics que s'anomenen xarxes. Una xarxa es descriu formalment com un conjunt de nodes que estan connectats entre si per arestes. Els nodes poden representar qualsevol entitat que ens interessi i les arestes generalment representen les relacions entre ells. La gran flexibilitat i generalització de la teoria de xarxes s'exemplifica pel fet que s'ha emprat amb èxit a nombrosos camps de la ciència, com ara l'economia, l'ecologia, les ciències socials o la semàntica. En el camp de la biologia molecular el seu ús es va estendre arran de la descripció de les primeres xarxes d'interaccions de proteïnes. D'aleshores ençà s'han emprat diversos algorismes derivats de la teoria de xarxes amb finalitats tant diverses com la descripció de les propietats de xarxes de regulació gènica, la predicció de la funció de proteïnes o la predicció de gens associats a diverses malalties.

8.2. Objectius

En aquesta tesi ens hem plantejat explorar l’ús de tècniques bioinformàtiques per tal d’identificar nous gens i mutacions associats a malaltia i els seus possibles mecanismes moleculars. Els objectius concrets per tal d’acostar-nos a aquesta meta són:

1. Desenvolupament d’una base de dades relacional amb informació sobre mutacions i gens associats a malalties i les propietats de les seves proteïnes

2. Implementar i utilitzar l’anàlisi d’enriquiment per tal d’identificar propietats de proteïnes que s’alteren específicament en determinades malalties

3. Ús d’algorismes i conceptes de la teoria de xarxes per extraure informació de xarxes biològiques i predir nous gens associats a malalties

8.3. Materials i mètodes

Per tal de poder dur a terme l'anàlisi d'enriquiment, és necessari disposar d'una base de dades amb informació de gens i mutacions associats a diverses malalties així com d'anotacions sobre diverses propietats moleculars dels mateixos. Tanmateix, tal com hem explicat a la introducció és important que aquestes anotacions estiguin normalitzades amb vocabularis controlats o ontologies.

Donat que cap de les bases de dades públiques existents amb informació sobre gens i mutacions associats a malaltia reunien tots els requisits, hem decidit generar una base de dades nova. Les mutacions d'aquesta base de dades provenen bàsicament de 3 fonts: COSMIC, un catàleg de mutacions somàtiques en càncer, OMIM, la base de dades de referència sobre mutacions i malalties hereditàries i GAD, una base de dades amb informació sobre polimorfismes associats a malalties complexes. A més a més, hem extret d'altres repositoris públics informació referent a la funció de les proteïnes mutades així com informació sobre la zona alterada: estructura secundària, modificacions post-traduccionals, possible activitat catalítica, dominis etc. A aquest efecte també hem utilitzat programari dedicat per tal de fer prediccions sobre aquestes regions. Totes les anotacions referents a les malalties associades a les mutacions s'han fet emprant l'ontologia de malalties (DO), aquelles referents a la funció de proteïnes amb l'ontologia de gens (GO), les relacionades amb dominis estructurals amb PFAM i les de propietats de la seqüència biològica amb l'ontologia de seqüències (SO).

L'anàlisi d'enriquiment s'ha dut a terme comparant les propietats dels gens associats a cadascun dels diferents termes de la DO amb les de la resta de gens associats a malaltia. Aquest és un punt important del nostre disseny experimental, ja que la majoria d'experiments que utilitzen l'anàlisi d'enriquiment comparen les propietats de les mutacions associades a malaltia amb aquelles propietats del genoma sencer, pel que s'identifiquen aquelles propietats que donen caràcter patogènic. Nosaltres, en canvi, al comparar propietats de mutacions associades a una malaltia en concret amb la resta de mutacions associades a malaltia hem pogut identificar propietats més específiques (allò que fa que una mutació provoqui càncer i no Alzheimer, per exemple). Per tal de simplificar la interpretació posterior dels resultats, hem decidit implementar l'algorisme "elim", descrit per Alexa i altres. Aquest algorisme descarta aquelles anotacions associades a gens que mostren un enriquiment o empobriment estadísticament significatiu per als posteriors termes analitzats. Això permet que només es donin com a significatives aquelles associacions més específiques (i per tant amb més informació) i evita que l'anàlisi d'enriquiment doni associacions involucrant termes massa genèrics (i per tant amb menys informació).

Pel que fa a l'ús d'algorismes derivats de la teoria de xarxes per a predir nous gens associats a malaltia, hem comparat l'eficiència de 5 algorismes diferents, en 4 xarxes biològiques alhora de predir 5 malalties diferents: càncer, càncer colorectal, diabetis, malalties neurodegeneratives i malalties hereditàries. Els algorismes són 3 variants del comptatge de veïns (DN, de l'acrònim en anglès), la difusió de nucli (DK, de l'acrònim en anglès) i el passejador a l'atzar (RWR de l'acrònim en anglès també). El funcionament de cada combinació malaltia/xarxa/algorisme s'ha mesurat utilitzant l'AUC, que significa l'àrea sota la corba ROC.

El comptatge de veïns és un algorisme conceptualment molt intuïtiu que consisteix en ordenar la llista de gens candidats a estar associats amb la malaltia d'interès en funció del nombre de veïns a la xarxa que es sap que estan associats a la malaltia. El nombre de veïns es compta fins a una distància predefinida "d". Nosaltres hem utilitzat 3 distàncies diferents: 1, 2 i 3. Els algorismes DK i RWR pertanyen a la categoria d'algorismes de difusió en xarxes. Ambdós tracten de simular la difusió d'informació al llarg de la xarxa, pel que són capaços d'explotar-ne la topologia i, d'aquesta manera, extraure'n més informació. El DK es basa en el càlcul infinitesimal i utilitza matrius que representen la xarxa per simular com la informació que surt des dels gens associats a la malaltia d'interès es propaga fins als gens candidats. Com més informació continguin els últims després de la simulació, més probable és que estiguin associats a la malaltia. El RWR simula infinits caminadors que van d'un node a un altre a l'atzar. Aquests caminadors surten des d'algun dels gens associats a la malaltia i vas caminant a l'atzar un temps infinit. Al final els gens candidats s'ordenen en funció de la probabilitat que el caminador es trobi en ells quan el temps tendeix a infinit.

Pel que fa a les xarxes biològiques, n'hem construït 4, cadascuna d'elles representant un tipus de relació biològica diferent entre els gens/proteïnes. La primera representa interaccions físiques entre proteïnes. Es tracta d'un tipus de xarxa àmpliament emprat en biologia computacional tant per a predir gens associats a malaltia com per a altres finalitats. Les dades d'aquesta xarxa provenen de la "Base de dades de Proteïnes de Referència Humanes" (HPRD de l'anglès). La segona connecta gens que, d'acord amb la informació a ENSEMBL, pertanyen a la mateixa família (són paràlegs). Una altra de les xarxes s'ha construït utilitzant informació de Reactome. En aquesta xarxa dues proteïnes estan connectades si formen part de la mateixa via biològica (metabòlica, de senyalització etc.). Finalment, la última de les xarxes s'ha creat utilitzant dades d'expressió en teixit humà sa dels diferents gens. Aquestes dades provenen de BioGPS i ens han permès connectar aquells gens que mostren un elevat grau de coexpressió (un coeficient d'R al quadrat superior a 0.7).

Una de les estratègies habituals per tal de millorar l'eficiència de les xarxes biològiques consisteix en combinar-les amb altres fonts d'informació. Per tal d'avaluar si aquesta és una bona estratègia hem intentat combinar la informació de les 4 xarxes de diferents maneres: utilitzant un classificador Bayesià, sobreposant-les, sumant-les directament o fent-ne una suma ponderada, els pesos de la qual s'han optimitzat amb un algorisme anomenat recuita simulada *(simulated annealing*).

8.4. Resultats

La nova base de dades relacional sobre gens i mutacions associats a malaltia conté 2716 gens, 9276 mutacions i 1195 termes de DO diferents. A més a més, hi ha almenys una propietat bioquímica per a 8529 mutacions. Pel que fa al seu origen, les mutacions venen en la seva majoria de COSMIC (6056), mentre que de OMIM i GAD n'hi ha un total de 3307.

Per tal d'estar segurs de no haver introduït cap error en el procés d'extracció de les mutacions hem representat el nombre de mutacions per gen i per malaltia, així com el nombre de malalties per mutació. Les tres representacions segueixen una escala logarítmica, el que significa que, per exemple, la majoria de malalties estan associades a una sola mutació mentre que hi ha molt poques malalties associades a moltes mutacions (a més a més es tracta sempre de neoplàsies, pel que la observació té sentit biològic). Aquests resultats concorden amb observacions prèvies d'altres grups.

Després de comprovar que les dades que hem extret i anotat amb les diverses ontologies tenen sentit i no estan esbiaixades hem dut a terme l'anàlisi d'enriquiment tal com s'ha explicat l'apartat de material i mètodes.

Les associacions entre termes de l'ontologia de malalties (DO) i la de funcions gèniques (GO) donen una bona primera idea del funcionament de l'algorisme, ja que com s'ha mencionat prèviament, la GO s'ha utilitzat de manera extensiva en anàlisi similars. Hem obtingut 3199 parelles de termes de la DO i de la GO que mostren una associació estadísticament significativa, ja sigui per una enriquiment o per empobriment de mutacions. La gran majoria de parelles tenen sentit des del punt de vista biològic. Així per exemple observem que mutacions en gens anotats amb la funció biològica "coagulació sanguínia" (GO:0007596) tendeixen a causar "malalties de coagulació sanguínia" (DOID:1247).

D'altra banda, també hem observat termes de GO que estan enriquits en una determinada malaltia i empobrits en una altra de relacionada. Un exemple seria el terme funcional "transducció de senyal", que mostra un fort enriquiment en mutacions relacionades amb càncer de pell, limfoma, sarcoma o càncer en general i, al mateix temps, un empobriment en mutacions associades a càncer de ronyó.

Un fenomen similar passa amb els dominis de proteïnes. Entre les 83 associacions estadísticament significatives entre dominis de proteïnes i malalties n'hi ha algunes que s'havien descrit prèviament i altres que mostren un comportament desigual en malalties, a principi, similars. Dins el primer grup hi trobem varies associacions entre càncer i dominis com ara Ras, Miro o PIP-3,4K. El representant més interessant del segon grup és el domini quinasa. Aquest domini s'ha associat prèviament, i nombroses vegades, amb càncer. Nosaltres hem pogut reproduir aquesta associació, no obstant també hem trobat que hi ha menys mutacions de les esperades en aquest domini que causin càncer de mama.

Per tal d'analitzar amb detall aquest fenomen hem separat les mutacions en funció dels diferents subdominis del domini quinasa on estan localitzades. Estudis previs havien demostrat que les mutacions en aquest domini que estan associades a càncer tendeixen a estar localitzades en els subdominis I i VIII del domini quinasa. Nosaltres hem pogut confirmar aquestes associacions per les mutacions que causen càncer, i a més a més hem observat un enriquiment en mutacions causals de càcner de mama en el subdomini V, que no havia estat associat prèviament a cap tipus de malaltia en particular.

Finalment, hem trobat 82 associacions entre termes de DO i SO. La majoria d'aquestes associacions involucren terms de SO que són força genèrics, pel que el seu anàlisi és complicat. No obstant, desprès de revisar manualment totes les associacions n'hem trobat 6 involucren càncer i algun terme de SO suficientment informatiu per a formular hipòtesis sobre les raons biològiques que provoquen aquesta associació. De les 6 associacions 3 ho són per enriquiment i 3 per empobriment de mutacions. Els 3 termes de SO enriquits en mutacions canceroses són "regions intrínsecament desestructurades", "regions riques en serina" i "regions de composició esbiaixada". D'altra banda, els termes SO empobrits en mutacions de càncer són "ponts disulfur", "pèptids de localització" i "regions transmembrana". Per tal d'assegurar-nos que les associacions són sòlides hem dut a terme dos controls addicionals, un de tècnic i un altre amb un nou set de mutacions de càncer que hem obtingut del "Atles Genòmic del Càncer" (TCGA). El primer control, de caràcter tècnic, ens ha permès confirmar les 3 associacions d'empobriment, però no 2 de les 3 associacions per enriquiment: regions de composició esbiaixada i regions riques en serina (l'associació amb les regions intrínsecament desestructurades no la vam poder comprovar en aquest control per raons conceptuals). El control amb el set de dades nou ha confirmat totes les associacions amb l'excepció de la que involucra les regions riques en serina.

Un cop analitzats els resultats de l'anàlisi d'enriquiment, hem dut a terme l'estudi de les xarxes biològiques. Si bé hi ha grups que suggereixen que els mètodes de difusió són millors que mètodes més simples com ara el comptatge de veïns, els nostres resultats no apunten en aquesta direcció, almenys com a norma general. En la majoria de xarxes no hem observat aquesta eficiència superior i, de fet, en la xarxa de paràlegs, els resultats de les diferents malalties són iguals independentment de l'algorisme emprat. La única xarxa que sembla que millora clarament amb els mètodes de difusió és la que integra les diverses vies cel·lulars.

Pel que fa al comportament de les diferents xarxes, és important destacar que en poques ocasions s'empren altres xarxes biològiques que no siguin la d'interacció proteïna-proteïna per si soles (si que es s'empren en combinació amb aquesta). No obstant, els nostres resultats suggereixen que aquestes xarxes poden ser tant bones o fins i tot millors, que la d'interaccions entre proteïnes alhora de predir gens relacionats amb malalties. Per exemple, la xarxa de paràlegs es superior, de manera estadísticament significativa, a la d'interaccions alhora de predir gens relacionats amb càncer.

A continuació, hem intentat combinar les xarxes de diverses maneres per veure si aconseguíem millorar-ne la capacitat predictiva. El primer resultat destacable és que no hi ha diferències entre combinar les xarxes per adició o per juxtaposició. Això probablement es deu a que les xarxes comparteixen poques arestes, pel que no hi ha diferència entre que aquestes arestes tinguin un pes igual o proporcional al nombre de xarxes en les que apareixen.

Si es comparen els resultats obtinguts per les xarxes soles o per la seva combinació alhora de predir les diferents malalties, s'observa que només en una malaltia, diabetis, hem pogut obtenir una combinació de xarxes que obtingués valors d'AUC que fossin millors de manera estadísticament significative que els de la millor xarxa individual en aquella malaltia. No obstant, sí que és possible obtenir combinacions de xarxes que funcionin tant bé com la millor xarxa individual. Donat que la combinació de xarxes inclou més gens relacionats amb la malaltia que les xarxes individuals, es pot dir que a igualtat de valors d'AUC, la combinació funciona millor.

Moltes de les publicacions que intenten combinar diferents fonts d'informació per tal de predir nous gens relacionats amb malalties descriuen un cert grau de correlació entre la quantitat d'informació emprada per fer la predicció i l'eficiència de la mateixa. Per tal d'estudiar aquest fenomen hem descarregat una xarxa funcional descrita per Marcotte i altres que inclou 21 fonts d'informació diferents i n'hem comparat el poder predictiu amb cadascuna de les nostres combinacions de xarxes en les diferents malalties. Els nostres resultats indiquen que no hi ha correlació entre la quantitat d'informació emprada per crear una xarxa de proteïnes i el seu poder predictiu, ja que xarxes resultants de la combinació de 2 o 3 xarxes biològiques diferents tenen tant o més poder predictiu que la xarxa de Marcotte o la combinació de les 4 xarxes.

La informació provinent de les xarxes no només es pot combinar *a priori*, sino que també es poden combinar els resultats obtinguts amb els diferents algorismes en les diferents xarxes *a posteriori*. Recentment s'han descrit diverses aproximacions en aquesta línia, com ara l'ús d'estadística d'ordre o classificadors Bayesians entre altres. Per tal de veure si és millor combinar la informació de les xarxes abans o després de còrrer els algorismes hem comparat l'eficàcia d'un classificador Bayesià que combina les puntuacions obtingudes en les diferents xarxes després d'utilitzar el RWR amb la de còrrer el RWR amb la xarxa resultant de la suma de les 4 xarxes individuals i en tots els casos el resultat ha estat a favor de la combinació de xarxes *a priori*.

Finalment hem tractat d'avaluar una de les combinacions de xarxes més prometedores, la derivada de la suma de HPRD, Paralogia i Reactome, utilitzant un set de dades extern al que havíem emprat per entrenar el mètode. Per a fer-ho ens hem descarregat el l'últim resum de mutacions en càncer colorectal disponible a COSMIC i hem escollit tots aquells gens que s'ha trobat mutats en, almenys, 15 mostres tumorals. Aquest primer filtre ens ha donat una llista de 482 gens que hem classificat com a causals. A continuació hem aplicat un segon filtre que ha consistit en extraure de la llista de 482 tots aquells gens que, d'acord amb el nostre set de dades ja estaven associats a (I) càncer colorectal i (II) qualsevol tipus of càncer. Aquest segon filtre ens ha donat una llista de 353 i 252 gens respectivament.

Hem utilitzat tots els gens de la nostra base de dades que estan associats a càncer colorectal o a càncer en general per fer la predicció dels 353 o 252 nous gens. Hem observat que som capaços de predir aquests gens amb un valor d'AUC de 0.65 i que no hi ha diferència entre utilitzar tots els nostres gens de càncer o només els de càncer colorectal alhora de predir els nous gens associats a càncer colorectal, el que suggereix que el nostre mètode permet utilitzar informació de malalties similars per a fer les prediccions.

A més a més, hem emprat un altre mètode d'avaluació que ha consistit en crear, per a cadascun dels nous gens de càncer colorectal grups de 10 o 25 gens, formats pel propi gen de càncer i els 9 o 24 gens més propers en el genoma. A continuació hem ordenat els diferents grups en funció de la puntuació obtinguda pel nostre mètode i hem observat que hi ha un enriquiment dels gens associats a càncer en les primeres posicions. En conclusió, tant els valors d'AUC com els de l'order dels grups de gens del genoma indiquen que el nostre mètode es capaç de predir gens coneguts associats a càncer colorectal.

8.5. Discussió

El desenvolupament de la base de dades relacionals ha estat complex i té unes certes limitacions que s'han de tenir en compte alhora d'avaluar els resultats. En priemr lloc no hem pogut extraure totes le smutacions disponibles, degut a problemes amb els repositoris originals (per exemple GAD no descriu les mutacions de manera sistemàtica), tant com per problemes de limitacions de l'ontologia (no totes les malalties estan presents a l'ontologia). Haguès sigut millor disposar de més dades referents a mutacions associades a malalties complexes, no obstant aquestes són complicades d'obtenir. No obstant, creiem que el nostre set de dades és prou representatiu d'almenys dos grans grups de malalties: càncers i malalties hereditàries.

Amb relació a l'anàlisi d'enriquiment, dir que hem detectat una sèrie de funcions biològiques, propietats de seqüència i de dominis que mostren associacions discordants entre fenotips similars. Aquestes associacions són les més interessants del nostre mètode, ja que només poden ser detectades amb el nostre disseny experimental que compara mutacions associades a diferents malalties entre si. A més a més es podrien emprar en el futur per tal d'entrenar un predictor de mutacions patogèniques.

Pel que fa a les associacions de càncer amb termes de SO, hem pogut generar una sèrie d'hipotesi sobre el seu possible origen biològic. Els tres termes empobrits en mutacions cancerígenes (ponts disulfur, pèptid senyal i domini transmembrana) creiem que estan relacionats amb una via cel·lular que s'ha implicat prèviament amb càncer, la resposta a proteïnes malplegades. Aquesta via pot tenir efectes apoptòtics o antiapoptòtics i s'activa quan detecta una acumulació de proteïnes malplegades al reticle endoplasmàtic. Les cèl·lules tumorals tenen un nivell d'activació basal d'aquesta via superior al normal degut a les condicions d'hipòxia i manca de nutrients. La mutació en qualsevol de les propietats de seqüència mencionades provoca l'acumulació de proteïnes malplegades dins el reticle i la sobreactivació de la via, el que finalment porta a la cèl·lula a morir per apoptosi. És per això que observem menys mutacions en aquestes propietats de seqüència de les esperades en càncer.

En referència a les 3 propietats enriquides en càncer, creiem que es poden analitzar en conjunt ja que les nostres dades suggereixen que les 3 estan interrelacionades i no són independents. Les regions intrínsicament desestructurades s'han relacionat prèviament amb càncer per diverses raons, tot i que la principal es sospita que és el seu paper en les interaccions proteïna-proteïna. És possible que les mutacions en aquestes regions alterin interaccions importants per al desenvolupament del càncer sense destruir la proteïna.

Respecte l'ús de xarxes biològiques, explicar que hem demostrat que xarxes biològiques derivades d'altres fonts que no siguin interaccions proteïna-proteïna es poden utilitzar amb èxit per a predir nous gens relacionats amb malaltia. A més a més no hem observat que els mètodes de difusió siguin superiors a mètodes simples com el comptatge de veïns, almenys de manera estadísticament significativa.

D'acord amb les nostres dades, sembla que alhora de predir gens relacionats amb malaltia és millor combinar la informació de les xarxes biològiques sumant-les que utilitzant mètodes estadístics per combinar-ne les puntuacions independents. A més a més és difícil combinar les xarxes de manera que s'obtingui un resultat millor que la millor xarxa independent.

Finalment, explicar que el nostre mètode és capaç de predir gens associats a càncer colorectal que provenen d'un set de dades independent de l'utilitzat per entrenar el mètode.

9. Supplementary Material

9.1. OCG tables

AA_features	Field	Type	Example
	Id_feature	Int(11)	50
	Description	Varchar(50)	Stretch
	Specification	Varchar(100)	Poly-Ala

AA_features_ontology	Field	Type	Example
	Id_aa_features_ontology	Int(11)	1083
	Id_feature	Int(11)	50
	Id_sequence_ontology	Varchar(30)	SO:60SO1

AA_features_source	Field	Type	Example
	Id_source	Int(11)	1
	Description	Varchar(50)	Netphos
	Status	Varchar(100)	NULL

Disease	Field	Type	Example
	Id_disease	Varchar(40)	DOID:162DOID1

Disease_protein	Field	Type	Example
	Id_disease_protein	Int(11)	1
	id_disease	Varchar(40)	DOID:3451DOID1
	Id_gene	Int(11)	1
	Id_mutation	Int(11)	1
	Id_reference	Int(3)	2

DO_GO_pairs	Field	Type	Example
	Id_pair_do_go	Int(11)	1
	Id_do	Varchar(30)	DOID:8499DOID1
	Id_go	Varchar(30)	GO:0007601
	P_val	double	2.227675 e-13
	Num_muts_do	Varchar(20)	14 / 18
	Num_muts_go	Varchar(20)	14 / 258
	Num_genes_do	Varchar(20)	6 / 8
	Num_genes_go	Varchar(20)	6 / 74
	Id_source	Int(3)	4
	Odds_ratio	Float	128.665
	95_perc_conf_int	Varchar(30)	40.06401 547.8941

DO_PFAM_pairs	Field	Type	Example
	Id_pair_do_pfam	Int(11)	1
	Id_do	Varchar(30)	DOID:4907DOID1
	Id_pfam	Varchar(30)	PF00069
	P_val	double	0.01656962
	Num_muts_do	Varchar(20)	16 / 57
	Num_muts_pfam	Varchar(20)	16 / 640
	Num_genes_do	Varchar(20)	3 / 9

	Num_genes_pfam	Varchar(20)	3 / 163
	Id_source	Int(3)	4
	Odds_ratio	Float	5.37326
	95_perc_conf_int	Varchar(30)	2.798045 9.850881

	Field	Type	Example
DO_SO_pairs	Id_pair_do_pfam	Int(11)	1
	Id_do	Varchar(30)	DOID:5041DOID1
	Id_so	Varchar(30)	SO:0001078SO1
	P_val	double	0.004729497
	Num_muts_do	Varchar(20)	36 / 73
	Num_muts_so	Varchar(20)	36 / 2044
	Num_genes_do	Varchar(20)	4 / 13
	Num_genes_so	Varchar(20)	4 / 385
	Id_source	Int(3)	4
	Odds_ratio	Float	3.4855
	95_perc_conf_int	Varchar(30)	2.134675 5.687335

	Field	Type	Example
DO_SO_pairs	Id_pair_do_pfam	Int(11)	1
	Id_do	Varchar(30)	DOID:5041DOID1
	Id_so	Varchar(30)	SO:0001078SO1
	P_val	double	0.004729497
	Num_muts_do	Varchar(20)	36 / 73
	Num_muts_so	Varchar(20)	36 / 2044
	Num_genes_do	Varchar(20)	4 / 13
	Num_genes_so	Varchar(20)	4 / 385
	Id_source	Int(3)	4
	Odds_ratio	Float	3.4855
	95_perc_conf_int	Varchar(30)	2.134675 5.687335

	Field	Type	Example
ENSEMBL_protein	Id_protein_ocg	Int(11)	3
	Ensembl_prot_id	Varchar(30)	ENSP00000388246
	Id_gene	Int(11)	2

	Field	Type	Example
ENSEMBL_prot_features	Id_ensembl_prot_feat	Int(11)	1
	Id_prot_OCG	Int(11)	1
	Id_feature	Int(11)	33
	Start	Int(6)	655
	Stop	Int(6)	661
	Id_source	Int(11)	3
	Score	Float	0

	Field	Type	Example
Gene	Id_gene	Int(11)	1
	Gene_name	Varchar(240)	BRAF
	ENSEMBL_id	Varchar(20)	ENSG00000157764

	Field	Type	Example
GO_terms	Id_go_term	Int(11)	1
	Go_term	Varchar(30)	GO:0005737

	Field	Type	Example
Info_reference_dict	Id_reference	Int(3)	1
	Reference_name	Varchar(30)	OMIM

	Field	Type	Example
Mutated_mut_sites	Id_mutated_mut_site	Int(11)	1
	Id_mut_site	Int(11)	2
	Id_mutation	Int(11)	1

	Field	Type	Example
Mutated_mut_site_features	Id_mut_features	Int(11)	29627
	Id_mutated_mut_site	Int(11)	27586
	Id_feature	Int(11)	3
	Id_source_feature	Int(11)	7
	Score	Float	0.876

	Field	Type	Example
Mutation_reference	Id_mutation_reference	Int(11)	1
	Id_mutation	Int(11)	1
	Id_reference	Int(3)	2

	Field	Type	Example
Mut_site	Id_mut_site	Int(11)	1
	Id_protein_ocg	Int(11)	1
	Coords	Int(11)	439
	Wt_allele	Varchar(10)	K

	Field	Type	Example
Protein_go_term	Id_go_protein	Int(11)	1
	Id_protein_ocg	Int(11)	1
	Id_go_term	Int(11)	1

	Field	Type	Example
Sequence_mutation	Id_mutation	Int(11)	1
	Id_gene	Int(11)	1
	Aa_mut_start	Mediumint(9)	439
	Aa_mut_allele_seq	Longtext	Q
	Aa_wt_allele_seq	Longtext	K

	Field	Type	Example
Mut_site_features	Id_mut_site_features	Int(11)	15247
	Id_mut_site	Int(11)	650
	Id_feature	Int(11)	970

	Id_source	Int(11)	6
	Score	Float	0.94
	Id_ensembl_prot_feat	Int(11)	NULL

9.2. CCBG tables

	Field	Type	Example
Disease	Id_disease	Varchar(40)	DOID:162DOID1

	Field	Type	Example
Disease_protein	Id_disease_protein	Int(11)	1
	id_disease	Varchar(40)	DOID:3451DOID1
	Id_gene	Int(11)	1

	Field	Type	Example
Disease_protein_reference	Id_disease_prot_reference	Int(11)	238
	Id_disease_protein	Int(11)	238
	Id_source	Int(3)	3
	Id_reference	Int(11)	12601293
	Score	Int(3)	1

	Field	Type	Example
Disease_source	Id_disease_source	Int(3)	2
	Source_name	Varchar(30)	OMIM

	Field	Type	Example
Gene	Id_gene	Int(11)	1
	Gene_symbol	Varchar(30)	ALDH1A1
	ENSEMBL_id_gene	Varchar(30)	ENSG00000165092
	ENSEMBL_version	Varchar(30)	homo_sapiens_core_62_37g

	Field	Type	Example
Genes_related	Id_genes_related	Int(11)	300855
	Id_gene1	Int(11)	16590
	Id_gene2	Int(11)	13666
	Id_relationship	Int(11)	6
	Relationship_value	Float	0.8954

	Field	Type	Example
Genes_related_pmed	Id_genes_related_pmed	Int(11)	3
	Id_genes_related_pmed	Int(11)	5
	Id_pubmed	Int(11)	9201297

	Field	Type	Example
Ontology_info	Id_ontology	Varchar(30	CCBGID:1
	Description	Varchar(30)	Genes physically interacting

	Field	Type	Example
Relationship	Id_relationship	Int(11)	2
	Id_ontology	Varchar(30)	CCBGID:4
	Id_source	Int(5)	2

	Field	Type	Example
Sources_info	Id_source	Int(5)	2
	Source_name	Varchar(30)	ENSEMBL

9.3. OLS tables

	Field	Type	Example
Term	Term_pk	Varchar(40)	DOID:162DOID1
	Term_name	Varchar(60)	Cancer
	Ontology_id	Int(11)	563002

	Field	Type	Example
Term_path	Term_path_pk	Int(11)	1
	Subject_term_pk	Varchar(40)	DOID:9256DOID1
	Predicate_term_pk	Varchar(40)	DOID:IS_ADOID1
	Object_term_pk	Varchar(40)	DOID:162DOID1
	Distance	Int(3)	1
	Ontology_id	Int(11)	563002

9.4. DO/SO associations

DO term	SO term	OR	P val
inborn errors metal metabolism	transmembrane_polypeptide_region	21.0929	3.88E-08
motor neuron disease	aminoacid_enriched_region	19.1216	4.85E-09
anterior horn cell disease	compositionally_biased_region_of_peptide	17.3757	1.75E-08
long QT syndrome	transmembrane_polypeptide_region	11.4346	6.51E-05
inborn errors renal tubular transport	transmembrane_polypeptide_region	11.4261	1.08E-06
Noonan syndrome	polypeptide_secondary_structure	8.90063	0.03460876
myopathy	coiled_coil	6.28292	0.002811855
muscle tissue disease	polypeptide_structural_motif	6.17192	0.003482194
congenital disorder	disulfide_bond	5.62758	0.000900182
severe combined immunodeficiency	polypeptide_secondary_structure	4.76066	0.01507244
disease of metabolism	protein_binding_site	4.41876	0.02999278
disease of metabolism	sequence_variant_causing_inactive_ligand_binding_site	4.41876	0.02999278
myopathy	transmembrane_polypeptide_region	4.38175	6.54E-09
muscle tissue disease	intramembrane_polypeptide_region	4.29636	1.15E-08
musculoskeletal system disease	disulfide_bond	3.97658	0.04667697
breast carcinoma	coiled_coil	3.97411	0.002519508
carcinoma	Ser-rich	3.94086	5.19E-07

intracranial neoplasm	polypeptide_secondary_structure	3.86152	8.60E-14
male genital cancer	alpha_helix	3.51659	0.0189492
male reproductive system disease	right_handed_peptide_helix	3.51659	0.0189492
esophageal neoplasm	polypeptide_secondary_structure	3.4855	0.004729497
genetic disorder	intramembrane_polypeptide_region	3.39768	0.01797461
endometrial carcinoma	alpha_helix	3.34845	2.62E-07
nervous system heterodegenerative disease	transmembrane_polypeptide_region	3.23602	0.000345457
brain neoplasm	beta_strand	3.13996	2.27E-10
neoplastic disease	compositionally_biased_region_of_peptide	3.09438	1.25E-06
brain neoplasm	alpha_helix	2.85495	3.67E-07
intracranial neoplasm	right_handed_peptide_helix	2.85495	3.67E-07
cancer	aminoacid_enriched_region	2.79248	0.000262729
endometrial carcinoma	beta_strand	2.7285	0.000702741
cancer of reproductive system	right_handed_peptide_helix	2.71723	0.003065169
colorectal cancer	aminoacid_enriched_region	2.66113	0.000247744
sphingolipidosis	polypeptide_secondary_structure	2.64929	0.000732642
neurodegenerative disease	intramembrane_polypeptide_region	2.57807	0.03620567
congenital disorder	membrane_structure	2.57782	0.000264116
bone marrow cancer	beta_strand	2.5578	1.63E-09
inborn errors of metabolism	intramembrane_polypeptide_region	2.48253	0.002940634
endocrine system disease	transmembrane_polypeptide_region	2.46725	0.000109441
lymphoma	beta_strand	2.41199	2.65E-11
large Intestine carcinoma	compositionally_biased_region_of_peptide	2.40239	0.002462735
simple genetic disease	membrane_structure	2.24123	0.000973413
reproductive system disease	polypeptide_secondary_structure	2.12031	0.00179245
lymphoid cancer	polypeptide_secondary_structure	2.08594	1.12E-11
bone marrow disease	polypeptide_secondary_structure	1.95547	2.97E-06
brain disease	right_handed_peptide_helix	1.93006	1.60E-05
brain disease	polypeptide_secondary_structure	1.86369	0.002784963
soft tissue disease	polypeptide_secondary_structure	1.75798	0.01261616
central nervous system neoplasm	peptide_helix	1.74419	0.04907455
cancer by anatomical entity	alpha_helix	1.56422	0.01003404
central nervous system neoplasm	polypeptide_structural_region	1.54158	0.01940891
pelvic cancer	polypeptide_secondary_structure	1.51004	0.01391864
cancer	intrinsically_unstructured_polypeptide_region	1.46751	0.000147879
neoplastic disease	polypeptide_structural_region	1.30499	0.000376776
disease of cellular proliferation	polypeptide_region	1.29831	0.000146788
carcinoma	intrinsically_unstructured_polypeptide_region	1.25414	0.03249705
endocrine system disease	intrinsically_unstructured_polypeptide_region	0.604194	0.000784021
carcinoma	transmembrane_polypeptide_region	0.537177	0.003694729
lymphoid cancer	polypeptide_structural_region	0.533869	0.003086615
bone marrow disease	polypeptide_structural_region	0.451679	0.000452141
inborn errors of metabolism	intrinsically_unstructured_polypeptide_region	0.387852	1.34E-06
retinal disease	polypeptide_secondary_structure	0.355451	0.02750617
breast carcinoma	beta_strand	0.350784	7.50E-09
adrenal gland disease	polypeptide_structural_region	0.330911	0.005149903
neuromuscular disease	polypeptide_secondary_structure	0.315643	0.002057483
eye disease	beta_strand	0.298127	0.02736738
cancer	peptide_localization_signal	0.296133	0.01964821
sphingolipidosis	intrinsically_unstructured_polypeptide_region	0.29438	0.01507502
myopathy	alpha_helix	0.26193	0.04833389

disease of anatomical entity	coiled_coil	0.25869	0.004289242
muscle tissue disease	right_handed_peptide_helix	0.257375	0.02356769
neoplastic disease	intramembrane_polypeptide_region	0.242143	1.32E-33
disease of cellular proliferation	membrane_structure	0.242143	1.32E-33
urologic neoplasm	transmembrane_polypeptide_region	0.214357	0.007497535
cancer	disulfide_bond	0.185306	0.04963278
cancer	transmembrane_polypeptide_region	0.174619	1.71E-15
inborn errors of amino acid metabolism	intrinsically_unstructured_polypeptide_region	0.167623	0.001303283
adrenal hyperplasia	polypeptide_region	0.155605	9.49E-05
gallbladder carcinoma	polypeptide_region	0.145595	0.01034421
carcinoma	disulfide_bond	0.123528	7.56E-06
simple genetic disease	aminoacid_enriched_region	0.115732	0.002828451
hereditary disease	compositionally_biased_region_of_peptide	0.105618	0.000264869
lung carcinoma	transmembrane_polypeptide_region	0.0971447	8.65E-08

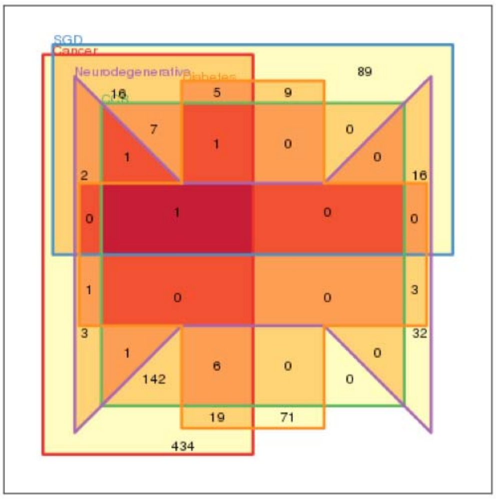
9.5. Intra-SO associations

Term 1	Term 2	OR	P val
Ser-rich	intrinsically_unstructured_polypeptide_region	5.68	6.40E-16
Ser-rich	phosphorylation	3.65	3.49E-07
disulfide_bond	beta_strand	3.52	1.81E-02
o_glycosylation	intrinsically_unstructured_polypeptide_region	5.61	2.04E-06
o_glycosylation	aminoacid_enriched_region	7.58	2.97E-02
o_glycosylation	phosphorylation	4.19	1.54E-03
protein_binding_site	beta_strand	3.77	8.34E-03
alpha_helix	transmembrane_polypeptide_region	0.19	2.58E-06
alpha_helix	phosphorylation	0.54	3.05E-06
signal_peptide	palmitoylation type III	55.81	4.69E-02
coiled_coil	intrinsically_unstructured_polypeptide_region	16.68	1.89E-28
coiled_coil	polypeptide_secondary_structure	0.16	7.19E-03
sequence_variant_causing_inactive_ligand_binding_site	beta_strand	3.77	8.34E-03
palmitoylation type III	signal_peptide	55.81	4.69E-02
n_glycosylation	intrinsically_unstructured_polypeptide_region	0.26	8.53E-03
transmembrane_polypeptide_region	alpha_helix	0.19	2.58E-06
transmembrane_polypeptide_region	intrinsically_unstructured_polypeptide_region	0.04	6.63E-37
transmembrane_polypeptide_region	phosphorylation	0.21	1.63E-10
beta_strand	disulfide_bond	3.52	1.81E-02
beta_strand	protein_binding_site	3.77	8.34E-03
beta_strand	sequence_variant_causing_inactive_ligand_binding_site	3.77	8.34E-03
beta_strand	beta_turn	2.86	1.47E-06
intrinsically_unstructured_polypeptide_region	Ser-rich	5.68	6.40E-16
intrinsically_unstructured_polypeptide_region	o_glycosylation	5.61	2.04E-06
intrinsically_unstructured_polypeptide_region	coiled_coil	16.68	1.89E-28
intrinsically_unstructured_polypeptide_region	n_glycosylation	0.26	8.53E-03
intrinsically_unstructured_polypeptide_region	transmembrane_polypeptide_region	0.04	6.63E-37

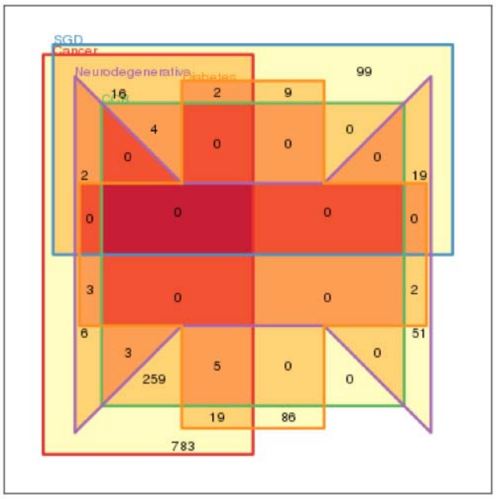
intrinsically_unstructured_polypeptide_region	peptide_localization_signal	0.17	3.35E-02
intrinsically_unstructured_polypeptide_region	phosphorylation	2.09	4.99E-32
intrinsically_unstructured_polypeptide_region	compositionally_biased_region_of_peptide	2.30	9.23E-03
intrinsically_unstructured_polypeptide_region	polypeptide_secondary_structure	0.79	3.68E-02
beta_turn	beta_strand	2.86	1.47E-06
right_handed_peptide_helix	intramembrane_polypeptide_region	0.19	2.58E-06
right_handed_peptide_helix	post_translationally_modified_region	0.50	1.59E-08
binding_site	polypeptide_secondary_structure	3.46	1.80E-03
aminoacid_enriched_region	o_glycosylation	7.58	2.97E-02
aminoacid_enriched_region	post_translationally_modified_region	1.97	1.49E-02
peptide_localization_signal	intrinsically_unstructured_polypeptide_region	0.17	3.35E-02
phosphorylation	Ser-rich	3.65	3.49E-07
phosphorylation	o_glycosylation	4.19	1.54E-03
phosphorylation	alpha_helix	0.54	3.05E-06
phosphorylation	transmembrane_polypeptide_region	0.21	1.63E-10
phosphorylation	intrinsically_unstructured_polypeptide_region	2.09	4.99E-32
intramembrane_polypeptide_region	right_handed_peptide_helix	0.19	2.58E-06
intramembrane_polypeptide_region	post_translationally_modified_region	0.27	4.00E-09
post_translationally_modified_region	right_handed_peptide_helix	0.50	1.59E-08
post_translationally_modified_region	aminoacid_enriched_region	1.97	1.49E-02
post_translationally_modified_region	intramembrane_polypeptide_region	0.27	4.00E-09
post_translationally_modified_region	polypeptide_structural_region	1.52	2.63E-10
membrane_structure	peptide_helix	0.19	2.58E-06
membrane_structure	biochemical_region_of_peptide	0.27	2.75E-09
compositionally_biased_region_of_peptide	intrinsically_unstructured_polypeptide_region	2.30	9.23E-03
compositionally_biased_region_of_peptide	polypeptide_secondary_structure	0.11	1.11E-14
compositionally_biased_region_of_peptide	biochemical_region_of_peptide	2.12	2.77E-04
peptide_helix	membrane_structure	0.19	2.58E-06
peptide_helix	biochemical_region_of_peptide	0.50	1.22E-08
polypeptide_secondary_structure	coiled_coil	0.16	7.19E-03
polypeptide_secondary_structure	intrinsically_unstructured_polypeptide_region	0.79	3.68E-02
polypeptide_secondary_structure	binding_site	3.46	1.80E-03
polypeptide_secondary_structure	compositionally_biased_region_of_peptide	0.11	1.11E-14
polypeptide_secondary_structure	polypeptide_motif	0.70	2.35E-04
biochemical_region_of_peptide	membrane_structure	0.27	2.75E-09
biochemical_region_of_peptide	compositionally_biased_region_of_peptide	2.12	2.77E-04
biochemical_region_of_peptide	peptide_helix	0.50	1.22E-08
polypeptide_structural_region	post_translationally_modified_region	1.52	2.63E-10
polypeptide_motif	polypeptide_secondary_structure	0.70	2.35E-04

9.6. Overlap between diseases in each network

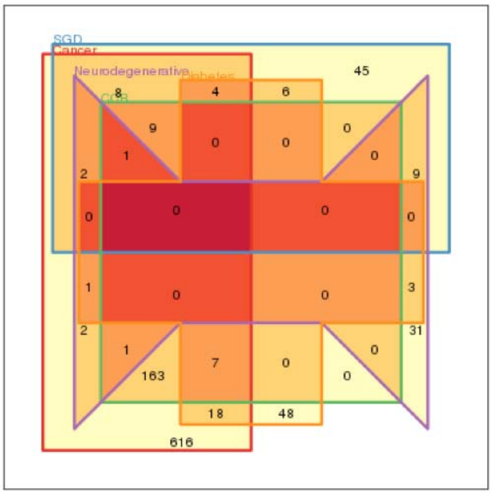
Reactome



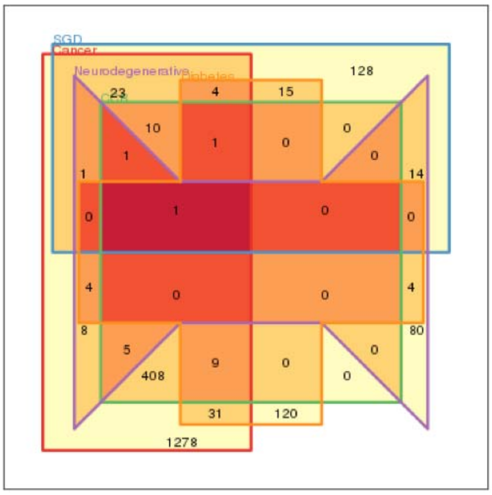
BioGPS



HPRD

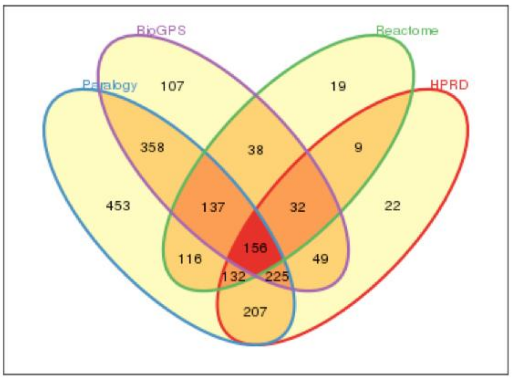


Paralogy

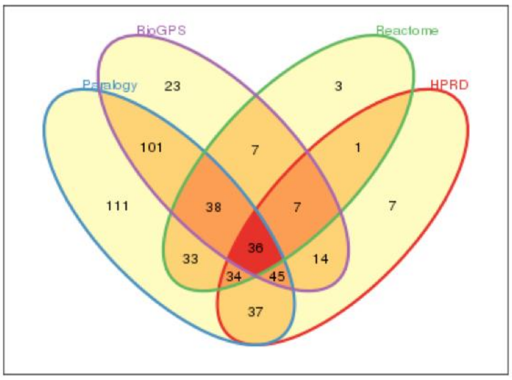


9.7. Overlap between networks in each disease

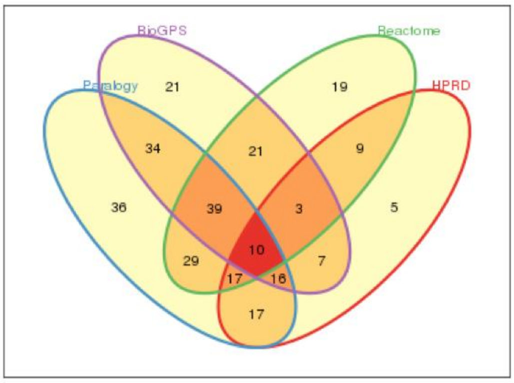
Cancer



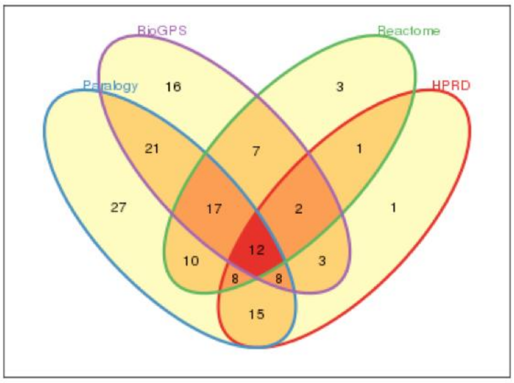
Colorectal Cancer



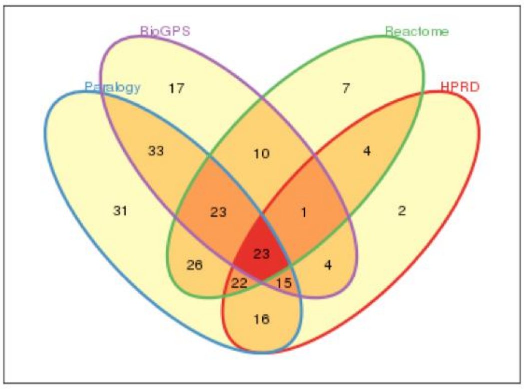
Simple genetic diseases



Neurodegenerative disorders

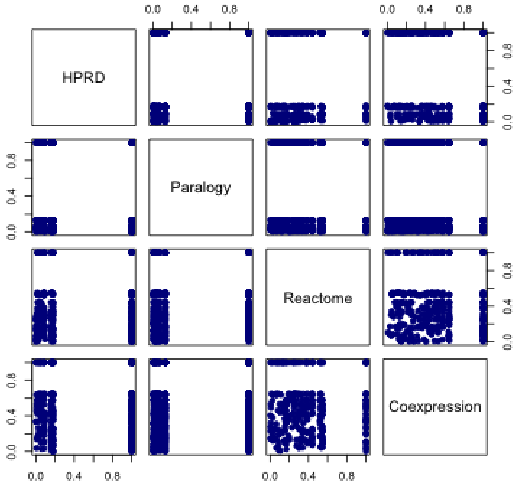


Diabetes

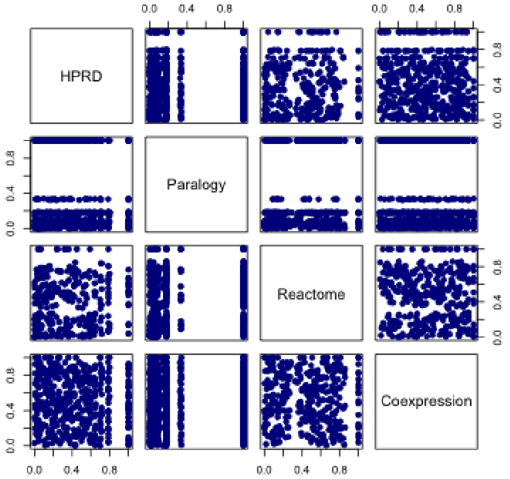


9.8. Correlations between scores of genes in the different networks

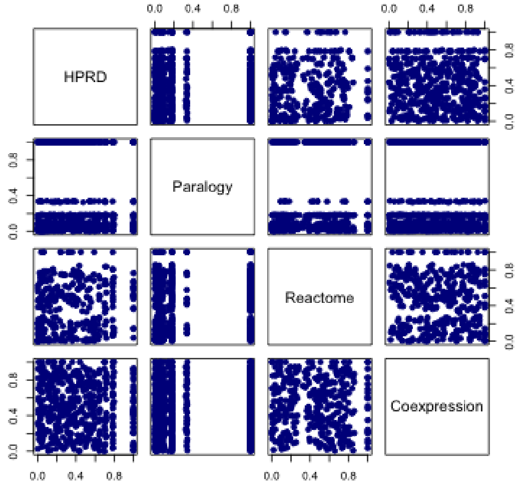
Cancer DN - 1



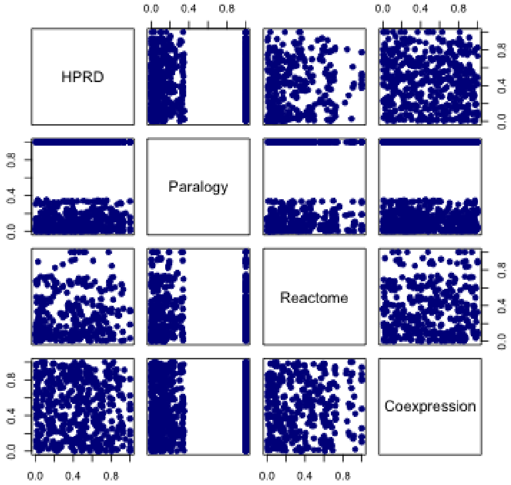
Cancer DN - 2



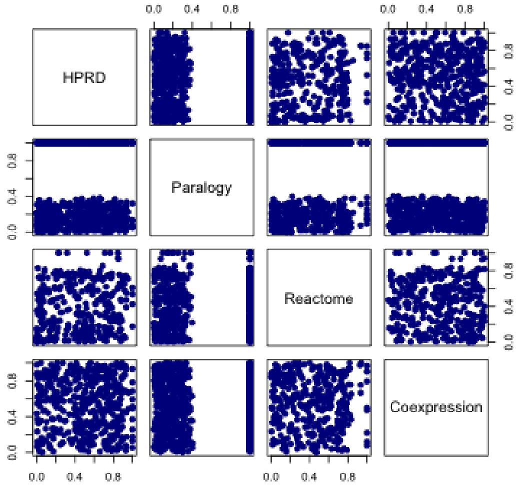
Cancer DN - 3



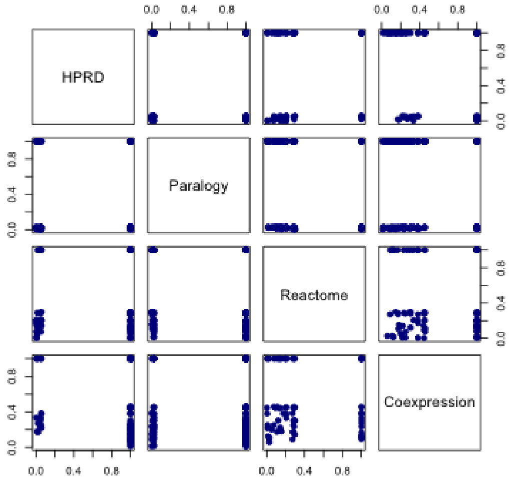
Cancer DK



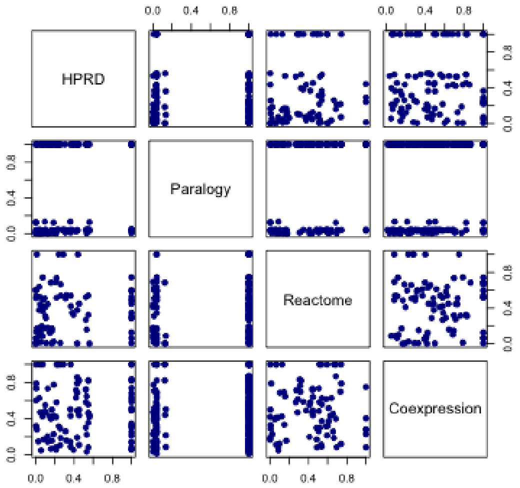
Cancer RWR



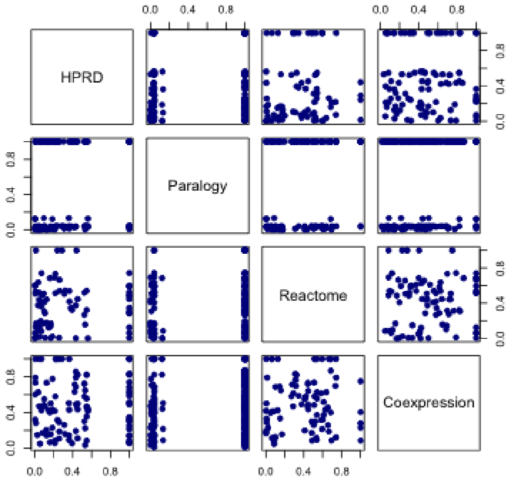
Colorectal Cancer DN - 1



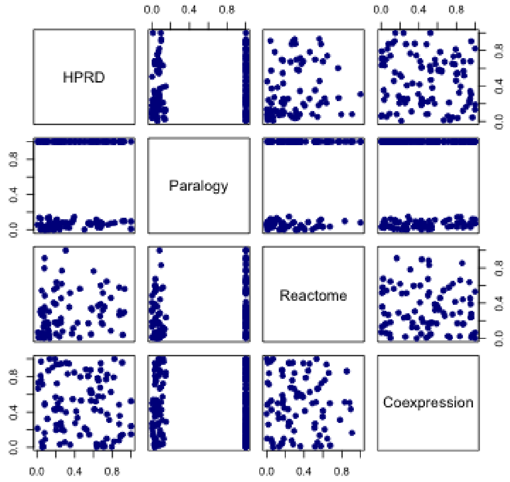
Colorectal Cancer DN - 2



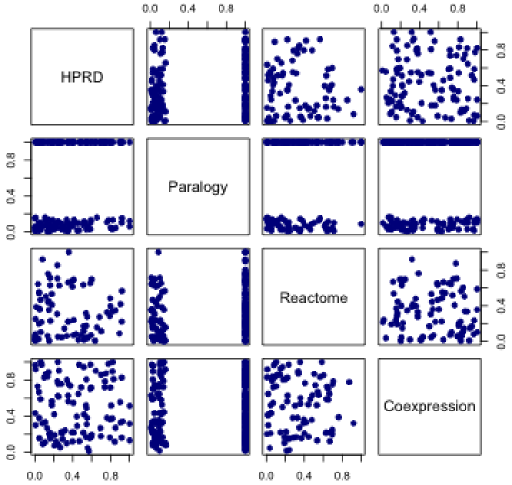
Colorectal Cancer DN - 3



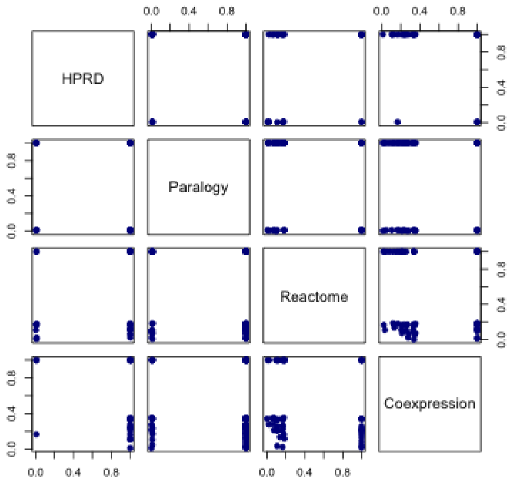
Colorectal Cancer DK



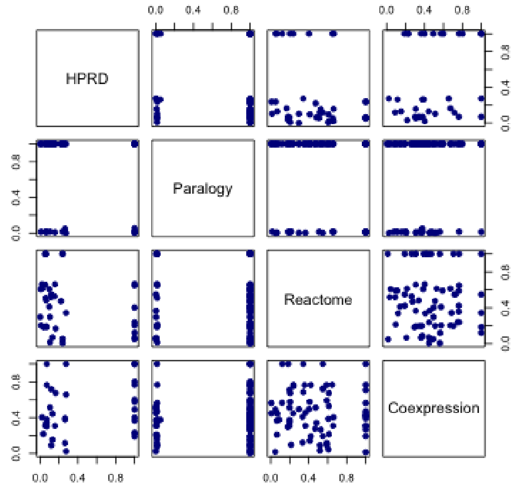
Colorectal Cancer RWR



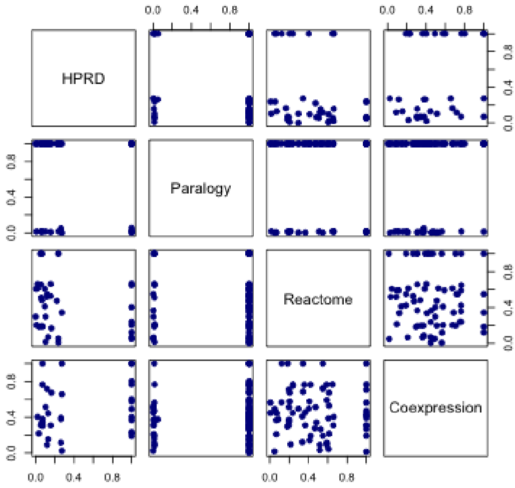
Simple Genetic Diseases DN - 1



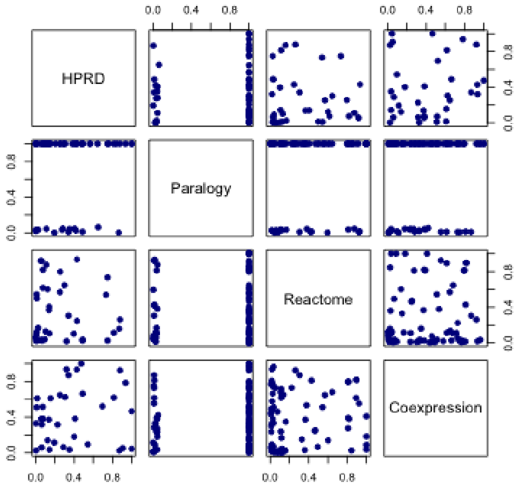
Simple Genetic Diseases DN - 2



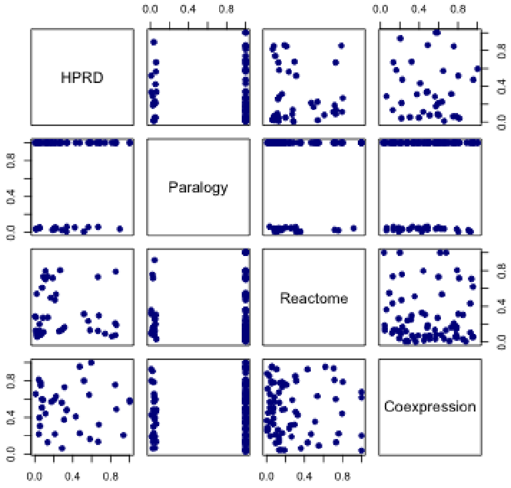
Simple Genetic Diseases DN - 3



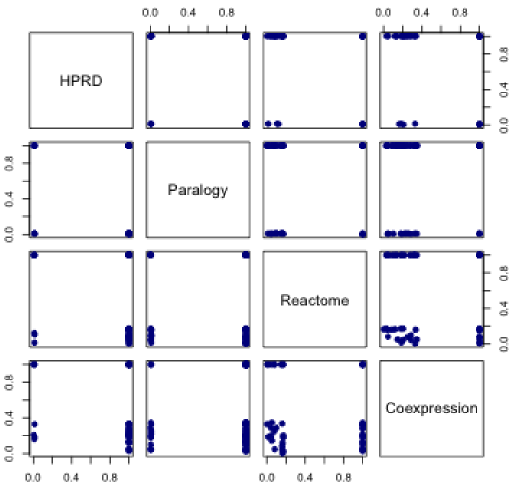
Simple Genetic Diseases DK



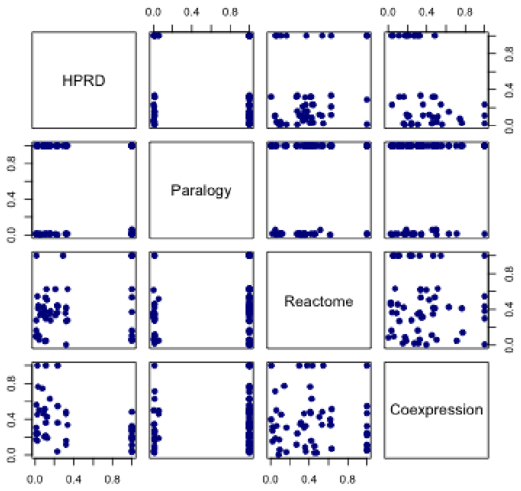
Simple Genetic Diseases RWR



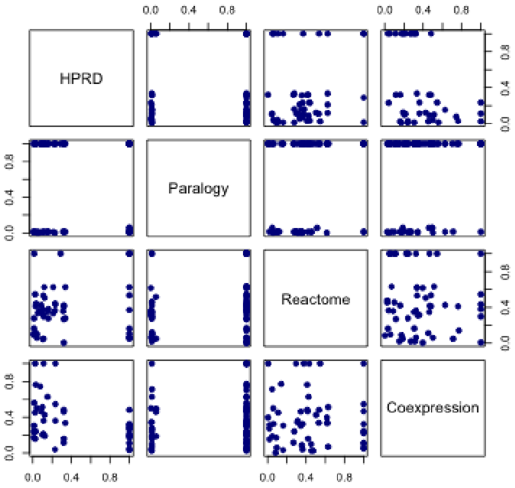
Diabetes DN - 1



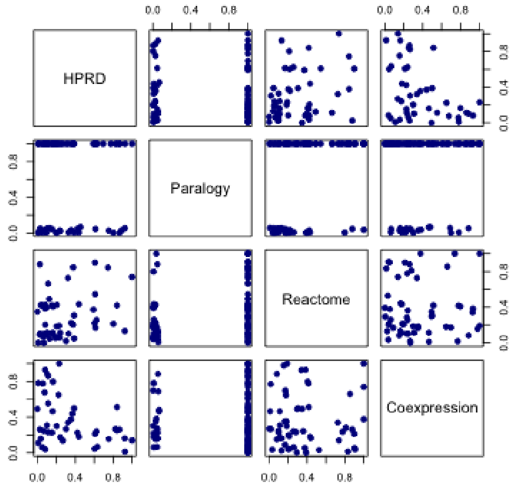
Diabetes DN - 2



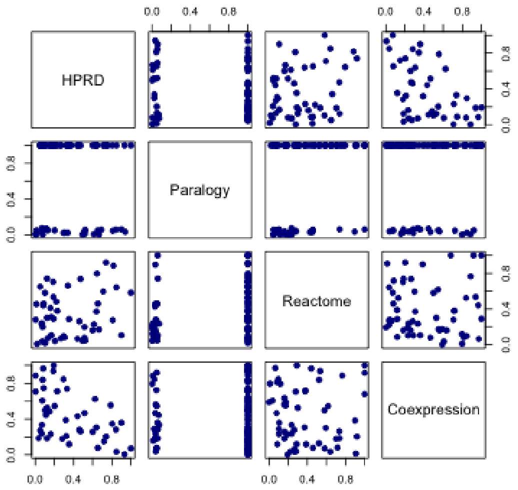
Diabetes DN - 3



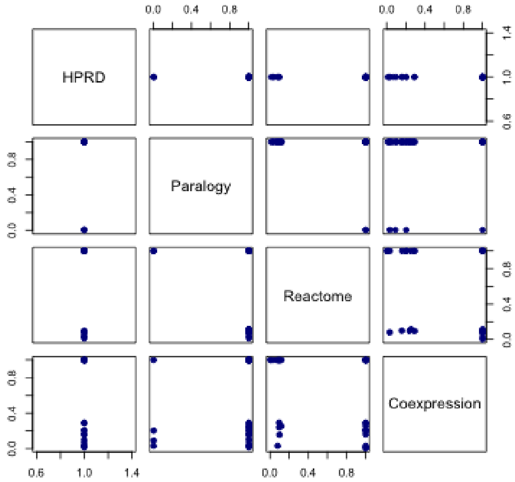
Diabetes DK



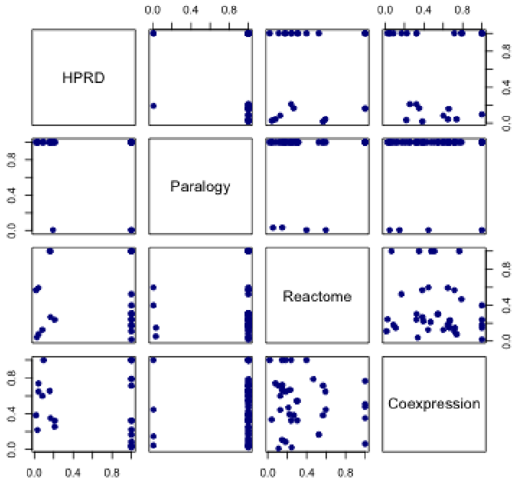
Diabetes RWR



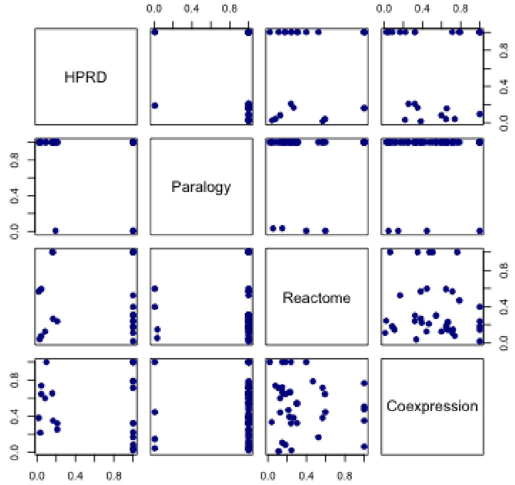
Neurodegenerative Diseases DN - 1



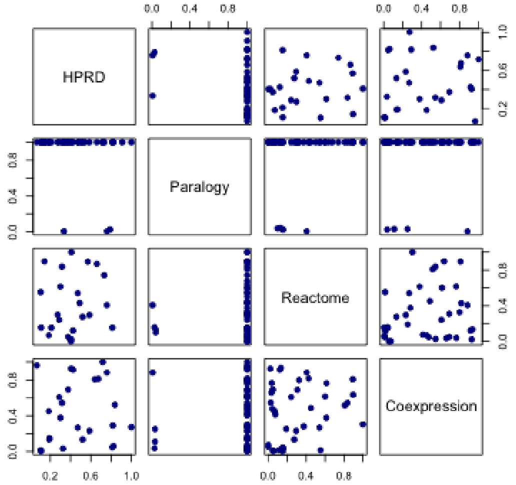
Neurodegenerative Diseases DN - 2



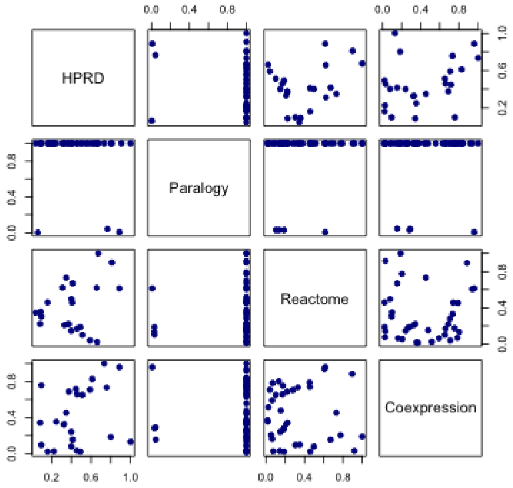
Neurodegenerative Diseases DN - 3



Neurodegenerative Diseases DK



Neurodegenerative Diseases RWR


HPRD
Paralogy
Reactome
Coexpression