An Exploratory Analysis on Drug Target Locality

Juan J. Cáceres* and Alberto Paccanaro*†
* Department of Computer Science - Centre for Systems and Synthetic Biology
Royal Holloway, University of London
Egham, Surrey, UK
† Departamento de Electrónica e Informática - Facultad de Ciencias y Tecnología
Universidad Católica "Nuestra Señora de la Asunción"
Asunción, Paraguay

***Abstract*—From a network medicine perspective, diseases are caused by perturbations in the dynamics of multiple interacting genes - a disease module. A drug that is a suitable candidate for re-purposing, should affect perturbed disease modules other than the one for which it was designed. In other words, it must act on various disease modules. A systematic analysis of re purposing suitability requires deeper understanding of drug target modularity. In this paper, we present a large-scale analysis of drug-target relationships, evaluating the locality of drug targets in protein-protein interaction networks. We show that the various drugs in each category affect different regions in biological networks, and present modular features. Additionally, multiple targets associated to the same drug appear close in the interactome. Our statistical analysis of the functions of the known drug targets reveals that peripheral functions of disease modules, such as signalling, are common targets for many drugs.**

I. Introduction

Biological networks are used to describe the internal operation of an organism; they encompass different levels of abstraction and scope. Common biological networks are: protein-protein interaction networks, gene transcription networks, gene regulatory networks, signalling networks, metabolic networks, and neural networks [1]. Each of this networks show a different aspect of the processes and interactions of biological components in an organism. In particular, molecular interactions of a cell define an interactome.

The availability of high throughput methods, new sequencing techniques and computational annotation greatly increased the number of high quality biological networks. High quality networks collect manually curated interactions from experimental data, and provide network methods the capability to extract significant results on analysis of complex biological systems.

Given the level of complexity of the physiological systems in play in human disease, the cause of a disease can rarely be traced to a single gene but should rather be considered a perturbation in the molecular interactions in the cell – the disease module [7]. This network view of diseases [7] involves a wider analysis that follows functional relationships, physical interaction and metabolic pathways to decode the mechanisms underlying complex diseases.

The modules of diseases with similar physiological characteristics tend to be close in the interactome. The modular nature of the diseases implies that diseases can be categorised with respect to the physiological system they affect [5]. Categories such as Cancer or Neurological include diseases that result from perturbations in close-by regions in the interactome [5]. Furthermore, the modular nature of diseases is also revealed through the analysis of disease phenotype [12], [13].

Network analysis for drug repurposing resulted also in the systemic analysis of drug and protein binding. Previous work has shown that drug targets are rarely protein products of essential genes, and a prevalence of intra modular protein targets [14]. A further look onto phenotype can serve to understand other aspects of drug target modularity.

Over 2,150 FDA (U.S. Food and Drug Administration) approved drugs are listed in DrugBank [10], with about 1,700 targeting some of the 4,200 known protein targets. The number of diseases greatly exceeds the indications available, and with average costs for developing a new drug exceeding £1,500 million, drug re-purposing of approved drugs is a promising option. The need for comprehensive understanding of the underlying mechanisms of diseases and drugs is still an unresolved problem [8]; only the among heritable diseases listed in the Online Mendelian Inheritance in Man (OMIM), there are 7,800 different diseases such as Alzheimer's and Cancer.

In this paper we perform an exploratory analysis of the genotypic and genotypic relation between drugs and heritable diseases. Since drugs and diseases are *a priori* incomparable sets, we propose a drug-disease category mapping, to define an intended objective of drug design. We establish metrics such as the locality and functional similarity of drug targets, to quantify their specificity and possible side effects. The locality of drug targets can be used to analyse the physical mechanism of action drugs use to treat diseases. On the other hand, functional similarity of drug targets reveals functions that can be exploited for drug repursosing.

II. Methods

First, we establish a mapping between diseases and drugs to create a coarse model of drug target locality. The coarse model will be useful to understand an underlying trend of drug design. This approach allows us to create a novel lax approximation to the drug target counterpart of disease modules, i.e. drug target modules. Then, we describe the

measures we established to compare these modules, from their structural genotype network properties, to the phenotype traits they express.

A. Category mapping

Goh *et. al.* manually curated OMIM, and produced 21 disease categories to group diseases based on the physiological system affected in 2005. The disease categories are: Bone, Cancer, Cardiovascular, Connective tissue disorder, Dermatological, Developmental, Ear-Nose-Throat, Endocrine, Gastrointestinal, Hematological, Immunological, Metabolic, Multiple, Muscular, Neurological, Nutritional, Ophthalmologic, Psychiatric, Renal, Respiratory and Skeletal. We bring up to date the mapping by including recent OMIM disease to gene associations (downloaded on March 30, 2017) for each of the curated diseases.

DrugBank groups drugs into Anatomical Therapeutic Chemical (ATC) categories, according to the physiological system affected by a drug and therapeutic properties [10]. The ATC classification is hierarchical, where the top level categories represent the most general elements. We considered that drugs belonging to a specific category in DrugBank have an effect on all the diseases in the corresponding physiological category presented by Goh *et.al*. Naturally, not all ATC categories will match a Goh category, such as the top level category J - *Antiparasitic products, insecticides and repellents*, as those drugs are not designed to target humans, but foreign organisms.

The analysis was performed in the 2017 released 5.05 DrugBank XML file. A summary of drugs per ATC category is included in Table I. Notice that each drug can belong to multiple ATC categories, so the total number of drugs is not the sum of the individual categories.

Category	Total	FDA approved
A	268	233
B	107	96
C	245	221
D	156	140
G	126	117
H	50	45
J	235	220
L	235	216
M	106	91
N	345	302
P	46	39
R	166	145
S	147	141
V	111	102
N/A	6298	414
Overall	8257	2157

TABLE I: **Drugs found in DrugBank per top level ATC category**.

We notice that not all Goh categories map naturally to ATC categories, thus DrugBank's pharmacological action categories are used as fallback whenever the Goh category did not match an ATC category. The *Multiple* Goh category has been left unmatched, as each disease requires an independent drug association. Furthermore, the *Muscular* and *Skeletal* were merged into a *Musculoskeletal* category, as we found no good way to separate the musculoskeletal drugs into each independent category. A summary of the FDA approved drugs, drug targets and disease proteins per category is shown in Table II.

Categories			Elements	
Goh	ATC	DBCat	Diseases	Drugs
Bone	-	89, 2165	52	25
Cancer	L01	-	198	138
Cardiovascular	C, V09G	-	98	223
Connective tissue	V03AK	-	49	0
Dermatological	D	-	95	140
Developmental	H01	-	51	21
Ear-Nose-Throat	S02	2159, 2106	52	62
Endocrine	L02	-	99	22
Gastrointestinal	A03	-	34	25
Hematological	B	-	157	96
Immunological	L[03,04]	-	105	58
Metabolic	A[09,14,15,16]	-	291	32
Musculoskeletal	M	-	167	91
Neurological	N	-	270	302
Nutritional	-	234, 2733	12	47
Ophthalmologic	S01	-	149	137
Psychiatric	-	529	26	47
Renal	-	629	59	14
Respiratory	R	-	22	145
Overall			1928	1302

TABLE II: **Category mapping**. The *Categories* column shows how diseases from Goh categories map drugs from ATC or DrugBank categories; - indicates no mapping between the Disease category and the Drug category. The *Elements* column shows the number of drugs and diseases per category.

Over 60% of FDA approved drugs were mapped to a Goh category. We believe that this is a good coverage of the drug set, and by broadening the mapping we may reduce the confidence in the associations of drugs to Goh categories.

B. Physical distance of proteins

The interactome can provide systemic information about a set of proteins. Neighbour proteins in the interactome are interactors, and clusters of proteins are likely to share function [13]. We consider that the distance between drug targets can offer perspective into mechanism of action for drug repurposing. Generally speaking, close targets will share the effect, while distant targets will have different effects.

In our analysis, the distance between two proteins is the minimum path length between a pair of proteins in an unweighted and undirected graph, that represents the protein-protein interaction (PPI) network given by the Human Protein Reference Database (HPRD) [15]. The HPRD PPI is hand curated, and provides high quality, experimentally verified protein-protein interactions.

Since many of the comparisons needed are among sets of targets, a fair measure must be established for sets of distances. We use an unequal variances t-test (or Welch's t-test), to compare if the distances from a pair of sets comes from the same distribution. It is formally defined as:

$$t = \frac{\bar{A} - \bar{B}}{\sqrt{\frac{\sigma_A^2}{|A|} + \frac{\sigma_B^2}{|B|}}},$$

where A, B are sets of values, and $\bar{X}$ is the mean, σ_X^2 is the variance and $|X|$ is the size of set X. The t-statistic can be later translated to a confidence p-value according to the degrees of freedom associated to the variance estimates. Typical confidence values to confirm that the sets come from different distributions are p-value < 0.05 or p-value < 0.01.

C. Functional relation of proteins

We characterise the phenotype of a gene by the aggregation of functions performed by the proteins it produces. We obtain gene functions from the Gene Ontology (GO), and calculate the most significant functions per category. Moreover, the functional relation of drug targets can be analysed by comparing mathematical models of the function distribution per category.

The Gene Ontology is a taxonomy designed with the aim to standardise the description of genes and gene products across databases, that is, it describes a unique vocabulary for all organisms. The taxonomy is divided in three major domains, namely: biological process, cellular component and molecular function. As with most taxonomies, it is presented as an up-rooted structure, in which the terms are arranged in ever increasing specificity, with the root being the less specific term and leaves the most specific terms [6]. This ontology is used to annotate proteins. That is, we can assign a set of terms to a protein that will describe it without ambiguity while allowing for a clear understanding of the current information available.

Running statistical tests on the annotated gene ontology terms is a common approach to find the most relevant functions performed by sets of proteins. The terms that appear significantly more annotated by the proteins in the set of interest, in contrast to the terms annotated by all human proteins, define the relevant functional categories of the set of proteins. This approach is known as a Gene Ontology over-representation analysis. Some of the common statistical models are the χ^2, and Fisher's exact test [2]. We use the Fisher's exact test as we want conservative estimates in the number of over-represented GO categories.

Furthermore, semantic similarity calculations in the Gene Ontology produce a quantifiable measure of relatedness of genes and gene products. That is, the produce is a single number that quantifies the functional similarity of two terms in the ontology, and since proteins are annotated by terms in the ontology, their function can be compared by way of the semantic similarity of the annotating terms [4]. We calculate the semantic similarity for every GO term in the Molecular Function and Biological Process domains, as they reveal functional aspects of the drug targets. The Resnik similarity is calculated with GOssTo [16] on the GO annotations downloaded on April 5, 2017.

To find the significance of the functional similarity within a set, we characterise pairwise semantic similarities between all terms. The difference between the domain and the category distribution of terms hints how specific is the mechanism of action of the drug category.

Term sets are characterised by a random variable distribution. However, to the best of our knowledge there is no known term distribution for these sets. Therefore, we tested every random variable distribution from the Python SciPY library to fit the model. We want a random variable which approximates both the cumulative distribution and the mean of similarities.

The real cumulative distribution for a collection of similarity values x is given by the cumulative histogram $H_i(x)$. Formally, the cumulative distribution of the first i similarity bins is:

$$H_i(x) = \sum_{j=1}^{i} h_j(x)$$

where $h_j(x)$ (bin j for the histogram of x) is the amount of elements in x between the boundaries b_{j-1} and b_j. The similarity boundaries $b = (b_0, \ldots, b_n) \in \mathbb{R}^n$ are equally distributed between the minimum and maximum similarity values, with $b_0 = \min\{x\}$ and $b_n = \max\{x\}$.

Therefore, we pick the random variable distribution which minimises J for both domains:

$$J = (0.5 - F(\bar{x}))^2 + \frac{1}{n} \sum_{i=1}^{n} \left(H_i(x) - F(b_i)\right)^2,$$

where $\bar{x}$ is the mean of x, and $F(b_i)$ is the fitted cumulative distribution function of the random variable at the similarity value b_i.

The half logistic distribution is the consensus best fit in the domains. The cumulative distribution function is:

$$F(k) = \frac{1 - e^{-k}}{1 + e^{-k}}$$

We consider the similarity of a category significant if the mean of the category is above 80% of the domain cumulative distribution. This is, the mean similarity of the category is among the top 20% percentile of similarity values in the sub ontology.

III. Analysis

The first analysis is intended to get a basic genotype relation between drugs. The second experiment is intended to select the most important molecular functions performed by drug targets. Finally, the third experiment characterises the phenotype relation between the most important molecular functions from the drug targets.

A. Drug targets and known disease proteins

To clarify the nature of the drug-disease relationship, we extracted the disease proteins from all known heritable diseases in OMIM and counted the number of drug targets that are known diseases associated proteins. Table III shows the number of shared proteins between diseases and drugs per category.

We observe that only 137 proteins from the 306 common proteins (e.g. disease proteins that are also drug targets) are in the same class. This means that more than half of the common proteins targeted by drugs, are specific to diseases from other categories.

Category	Disease Proteins	Drug Targets	Overlap
Bone	42	51	4
Cancer	195	298	26
Cardiovascular	129	318	18
Connective tissue	98	0	0
Dermatological	147	316	2
Developmental	51	38	2
Ear-Nose-Throat	40	200	0
Endocrine	98	46	5
Gastrointestinal	34	29	0
Hematological	101	233	22
Immunological	163	140	10
Metabolic	247	104	13
Musculoskeletal	115	179	3
Neurological	251	367	22
Nutritional	19	195	1
Ophthalmologic	97	308	6
Psychiatric	31	81	7
Renal	48	56	1
Respiratory	45	232	6
Overall	1554	1280	306

TABLE III: **Break-down of proteins per category**. The *Disease proteins* column shows the number of different proteins that belong to a disease from each category. Correspondingly, the *Drug Targets* category shows the number of different proteins targeted by a drug from each category.

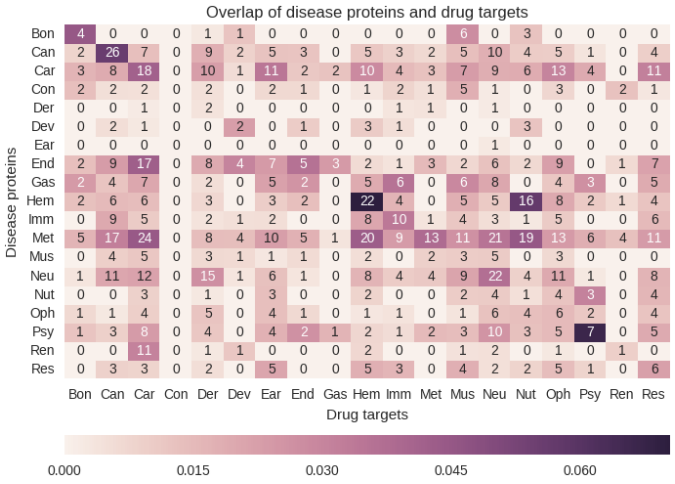


Fig. 1: Heatmap of the Jaccard coefficient between disease proteins and drug targets by category. Colors are based on the Jaccard coefficient, while labels indicate the actual number of common elements in the categories.

We calculate the Jaccard coefficient between all disease and drug categories the to further explore how common proteins are distributed. The Jaccard coefficient measures the diversity of the elements of two sets. Formally:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|},$$

where A and B are sets of elements. Figure 1 shows the resulting comparison (notice that the diagonal of the matrix is equal to the *Overlap* column from Table III).

We can clearly see that drugs do not exclusively target diseases from the same category. While some disease categories are preferentially targeted by drugs from the same category, most do not show this behaviour. In particular, Bone, Cancer, Cardiovascular, Developmental, Hematological, Immunological, Neurological, and Psychiatric diseases proteins are preferentially targeted by drugs from the same category. On the other hand, Endocrine, Nutritional, and Renal diseases, are preferentially targeted by Cardiovascular, Psychiatric and Cardiovascular drugs respectively. Metabolic diseases are targeted more preferentially by Cardiovascular, Hematological and Nutritional drugs than by Metabolic drugs.

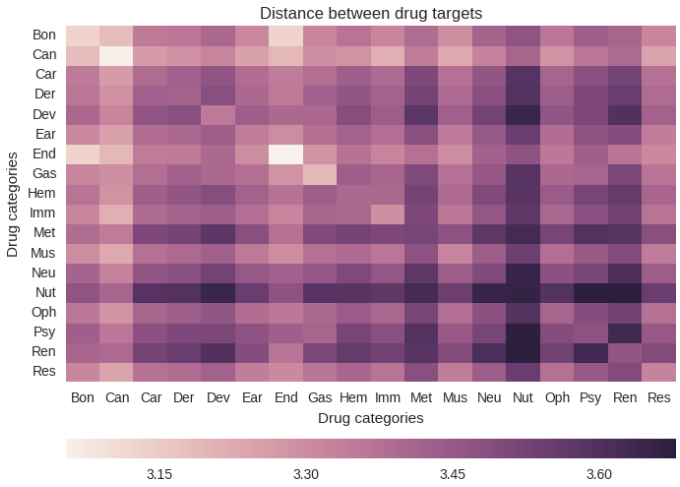


Fig. 2: Heatmap of target distance for category pairs. Colors indicate the average distance between all drug targets from category A to targets from category B.

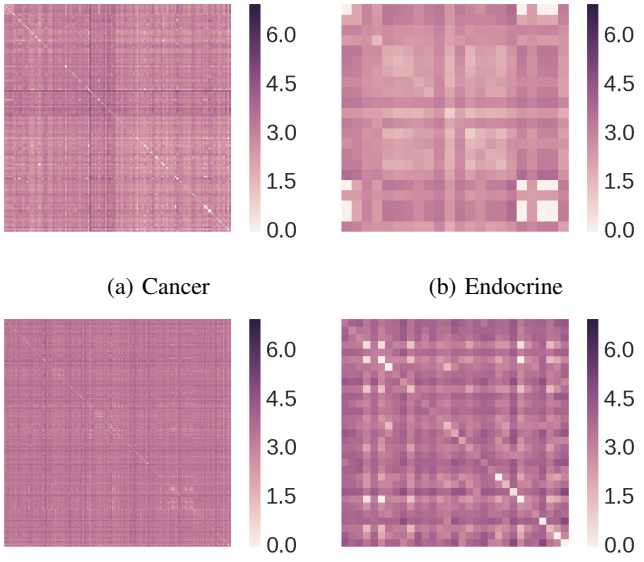
B. Physical distance between drug proteins

According to the network view of diseases, a disease is a wider perturbation in the interactome. To verify whether a drug was targeting the wider disease module, we analysed the distance of the target proteins for the mapped drug categories. Figure 2 shows that intra-category proteins are not in general closer than inter-category proteins, with the exception of Cancer and Endocrine drugs. This seems to indicate that most drugs do not affect specific modules in the network, but rather target proteins which have an indirect effect on the disease.

Although disease genes cluster in the interactome [7], the entire set of disease genes is not much closer than random proteins from the interactome. The average distance between all disease genes is 3.45, while the average interactome distance is 3.58. Drug targets appear to be somewhat closer than random genes, with an average distance of 3.25. Welch's t-test confirms that the distance distribution from drug targets is significantly different from a random choice in the interactome (p-value $< 1.0 \times 10^{-300}$). The slight modularity suggests that there are some functions which are particularly good as gene targets.

We focus on particular categories that are representative among all drugs to describe the intra-category relation of drug targets. Cancer and Endocrine drugs are selected because they have closer intra-category targets. On the other hand,

Nutritional drugs are selected because they have the furthest intra-category targets. Finally, Neurological drugs are selected as the largest drug category.



(a) Cancer (b) Endocrine

(c) Neurological (d) Nutritional

Fig. 3: Heatmaps target distance for drug pairs within a category. Colors indicate the average distance between targets from drug A to targets from drug B.

Figure 3 shows the details of the intra category distances of the drugs. Note that the main diagonal of each heatmap is significantly closer than the average. This is expected as in the main diagonal, the targets of a single drug are compared between each other. This distance will be zero only in the case that the drugs have a single target. Notice that these figures are not "zoom-in"s into diagonal elements from Figure 2; while in Figure 2 each target is included only once, in Figure 3 drugs can that share targets become closer. Table IV shows how the intra-category distances look from both perspectives.

Cancer drugs not only share the largest amount of targets with other categories (Figure 1), but also have closer targets with drugs from other categories (Figure 2).

The largest group of drugs from a single category that target the same protein belongs to the Endocrine category. Four drugs target Gonadotropin-releasing hormone receptor (GNRHR): Degarelix, Histrelin, Leuprolide, and Triptorelin. Endocrine drugs act in the tightest module among drug categories.

Drug targets from Neurological drugs appear to be randomly distributed in the interactome. Even if some drugs target specific disease genes, most targets do not lie in a module.

Nutritional drugs target genes that are more distant than the average of the interactome. Manual curation of the drugs reveals a clearly significant group of drugs within this category. Many Nutritional drugs are dietary supplements, which on their own are different enough to not target any specific module.

Category	by Category	by Drug
Bone	3.13	2.54
Cancer	3.06	2.95
Cardiovascular	3.39	3.27
Dermatological	3.42	2.86
Developmental	3.35	3.06
Ear-Nose-Throat	3.34	3.06
Endocrine	3.05	2.50
Gastrointestinal	3.19	2.52
Hematological	3.40	3.12
Immunological	3.30	2.99
Metabolic	3.52	3.42
Musculoskeletal	3.32	3.44
Neurological	3.50	3.33
Nutritional	3.66	3.40
Ophthamological	3.41	3.30
Psychiatric	3.48	3.12
Renal	3.46	3.02
Respiratory	3.33	2.96
Total	3.38	3.26

TABLE IV: **Average distance per category**. The columns show how target genes are grouped. Every target from any drug is included once when grouped *by Category*, while every target is included as many times as it appears like a drug target when grouped *by Drug*.

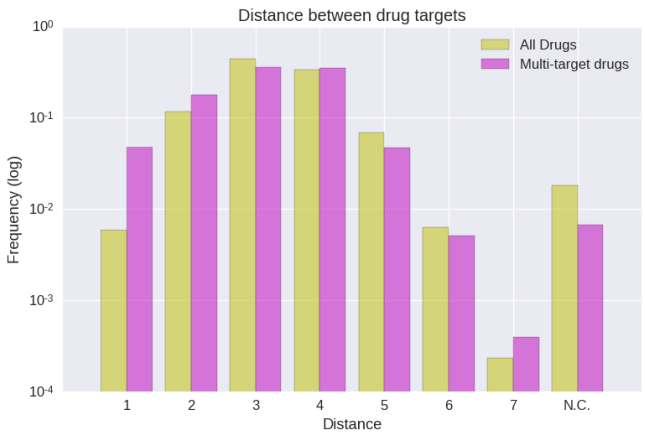


Fig. 4: Histogram of target distance for drugs with multiple targets. Lower distance values are better. N.C. means that the target proteins are not connected in the interactome.

Another interesting analysis is checking the distance of the protein targets for drugs that have multiple targets. If the targets of a drug are close in the interactome, the drug affects a module. Otherwise, the drug is affecting multiple modules and hints to other possible effects of that drug — e.g. side-effects. Figure 4 shows that the proteins from multi target drugs are closer than average proteins in the human interactome.

C. *Functional relation of target proteins*

Since drugs appear to affect non cohesive portions of the network, it is important to analyse the target proteins with respect to their function.

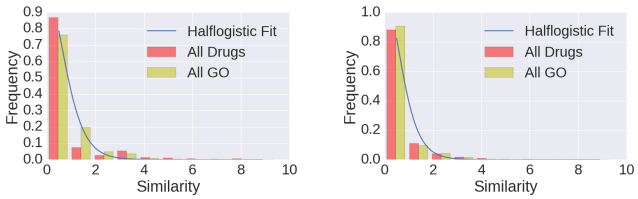
We have performed a Gene Ontology over-representation analysis using the Fisher's exact model with a Bonferroni

correction [3] to obtain the significant functional categories that represent the drug target proteins. Table V shows the 10 most over-represented GO terms for the proteins in each category, and their information content. It can readily be seen that most categories represent signalling, transport or binding processes, indicating that drug compounds target peripheral signalling mechanisms to produce their effect.

Category	GO Term	Name	Inf. Cont.
Cancer	GO:0044323	retinoic acid-respon	8.71
	GO:0015858	nucleoside transport	8.71
	GO:0006145	purine nucleobase ca	8.71
	GO:0004666	prostaglandin-endope	8.71
	GO:0003918	DNA topoisomerase ty	8.71
	GO:0004517	nitric-oxide synthas	8.30
	GO:0034875	caffeine oxidase act	8.01
	GO:0009820	alkaloid metabolic p	8.01
	GO:0005452	inorganic anion exch	8.01
	GO:0048384	retinoic acid recept	7.79
Endocrine	GO:0004769	steroid delta-isomer	8.71
	GO:0050294	steroid sulfotransfe	8.01
	GO:0034875	caffeine oxidase act	8.01
	GO:0034056	estrogen response el	8.01
	GO:0009820	alkaloid metabolic p	8.01
	GO:0030284	estrogen receptor ac	7.79
	GO:0009404	toxin metabolic proc	7.79
	GO:0002933	lipid hydroxylation	7.79
	GO:0016098	monoterpenoid metabo	7.61
	GO:0006068	ethanol catabolic pr	7.45
Neurological	GO:0086043	bundle of His cell a	8.71
	GO:0071886	1-(4-iodo-2,5-dimeth	8.71
	GO:0052834	inositol monophospha	8.71
	GO:0048630	skeletal muscle tiss	8.71
	GO:0045777	positive regulation	8.71
	GO:0019371	cyclooxygenase pathw	8.71
	GO:0014827	intestine smooth mus	8.71
	GO:0009804	coumarin metabolic p	8.71
	GO:0008292	acetylcholine biosyn	8.71
	GO:0005219	ryanodine-sensitive	8.71
Nutritional	GO:0070814	hydrogen sulfide bio	8.71
	GO:0047057	vitamin-K-epoxide re	8.71
	GO:0042167	heme catabolic proce	8.71
	GO:0035408	histone H3-T6 phosph	8.71
	GO:0035403	histone kinase activ	8.71
	GO:0017187	peptidyl-glutamic ac	8.71
	GO:0009804	coumarin metabolic p	8.71
	GO:0004666	prostaglandin-endope	8.71
	GO:0004485	methylcrotonoyl-CoA	8.71
	GO:0004357	glutamate-cysteine l	8.71

TABLE V: **Over-represented terms per category**. The *GO Term* and *Name* columns identify the over-represented GO term, and *Inf. Cont.* indicates the information content of the term (the higher the better).

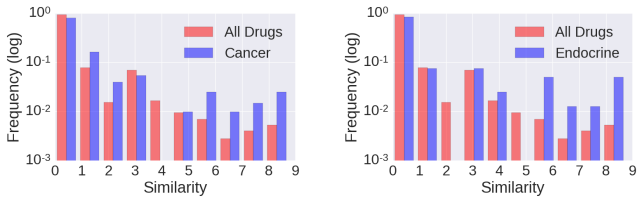
To verify that the over-represented proteins form a functionally coherent group, we have calculated the semantic similarity of the over-represented terms. We show that the function of the proteins in the over-represented group is significantly different than the overall function of the proteins in the human genome. Figure 5 compares how the significant terms from the molecular function and biological process domains separate from all annotations in the Gene Ontology. The significant terms from the drugs appear closer than expected by random chance; therefore, drugs target proteins from genes which are functionally related. Furthermore, figures 6 and 7 show the



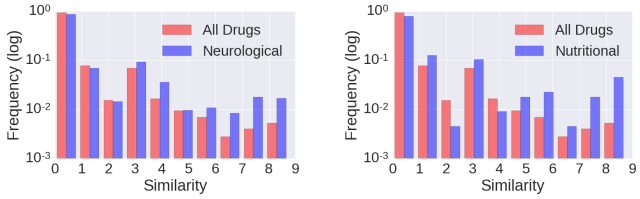
(a) Molecular Function (b) Biological Process

Fig. 5: Histogram of semantic similarity in the domains. Higher values of similarity indicates that the function is closer.

distribution of semantic similarity scores for the preselected drug categories.



(a) Cancer (b) Endocrine



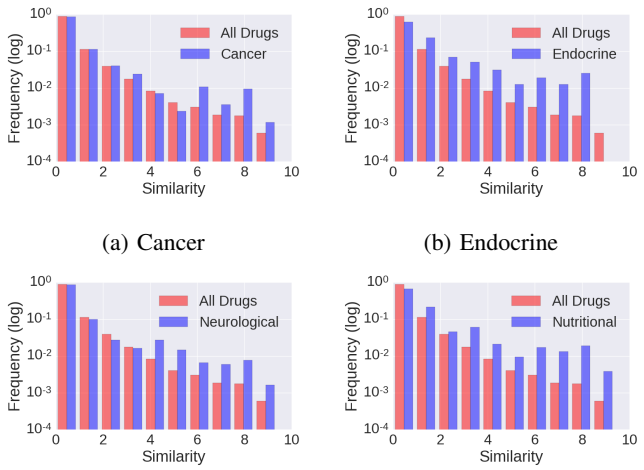
(c) Neurological (d) Nutritional

Fig. 6: Histogram of semantic similarity for specific groups in the molecular function domain. Higher values of similarity indicates that the function is closer.

The p-value column from Tables VI and VII show the confidence percentile by which we can reject that the sample, and the set of all the similarity values, come from the same distribution. The samples are the similarity values between every pair of over represented GO term in each category. This was done performing a one tailed t-test (since each distribution has a lower mean than the set of all the similarity values), on the samples and the set of all the similarity values for the biological process and the molecular function domains.

A review of Tables VI and VII reveals that all over represented categories have some preferential functions compared to the entire domain. This supports the evidence found by the physical distance between drug targets. Furthermore, some categories such as Bone and Renal are extremely specific in target functions.

The function of Cancer drugs is not very specific, in fact it shows as the least specific category in the biological process domain (Table VII). This fact is supported by the closeness of Cancer drug targets to other category targets presented in



(a) Cancer (b) Endocrine

(c) Neurological (d) Nutritional

Fig. 7: Histogram of semantic similarity for specific groups in the biological process domain. Higher values of similarity indicates that the function is closer.

Category	Mean	Percentile	p-value
Bone	1.92	4.3	1.71×10^{-02}
Cancer	0.91	28.1	2.87×10^{-03}
Cardiovascular	0.77	35.3	7.48×10^{-05}
Dermatological	1.28	14.6	4.76×10^{-18}
Developmental	1.64	7.4	4.65×10^{-02}
Ear-Nose-Throat	0.92	27.8	4.93×10^{-04}
Endocrine	1.16	18.0	1.13×10^{-02}
Gastrointestinal	1.79	5.5	1.23×10^{-02}
Hematological	0.85	31.1	3.72×10^{-05}
Immunological	0.98	25.1	5.09×10^{-03}
Metabolic	1.17	17.7	3.82×10^{-03}
Musculoskeletal	0.88	29.7	8.10×10^{-05}
Neurological	0.88	29.6	2.92×10^{-08}
Nutritional	1.13	18.9	2.31×10^{-05}
Ophthamological	0.88	29.8	1.08×10^{-05}
Psychiatric	1.17	17.9	2.90×10^{-04}
Renal	1.63	7.5	4.01×10^{-04}
Respiratory	0.85	31.2	2.78×10^{-04}

TABLE VI: **Semantic similarity significance in the molecular function domain**. We consider categories with means within the top 20 percentile to be significantly closer than the domain average.

the previous section. Cancer drugs are known to have large amounts of side effects and are used as an extreme measure [17].

IV. Conclusion

The construction of a categorical drug to disease mapping allows the usage of an evolving platform for drug and disease comparison. As DrugBank annotates drugs, they can be included into the defined categories. Although our analysis is centred on FDA approved drugs, the category mapping is not bounded to that restriction. These features allow the mapping to be useful also in experimental drug analysis.

Category	Mean	Percentile	p-value
Bone	1.46	7.8	3.40×10^{-10}
Cancer	0.66	38.0	6.24×10^{-03}
Cardiovascular	0.67	37.5	2.67×10^{-06}
Dermatological	0.79	30.1	5.63×10^{-10}
Developmental	1.30	10.8	1.16×10^{-03}
Ear-Nose-Throat	0.90	24.4	4.08×10^{-12}
Endocrine	1.57	6.1	3.23×10^{-14}
Gastrointestinal	1.33	10.3	6.61×10^{-06}
Hematological	1.01	19.5	3.35×10^{-26}
Immunological	0.85	26.6	1.23×10^{-11}
Metabolic	1.68	4.8	1.59×10^{-07}
Musculoskeletal	0.78	30.7	1.59×10^{-05}
Neurological	0.72	34.0	2.46×10^{-08}
Nutritional	1.29	11.1	1.44×10^{-23}
Ophthamological	0.84	27.1	4.61×10^{-11}
Psychiatric	0.98	21.0	2.98×10^{-07}
Renal	1.60	5.8	8.50×10^{-06}
Respiratory	0.86	26.3	4.25×10^{-10}

TABLE VII: **Semantic similarity significance in the biological process domain**. We consider categories with means within the top 20 percentile to be significantly closer than the domain average.

While the principles behind disease modules have been studied in recent years, little is known about how drugs interact with them. The formalisation of concepts and procedures to approach their drug target counterparts provides a computational basis to analyse this relation. Moreover, the study of drug targets can help in the understanding and characterisation of disease modules and functional gene modules.

The initial analysis of drug categories shows that while drugs targets from specific categories are modular, the mechanisms used to produce an effect do not strictly co-localise with the disease modules. Additionally, we find that the preferred target are proteins used for signalling, transport or binding processes. This suggests that the comprehension of disease modules might benefit from the usage of different biological networks, such as signalling networks. Drugs that have common functions are interesting for repurposing effects.

The analysis of multi target drugs could also be useful for drug repurposing, as their multiple targets can relate to different diseases. Multi target drugs are also interesting when the targets are close, since they target the same module. This further adds to the potential of drugs to enhance the disease modules.

Acknowledgment

This work was supported, in part, by the CONACYT Paraguay Grant 14-INV-088 and PINV15-315. We are thankful to our colleague Horacio Caniza who provided expertise that greatly assisted the research.

References

[1] Junker, Björn H., and Falk Schreiber. *Analysis of biological networks*. Vol. 2, John Wiley & Sons, 2011.

[2] Khatri, Purvesh and Sorin Drăghici. *Ontological analysis of gene expression data: current tools, limitations, and open problems* Bioinformatics 21.18, pp. 3587-3595, 2005.

[3] Bonferroni, C. E., *Teoria statistica delle classi e calcolo delle probabilita*, Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 1936.

[4] Yang, Haixuan, Nepusz, Tamás and Paccanaro, Alberto. *Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty* Bioinformatics 28 (10): 1383-1389, 2012.

[5] Goh, Kwang-Il, et al. *The human disease network*. Proceedings of the National Academy of Sciences 104.21: 8685-8690, 2007.

[6] Ashburner, Michael, et al. *Gene Ontology: tool for the unification of biology*. Nature genetics 25.1: 25-29, 2000.

[7] Barabási, Albert-László, Natali Gulbahce, and Joseph Loscalzo. *Network medicine: a network-based approach to human disease*. Nature Reviews Genetics 12.1: 56-68, 2011.

[8] Li, Lin, and Christian Hölscher. *Common pathological processes in Alzheimer disease and type 2 diabetes: a review*. Brain research reviews 56.2: 384-402, 2007.

[9] Hamosh, Ada, et al. *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders*. Nucleic acids research 33.suppl 1: D514-D517, 2005.

[10] Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. *DrugBank: a comprehensive resource for in silico drug discovery and exploration*. Nucleic Acids Res. Jan 1;34(Database issue):D668-72. 16381955, 2006.

[11] César A. Hidalgo, Nicholas Blumm, Albert-Lászl Barabási and Nicholas A. Christakis. *A Dynamic Network Approach for the Study of Human Phenotypes*. Plos Computational Biology 5(4), e1000353, 2009.

[12] Van Driel, Marc A., et al. *A text-mining analysis of the human phenome*. European journal of human genetics 14.5: 535-542, 2006.

[13] Caniza, Horacio, Alfonso E. Romero, and Alberto Paccanaro. *A network medicine approach to quantify distance between hereditary disease modules on the interactome*. Scientific reports 5, 2015.

[14] Yildirim, Muhammed A., et al. *Drug–target network*. Nature biotechnology 25.10: 1119, 2007.

[15] Prasad, T. S. K. et al. *Human Protein Reference Database - 2009 Update*. Nucleic Acids Research 37, D767-72, 2009.

[16] Caniza, Horacio, et al. *GOssTo: a stand-alone application and a web tool for calculating semantic similarities on the Gene Ontology*. Bioinformatics 30.15: 2235-2236, 2014.

[17] Evans, William E., and Howard L. McLeod. *Pharmacogenomicsdrug disposition, drug targets, and side effects.* New England Journal of Medicine 348.6: 538-549, 2003.