SN Computing manuscript No.
(will be inserted by the editor)

A method to explore the connectivity patterns of proteins and drugs for identifying disease communities

Ken McGarry* · David Nelson · Mark Ashton

Received: date / Accepted: date

Abstract Diseases are often caused by defective proteins, these proteins rarely operate in isolation and may have several roles in the cell. Thus over time a defective protein may be involved in several disorders, either directly or indirectly. The multiple roles leads to the concept of a disease module or cluster. This work describes how we generate overlapping clusters from complex networks to explore the dynamic nature of diseases, the genes implicated with them and the drugs used to treat them. Link clustering is vital for community detection as it enables the integration of disparate sources of data and provides a better understanding of community hierarchy and community dynamics than non-link methods. Furthermore, we view not just the genes directly shared between diseases but also indirectly connected genes in the network neighborhood. We use data and information from the STITCH protein and drug interaction databases, OMIM disease database, lists of diseases categorized by MeSH and the drugbank repository. The Gene Ontology, Disease Ontology and KEGG provide biological validity for the disease communities. We demonstrate how the detection of overlapping clusters enables the identification of biologically plausible communities consisting of cooperating proteins. We verify their role in disease with respect to targeting drugs more effectively with expert opinion. We have been able to identify various modules that make sense from a biological and medical perspective and validate drug repositioning candidates with clinicaltrials.gov.

Ken McGarry* and David Nelson
School of Computer Science, University of Sunderland, UK E-mail: ken.mcgarry@sunderland.ac.uk

Mark Ashton
Faculty of Medical Sciences, University of Newcastle, UK.

Keywords Link clustering · gene ontology · MeSH · disease modules

1 Introduction

In this work we describe our methods for building complex networks (graph theory) of protein interactions linked with various diseases which we then use to generate overlapping communities of diseases. Many real-world problems naturally lend themselves to be modeled by graph theoretic methods, which involve calculating statistical summaries from the pair-wise connections between entities or nodes [2]. The nodes are typically some important object, the links refer to particular relationships between them [42].

A natural extension of complex network/graph theory is the search for viable and plausible communities that may coexist in a given network. Rather than assign the nodes in a given network uniquely to a single group or cluster, it is more realistic and useful to search for all groups that each node may participate in [35]. Community detection from graphs/complex networks has attracted much research recently, notably in the social networking context [17, 16, 39]. Community detection algorithms have played a major role in developing a deeper understanding of diseases through protein interactions [30]. The so called *diseasome* has initiated a quantum shift in medical research, researchers are now tackling diseases with new insights into how the same proteins may be involved in many different diseases [14]. This in turn has led to a greater understanding of treatments and that drugs developed for one disease may be suitable for targeting at another seemingly unrelated disease

[3, 37] and this has led to a greater knowledge of drug side-effects on patients [4].

However, there are many challenges to overcome in community detection. The majority of community detection systems allocate each node to a single community, this is unrealistic because very few real-world entities can be uniquely identified with a single affiliation. In our approach we employ link detection to form overlapping clusters or communities. The advantage of detecting overlapping clusters (communities) enables the handling of different types of data as well as capture the hierarchy and community dynamics. Thus we are able to integrate protein interactions with disease associations and prior biological knowledge from ontologies.

We use the Medical Subject Headings (MeSH) database as a starting point for our investigations. This was developed by the National Library of Medicine and is a controlled vocabulary thesaurus, used for indexing articles for the MEDLINE/PubMED database. It is a controlled vocabulary (thesaurus), providing uniformity and consistency to the indexing and cataloging of biomedical literature. However, the relatedness of diseases is reflected in the MeSH hierarchy. The following list of 16 medical subject headings gives a broad overview of the categories:

- Anatomy [A]
- Organisms [B]
- **Diseases [C]**
- Chemicals and Drugs [D]
- Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E]
- Psychiatry and Psychology [F]
- Phenomena and Processes [G]
- Disciplines and Occupations [H]
- Anthropology, Education, Sociology, and Social Phenomena [I]
- Technology, Industry, and Agriculture [J]
- Humanities [K]
- Information Science [L]
- Named Groups [M]
- Health Care [N]
- Publication Characteristics [V]
- Geographicals [Z]

Of this list we are concerned with the Diseases [C] category, this is further broken down into 26 types of diseases shown in the list below. We are interested in the C06 category, as it represents the digestive system diseases.

- Bacterial Infections and Mycoses [C01]
- Virus Diseases [C02]
- Parasitic Diseases [C03]
- Neoplasms [C04]
- Musculoskeletal Diseases [C05]
- **Digestive System Diseases [C06]**
- Stomatognathic Diseases [C07]
- Respiratory Tract Diseases [C08]
- Otorhinolaryngologic Diseases [C09]
- Nervous System Diseases [C10]
- Eye Diseases [C11]
- Male Urogenital Diseases [C12]
- Female Urogenital Diseases and Pregnancy Complications [C13]
- Cardiovascular Diseases [C14]
- Hemic and Lymphatic Diseases [C15]
- Congenital, Hereditary, and Neonatal Diseases and Abnormalities [C16]
- Skin and Connective Tissue Diseases [C17]
- Nutritional and Metabolic Diseases [C18]
- Endocrine System Diseases [C19]
- Immune System Diseases [C20]
- Disorders of Environmental Origin [C21]
- Animal Diseases [C22]
- Pathological Conditions, Signs and Symptoms [C23]
- Occupational Diseases [C24]
- Chemically-Induced Disorders [C25]
- Wounds and Injuries [C26]

The C06 digestive disease category is further divided into eight main sub-categories, using two levels of annotations.

- Biliary Tract Diseases [C06.130]
- Digestive System Abnormalities [C06.198]
- Digestive System Fistula [C06.267]
- Digestive System Neoplasms [C06.301]
- Gastrointestinal Diseases [C06.405]
- Liver Diseases [C06.552]
- Pancreatic Diseases [C06.689]
- Peritoneal Diseases [C06.844]

The MeSH system is a tree based method of linking the general categories of disease all the way down to highly specific terms, potentially using up to 8 levels of numbers. The C06 group contain nearly 300 unique identifiable disease types. The MeSH system is a useful resource for structuring diseases into related groups but provides little in the way of indicating how they are linked to other groups.

A major factor in understanding the effects of diseases and their inter-relationships, is to appreciate the role and functionality of protein interactions [33]. The *interactome* defines the connectivity patterns of proteins revealing a complex pattern of relationships and associations. Recently, researchers applying machine learning methods such as clustering to reveal the multi-functionality and overlapping activities of proteins as they cooperate in various functions [7, 27]. Other researchers have applied complex networks or graph theory methods to construct and analyze protein interaction networks, or have implemented predicative algorithms for identifying protein function [28] and in particular identifying interesting sub-graphs using a combination of clustering and classification methods has received increased interest [22, 19].

A high degree of heterogeneity exits both in genes and disorders, in particular some diseases are implicated with a small number of genes, while some medical problems such as colon cancer and deafness have been associated with more than thirty genes [30]. Highly connected hub genes play an important role, as diseases appear to be correlated with them causing several health problems when they fail in their cellular functions [12]. Considering human diseases as a network (diseasome) is a relatively new way of assessing how diseases and comorbidities occur and the relationships between them [8]. However, it is

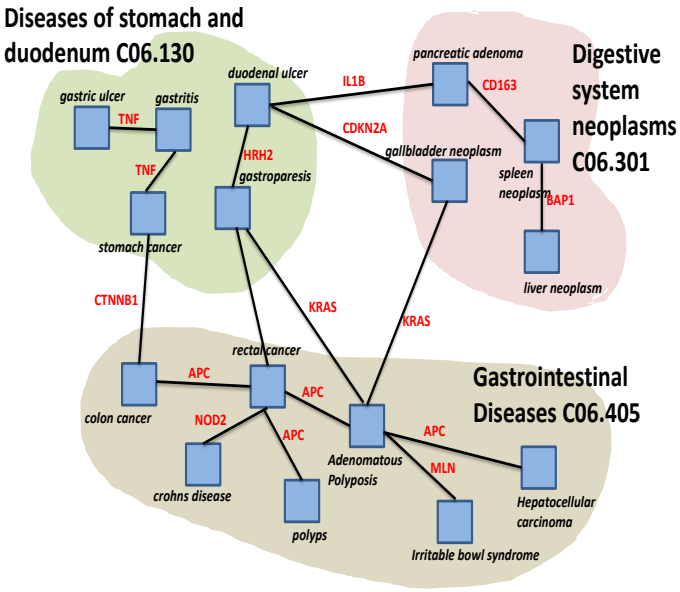


Fig. 1: Highly simplified disease network showing only three of the eight classes of Digestive System diseases, each square node indicates a specific disease linked to the others by at least one shared protein.

novel and though ill-defined is generally understood to be a method to tackle diseases at a more holistic level [15]. The computational challenges are great but so are the opportunities for deeper insights to develop novel drug products and to potentially reposition existing drugs to new targets [15].

Our work builds upon recent discoveries on the modular nature of protein networks where there is overlap or crosstalk of function [9]. We take these ideas forward and define the notion of overlapping disease modules with shared proteins and protein modules all cooperating in various cellular processes. Figure 1 highlights a small fraction of the linkage patterns of only one category of disease (Gastrointestinal diseases) [13]. Initial work conducted by Goh devised a network using mutation information from the OMIM database which linked pairs of diseases when any mutation of a gene was identified in both disorders [14].

Numerous issues and challenges are involved in generating disease networks, the main problem is the sheer size of the human genome in which there are perhaps 25,000 genes expressing proteins of which 10% are implicated to a disease [6]. In addition, about 1,000 metabolites, compounds and chemicals can interfere and alter the functionality of the proteins they interact with. The other issue, which we address is the identification and cooperation of clusters of genes in the disease network that have more than one role in the cell [43].

In figure 2 we present an overview of our system, on the left of the diagram we have five databases providing information to construct drug similarity structures, networks of disease implicated genes, networks of neighboring genes not directly linked to the diseases, lists of diseases and their similarities based on ontological terms. The core or inner shell of the system is the known linkages between proteins and specific "C06" diseases. These "C06" proteins link out to the outer shell where non"C06" diseases are linked to them, these are shown as colored areas with dashes. The yellow squares refer to proteins known to be implicated in these non"C06" diseases. Thus a chain of interacting proteins may interlink diseases with each other. The outputs from the analysis are overlapping disease modules and perhaps more usefully (but tentatively) are the drug modules - whereby we hope to repurpose a drug targeted at one disease to another seemingly unrelated disease.

Although explained in greater detail in the methods section it will be useful to enumerate the data sources.

- DrugBank - contains 13,536 drug entries including 2,630 approved small molecule drugs, 1,372 approved biologics .
- MeSH - Medical Subject Headings of all known diseases. Useful, for structuring and categorizing known medical relationships.
- OMIM - Online Mendelian Inheritance in Man, contains lists of all the genes known to be implicated in diseases.
- STITCH - Protein to protein pairwise interactions and interactions between proteins and chemicals. Allows us to link disease implicated proteins with the rest of the protein interaction network.
- CHemBl - CHemical database of Bioactive molecules with drug-like properties.

We define a disease module (or community) as a function of the inter-connectivity patterns between disease implicated genes and genes in their neighborhood. Furthermore, the density of the modules represents a subgraph within the larger connected component. Thus, it is important to note that in our system any specific disease may produce several disease modules that interact with and overlap with other diseases. Thanks to graph theory, systems biology and the growing body of medical evidence at the level of genes/proteins, science is now starting to reevaluate the definitions of disease and to look at the interconnectedness of disease [7]. We use the Medical Subject Headings (MeSH) as a framework for relating the interconnectedness of diseases because

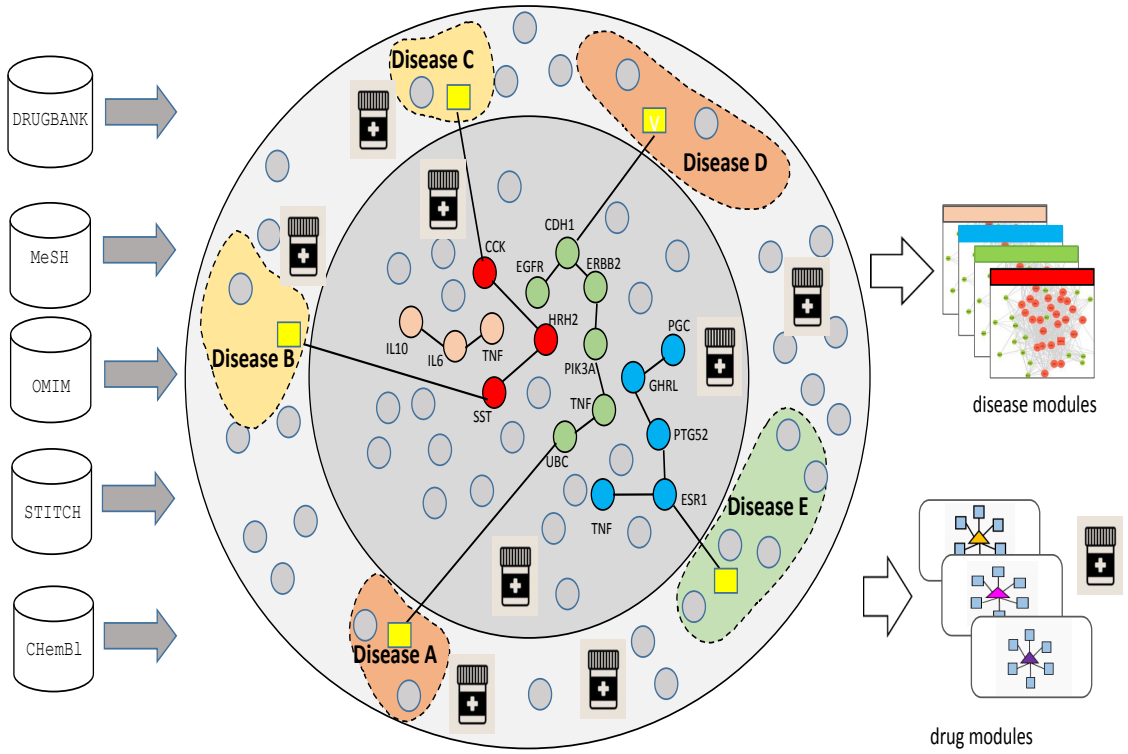


Fig. 2: Disease module system operation, indicating data and processing. The inner circle or shell contains the C06 related diseases linked by shared genes to other C06 diseases and some non-C06 diseases. The outer circle or shell holds the non-C06 diseases (A to E) that are linked indirectly by neighborhood genes. Using this information of connectivity including the known drug treatments we can deduce co-morbidities and disease module structure.

this is how the doctors and the medical profession view diseases. The MeSH system thesaurus is a controlled vocabulary originally intended for searching PubMed based on annotations, it is a natural way to express the hierarchy of diseases and sub-diseases. This does not mean it is the best or most accurate method, especially now that we think the *diseaseome* approach is likely to be superior but until it gains more acceptance the MeSH framework is useful.

2 Related work

Previous systems investigating overlapping modules have been developed such as the CIDeR network by Lechner et al [21]. CIDeR is a manually curated knowledge base of protein interactions between disease-related elements such as biomolecules and biological processes and phenotypes. The aim of CIDeR is to serve as a one-stop source for bioinformatics applications. The database contents were developed by examining publications from the biomedical literature and transferring the appropriate information from experimental sources into a structured form that can be processed by computational approaches. In addition, important information, such as the cell type used in experiments, is described because there are differences in the cellular processes in different tissues. CIDeR contains many of the most common diseases but is not an exhaustive list of resources given its manual nature.

The system developed by Ghiassian, detected the connectivity patterns of proteins associated with diseases DIseAse MOdule Detection (DIAMOnD) [12]. DIAMOnD was based on a number of assumptions: that although the topology of a network represents functionality it cannot capture the essence of a disease module; that disease implicated proteins have distinct and predictable interaction connectivity patterns that can be used to determine disease modules; that the significance of disease proteins connections are more important than simply considering the number (density) of connections. Ghiassian considered over 70 diseases and identified the key disease proteins how they cooperate in disease modules. What the DIAMOnD system lacks is the ability to rank the

discovered disease modules in a biologically meaningful way.

The algorithm designed by Yu formulates the creation of modules by estimating the distances between functional blocks of proteins only the basis that knowledge of the human protein interactome is very incomplete and biased with a great deal of uncertainty present [48]. Therefore, Yu built a probabilistic approach into module estimation. The final goal was to reposition drugs for diseases that were linked to the primary seed disease. The modules are formed from drug-protein pairs and all are related to cancer specific problems. For each and every disease module formed, the distance metric when overlaid on the drug network identified several candidate drugs. Yu then confirmed these drugs were viable candidates by accessing clinicaltrials.gov which indicated they were currently undergoing repurposing trials.

The MBiRW algorithm designed by Luo et al, employed a bi-random walk that assessed the similarity of diseases and the drugs that could potentially act as therapies for the disease implicated proteins [25, 26]. The MBiRW algorithm utilized a number of similarity measures employed against gold standard data which provided a degree of validation. However, MBiRW does not use target information effectively and cannot include biologically relevant knowledge. The CommWalker method developed by Luecken is another random walk method that samples proteins allocated to its disease modules but also used GO annotation to improve biological plausibility of the generated modules [24]. To reduce bias and improve accuracy, every module created from the random walk was assessed by three different link-analysis algorithms. Each random walk was terminated when they reached a dynamically determined cut-off value. For each step of the random walk the functional GO annotations were averaged to determine the module homogeneity, thus allowing each module to be ranked in terms of biological plausibility.

Other approaches such as the Ravasz algorithm analyzes the pair-wise connectivity patterns of nodes and is part of the agglomerative hierarchical clustering class of algorithms [38]. It uses four stages to calculate community membership, node similarity must be substantially greater for node-pairs belonging in the within the same cluster or community and low values for node value pairs that are assigned to other communities. Thus communities were grown from the Ravasz algorithm. The opposite approach was taken by Girvan and Newman who used a modified divisive algorithm that systematically removed links and thus broke the network into separate communities [34].

Usually, clustering nodes is performed to create their network structure, unfortunately using this approach each node belongs to a single cluster or group. This can be counter-intuitive to the way nodes or entities operate in the real-world. A better approach is not to cluster the nodes themselves but rather the links between them indicating their coexistence with other clusters. Several algorithms have been proposed along these lines [1], these methods typically cluster the links connecting the nodes indicating their involvement with other groups. The overlapping links connection algorithm developed by Ahn [1] employs the Jaccard similarity measure for determination of link suitability to be clustered together.

Work by Dissez *et al* considers the use of non-negative matrix tri-factorization (NMTF) to integrate data from several sources for drug repositioning [10]. Non-negative matrix factorization (NMF) has been successfully used across several domains that require the integration of heterogeneous and disparate data such as community detection within complex networks [49, 29, 23]. However, the NMTF approach produces three matrices which fuse the data together in a robust manner and appears to have superior performance over NMF applications. NMTF uses a series of hyper-parameters are fine tuned over a series of iterations, the system can predict missing links and identify candidate drugs for consideration.

Methods

Data and knowledge sources

A number of databases and ontologies were accessed. The chemical structure of the drugs was obtained from the NCBI in SDF (structure data files) format. These consist of a series of molfiles joined together, together with some further information about the compounds. They are frequently used for sharing libraries of compound structure data. [45]. For each drug a fingerprint was created consisting of an atom-pair arrangement of 1024 bits, this is effectively a matrix-like representation where every molecule is encoded as a fingerprint of the same type and length. The presence or absence of a particular structure is represented by a binary bit, either ‘1’ or ‘0’. The chemical structures are used to assess drug similarity and role in treating the disease network, potentially identifying drugs for repositioning based on their similarity to other drugs.

For each drug we identify their on-target proteins and also their off-target proteins. We enhance this information using drug-to-protein interactions and protein-to-protein interactions residing in the STITCH and HINT

databases [32]. The STITCH database contains use of over 6,000,000 protein-to-protein interactions annotated by experiments from the literature and through text mining. The HINT database contains high quality chemical to protein associations annotated by experts and supports any shortfalls in the drug data residing in Drugbank, which often is not complete in identifying every known drug-to-protein association.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a highly respected database that provides biological pathways involvement for any gene or protein, here we use interactions associated with on-target and off-target proteins known to be influenced the drugs [46]. Pathway involvement often gives insights into the biological activity associated with genes or proteins of interest. For example, deeper insights into the pathogenesis of neonatal sepsis was recently discovered by using pathway information [31]. Gene ontology (GO) is a useful resource for highly detailed biological information for gene product annotation. The information is available in a hierarchy, leading downwards from generic terms to the highly specific for cellular function, molecular function and biological process.

Annotating gene products with GO ensures an element of biological plausibility instead of risking potentially spurious or random correlations. For every protein residing in a disease module we performed GO enrichment, using similarity measures based on information theoretic algorithms. The annotation information is assessed by calculating the negative log and probability of the term t annotated to the proteins. There are many measures available to assess semantic similarity, however we use the *Wang* measure because it gives more credence to biological similarity than most measures since it uses the positions of each term in the GO structure but also the association and hierarchical level with previous ancestor terms [44]. In equation 1, the details of the Wang equation for two terms A and B:

$$Wang(A,B) = \frac{\sum_{t \in T_A \cap T_B} S_A(t) + S_B(t)}{SV(A) + SV(B)} \tag{1}$$

Where: $S_B(t)$ is the Similarity-value of the gene ontology term t related to term B and $S_A(t)$ represents the Similarity-value of gene ontology term t related to term A. The term SV is the semantic value of GO terms A and B. The locations and semantic relations between ancestors of the A and B terms in the GO graph needs to be taken into account by $t \in T_A \cap T_B$, this represents a major advantage of the Wang method over previous techniques.

3 Complex networks

Given a list of pair-wise connections between nodes, graphs can be constructed through the creation of an adjacency matrix which specifies the connections for the entire network. A graph can be defined as G = (V,E) where the nodes or vertices (V) are linked to other nodes via edges (E). The most efficient data structure to hold the connectivity information is the adjacency matrix **A**. For graph $G = (V, E)$ this entails $N_v \times V_v$ such that **A** as defined by Kolaczyk [20]:

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E, \\ 0 & \text{for no connections between nodes.} \end{cases} \tag{2}$$

Where: for indices (i, j) represent the vertices V in the graph G connected by an edge E, from i to j, A is non-zero for those instances and zero for no connections. Once graph G has been generated a number of statistics can be applied. The process of detecting the communities residing in a complex network requires an analysis of the structures and characteristics present [39]. The challenges are difficult, since it is by no means clear how many communities actually exist in a complex network [36]. The algorithms used must be able to avoid false positives or noise in the linkage patterns.

4 Defining disease communities through link clustering

We view the concepts of module and community as synonymous, generally accepted definitions of a community include the notion they are a locally *dense* set of connections that form a cohesive subgroup or subgraph [5]. Specifically, we can state that every member of a community should link to other members of that same community with a much higher probability than members belonging to other communities.This is possible through link clustering enables the detection of overlapping clusters. Cluster similarity or overlap can be calculated using the Jaccard score, shown in equation 3. We take this approach further by calculating measures for node centrality through analysis of the weights localized by each community based on their pair-wise similarity [18].

$$S(e_{ik}, e_{jk}) = \frac{\mid n + (i) \cap n + (j) \mid}{\mid n + (i) \cup n + (j) \mid} \tag{3}$$

Similarity is determined by the jaccard coefficient through checking node links, e_{ik} and e_{jk} which reveal

shared nodes k. For each node the neighborhood of i, similarity is determined by observing the next node $n+(i)$, once coefficient values are determined $n+(i)$ can be clustered. The cluster dendrogram determines a cut-off point where the link density is maximized between all shared (overlapping) communities.

Community measures such as that shown in equation 4 measures the weighting of the clusters or communities using each groups pair-wise similarity, allowing the node centrality to be calculated.

$$CC(i) = \sum_{i \in j}^{N} \left(1 - \frac{1}{m} \sum_{i \in j \cap k}^{m} S(j,k) \right) \tag{4}$$

Where: j and k denotes the similarity between two communities and is defined by the function $S(j,k)$ which is calculated by taking the Jaccard measure for the sum of nodes that are shared between the two clusters. N represents the number of clusters detected with respect to each node i in the network.

Each of the discovered communities will have a certain amount of *coverage*, this refers to the nodes belonging to nontrivial communities. The amount of overlap coverage, can be difficult to determine because methods with similar cluster coverage can calculate different amounts of overlap. Determining nontrivial communities is also problematic, since a small number of nodes may be a viable community in some applications but not others.

The criteria or cut-off point we used to build and calculate the rating of disease modules is taken from previous research discoveries [12]. We conjecture that disease implicated genes and essential genes will be encoded as hub proteins with numerous connections, furthermore simple network topology is unlikely to successfully identify disease modules [41]. In algorithm 1 the details are presented for disease module discovery and assessment.

In algorithm 1, lines 2-9 initialize key values: the UMLS code for the disease of interest, the BP (biological plausibility) cut-off point was discovered empirically and for every disease module that is generated it must score a value of 4.5 or greater to be retained, else it is discarded. Each disease module will have a score and they are later ranked in terms of importance.

Lines 12-19 create a hierarchical, meshtree structure that is used to hold the C06 diseases, after the main term C06, there are eight next levels terms (i.e. C06.130, C06.405 etc) until the 8th level terms like C06.405.117.119.500.484.500.500 that identifies "Respiratory aspiration of gastric contents" are reached. There are 299 terms in total for Digestive System Diseases. Further processing assigns drugs to each disease (if any) from DRUGBANK database, identifies any known disease genes for each disease from OMIM. The second shell proteins are identified using the HINT database. For each drug, the SDF file containing the chemical structures are downloaded from CHeMBL and binary fingerprints (1024 bits) are generated.

Lines 21-25 transform the lists of genes, drugs and diseases into pair-wise interaction lists that are suitable for graph functions and link analysis functions to operate on. Lines 27-34 build disease modules using equations 3 and 4. They perform tests for biological relevance on these modules and rank them using the gene ontology resource, this ensures the modules have some basis in biological fact and are not a random collection of genes or drugs. Lines 35-38 merge the disease modules based on a biological overlap of 75% or greater, linking diseases of the same type and also different diseases.

In algorithm 2 we expand the details of the merging process, where *Proteins A* refers to the 1st C06 disease module and *Proteins B* refers to the next disease module. Starting with the C06 diseases we merge modules within this group with any non-C06 disease that has sufficient protein interaction similarity based on the cosine measure. If the modules (from any disease) also have a biological similarity of 75% or greater with jointly shared proteins, these are merged into a single module. This removes weak, noisy and potentially false positive modules from the list generated by the link analysis process. Details are kept of the merged diseases via the key implicated proteins and updated statistics.

The C06 disease modules are annotated with gene ontology and KEGG for biological plausibility using the ClusterProfiler package [47]. They are also compared with localized non-C06 disease modules, the comparison gives insights into the potential relatedness of many disorders. We use an adapted form of the Cosine similarity measure, originally developed for information retrieval, but it is finding increased use in bioinformatics applications [40].

$$cos(x,y) = \frac{x \cdot y}{||x|| \cdot ||y||} \tag{5}$$

The Cosine similarity function compares the structure of two vectors without respect to the numerical magnitude. The dot product produces a scalar (of two vectors) which is normalized as function of their lengths. The Cosine function will output values close to unity indicating high similarity and values close to zero indicates low similarity.

Algorithm 1 Identification of disease modules

procedure SeekDiseaseModules($DRUGBANK, HINT, OMIM, CHeMBL, MeSH, GO$) ▷ Databases used
  do initialize
    $umls \leftarrow$ get UMLS code for disease ▷ e.g. C06 for Digestive System
    $DiseaseModList = 0$ ▷ set to zero, no modules yet
    $drugstructure = 0$ ▷ no drug structures found
    $drugs = 0$ ▷ no drugs found
    $disgenes = 0$ ▷ no implicated genes found
    $shell2 = 0$ ▷ no 2nd shell genes found
    $BP = 4.5$ ▷ Biological plausibility cut-off
  end initialize

  $MeshTree \leftarrow MeSH[umls]$ ▷ Generate a meshtree structure for this disease
  $meshCount \leftarrow len(MeshTree)$ ▷ How many sub diseases do we have in C06?
  for $i \leq meshCount$ **do**
    $druglist[i] \leftarrow DRUGBANK[i]$ ▷ get known drug treatments for each disorder
    $disgenes[i] \leftarrow OMIM[i_X]$ ▷ get the known disease genes
    $drugstructure[i] \leftarrow CHeMBL[i]$ ▷ get chemical structure if available
    $shell2[i] \leftarrow HINT[i]$ ▷ get 2nd shell genes and drugs
  end for

  do build_edgelists ▷ pairwise lists of interactions to build graph networks
    $drug2drug_{el} \leftarrow convert[druglist]$
    $disgenes_{el} \leftarrow convert[disgenes]$
    $shell2_{el} \leftarrow convert[shell2]$
  end build_edgelists

  do build_linkcommunity ▷ Disease module construction and assessment
    $mainlist_{EL} \leftarrow merge[drug2drug_{el}, disgenes_{el}, shell2_{el}]$
  if $GOannotate \geq BP$ **then** ▷ Gene Ontology annotation assesses biological meaning
    $DiseaseModuleList[i] \leftarrow mainlist_{EL}$ ▷ add to list of modules
    $DMstatistics \leftarrow CalcStatistics[DiseaseModuleList]$ ▷ calculate module statistics
  end if
  end build_linkcommunity

  do Merge_DiseaseModules ▷ Merge disease modules into coherent meta-modules
    $Overlap \geq GOannotate$ `at 75%`
  end Merge_DiseaseModules

  return $DiseaseModuleList, DMstatistics$ ▷ Return disease modules and statistics
end procedure

Algorithm 2 Merging functional disease modules

procedure MergeDiseaseModules(C06)
  for *all modules 1:i* **do**
    $\forall ProteinsA\, \exists ProteinsB : usingCosine$ ▷ as in eqn 5
    if $GOannotateModule_i \geq 75\%$ **then**
      $MergedModuleList \leftarrow$ add module
      $MergedModuleList \leftarrow$ add module statistics
    end if
  end for
  return $MegedModuleList$
end procedure

Software availability

The analysis and data processing, along with graphics was performed using the R language with the RStudio development environment. All R code and data files that generate the tables, diagrams and functional code presented in this paper can be downloaded from GitHub for download:
https://github.com/kenmcgarry/Disease-Modules

5 Results

The MeSH database was searched using "C06" as the root starting point of "Digestive System Diseases". This returned a list of 299 related disorders in a hierarchy based on the eight level coding system, table 1 highlights the first 15.

The "ID" column is the unique identifier used to access the DrugBank database to return a list of known drugs to treat each C06 disease. This produces a list of 194 known drugs and treatments, in table 2 a small number of drugs used to treat Gastrointestinal disorder are shown. The drug target genes are used as search patterns to access HINT database to provide a list of known protein interactions between these genes. Unfortunately, only 189 drugs have chemical structures available in SDF format to analyze further.

In figure 3 we show the results of applying the Tanimoto similarity score to the 189 available drugs used to treat C06 disorders and their chemical fingerprints (each binary 1024 bits in length) indicating the presence or absence of various structural keys or subfragments. The aim of clustering and similarity matching is

Table 1: List of 15 C06 disorders, concentrating on Gastrointestinal disease (C06.405), there are 299 in total and arranged in a hierarchy, the top levels are general and the lower levels are highly specific.

MeSH	ID	Term
C06	D004066	Digestive System Diseases
C06.405	D005767	Gastrointestinal Diseases
C06.405.117	D004935	Esophageal Diseases
C06.405.117.119	D003680	Deglutition Disorders
C06.405.117.119.500	D015154	Esophageal Motility Disorders
C06.405.117.119.500.204	D017675	CREST Syndrome
C06.405.117.119.500.432	D004931	Esophageal Achalasia
C06.405.117.119.500.450	D015155	Esophageal Spasm, Diffuse
C06.405.117.119.500.484	D005764	Gastroesophageal Reflux
C06.405.117.119.500.484.500	D057045	Laryngopharyngeal Reflux
C06.405.117.119.500.484.500.500	D063466	Respiratory Aspiration of Gastric Contents

Table 2: 10 drugs used to treat C06 Gastrointestinal disorders. Note, only one gene target is illustrated and each drug may have several in addition to treating several disorders.

	Drugbank_id	Drugbank_name	Target Type	Gene Target	Disorder
1	DB00152	Thiamine	Enzyme	TKTL1	Beri-Beri
2	DB00213	Pantoprazole	Transporter	SLC22A2	Systemic Mastocytosis
3	DB00424	Hyoscyamine	GPCR	CHRM2	Irritable Bowel Syndrome
4	DB00443	Betamethasone	Nuclear hormone receptor	NR3C1	cutaneous T-cell lymphoma
5	DB00572	Omeprazole	Enzyme	CYP2C19	Gastritis
6	DB00586	Mesalazine	Enzyme	ALOX5	Crohn's disease
7	DB00620	Pralatrexate	Enzyme	DHFR	Mucosal inflammation
8	DB00630	Moxifloxacin	Enzyme	gyrA	Hepatic cirrhosis
9	DB00635	Prednisone	Transporter	SLC28A1	Inflammatory bowel disease
10	DB00710	Erythromycin	Ion channel	KCNH2	Amoebic colitis

devise a list of drugs that may be potentially repositioned (based on chemical similarity) for disorders within C06 class. Furthermore, we use these similarities to explore the drug space of non-C06 disorders for any potential two-way repositioning [29]. This involves a similar process of downloading SDF data and building chemical fingerprints for the non-C06 drugs. The non-C06 drugs are identified at a later stage when we compare the disease modules.

In fig 3 the similarity of the drugs is shown as a heatmap, this matrix of 189 x 189 (there are 189 drugs in total) holds scores from zero to unity, indicating minimum or maximum similarity, and is encoded by yellow and dark blue areas respectively. There are a small number of well defined clusters along with patches of smaller clusters. The chemical similarities are important as seemingly different drugs may be applied to disorders other than they were designed for. In fig 3 a silhouette plot shows the goodness of cluster fit where cluster coefficients approaching unity suggests the observation is a good match for the cluster, while values approaching zero indicate the observation would better fit into a different cluster.

The cut-off criteria for cluster goodness of fit:

- 0.71-1.0 - Structure is very strong and plausible
- 0.51-0.70 - Structure is valid and strong
- 0.26-0.50 - Weak structure possibly artificially introduced
- < 0.25 - No structure is evident

We relate the drug clusters to the C06 community/disease module structures, investigating how the importance of their chemical similarity is related to MeSH structure and the shared protein targets. Later we cluster the drugs linked to diseases that are not C06 labeled, and compare their chemical similarity to the C06 drugs. The intention is to identify potential candidates for drug repositioning - either C06 to non-C06 diseases and vice-versa. We check the validity of this approach by viewing the research literature and ClinicalTrials.Gov for evidence.

The next stage was to create three networks of communities by using the link clustering equations 3 and 4, the three networks comprise: a). for the drug interactions, b). for 1st shell protein interactions, c). for 2nd shell protein interactions. The 1st shell protein interactions are direct connections between the disease proteins and target proteins. The 2nd shell interactions are not directly connected but are in the neighborhood of these interactions. Within each community we generate many disease modules.

In figure 4 the validation statistics of the three networks are displayed, the information presented in the nine graphs indicates the overall structure and size. The 1st column is the 1st shell genes, 2nd column is the 2nd shell of genes, third column is the drug network. For each column, there are three plots, the first graph in red indicates the community modularity versus the community connectedness. This graph represents how tightly the modules are coupled and indicates the community

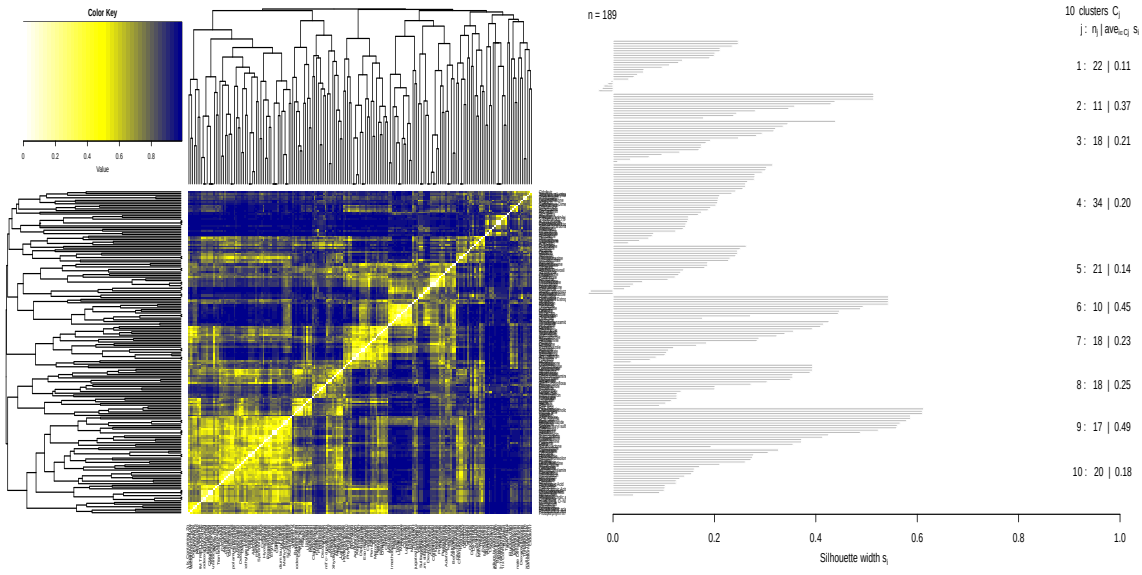


Fig. 3: Clustering the 189 drugs treating the C06 disorders based on chemical structure similarity.

modularity with respect to the number of links shared within the community versus links external to the community. The opposite of modularity is community connectedness.

The second plot in blue, indicates the clustering threshold and partition density, for all three communities the threshold is very similar at about 0.5 to 0.8, indicating that structurally where the resulting dendrogram can be cut at a point that optimizes the clusters density of links, it also takes into account normalizing the maximum and minimum number links attached to each cluster. It is determined automatically by our algorithm.

The third graph in green identifies membership criteria, where each node has a weighted membership value calculated from the community by how unusual that community is compared with the other communities in which that exact node resides versus the number of genes or drugs. Nodes that coexist within many dissimilar communities will obtain significantly larger scores of community centrality. However, those nodes belonging to communities that are highly nested or overlapped, or belong to only a few communities will get smaller scores.

The drug network consists of 189 drugs and associated target proteins along with other drug-to-drug interactions. In total there are 266 nodes and 2,482 interactions between them, forming 122 communities or disease modules. The maximum partition density was 0.31 and the number of nodes in largest cluster was 32. The inner shell network formed a network of 873 nodes with 15,787 interactions forming 509 communities or modules. The partition density was 0.17 with the largest cluster containing 120 nodes. The outer shell network consists 204 nodes with 4,178 interactions forming 106 communities. The maximum partition density is 0.55 with the largest cluster containing 73 nodes. The inner and outer networks are purely protein to protein interactions and can be formed into viable disease modules by data processing and ontology annotations, the drug network cannot be annotated since drugs are not part of gene ontology. However, the drug network is used to provide potential therapeutic effects based on chemical similarity and targets.

In acknowledging recent discoveries, we have pruned our list of disease modules to remove any module with fewer than 20 genes [30]. Percolation theory of networks predicts that any disease module with fewer than 20 genes is too fragmented and unlikely to be observed and reflect the biological reality. The inner shell and outer shell disease modules were reduced respectively from 509 to 83 and from 106 modules to 13 as a result of pruning. The majority of the modules having few genes and likely reflect noise.

Referring to the algorithm, the next stage was annotate the disease modules with gene ontology terms. The biological plausibility method of using gene ontology was inspired by Gamalielsson and the templates made from the binary relationships formed from the terms [11]. The disease modules are validated by the enrichment process, that is to say the use of gene ontology and KEGG ensures the disease modules correspond to logically, coherent and plausible biological activities.

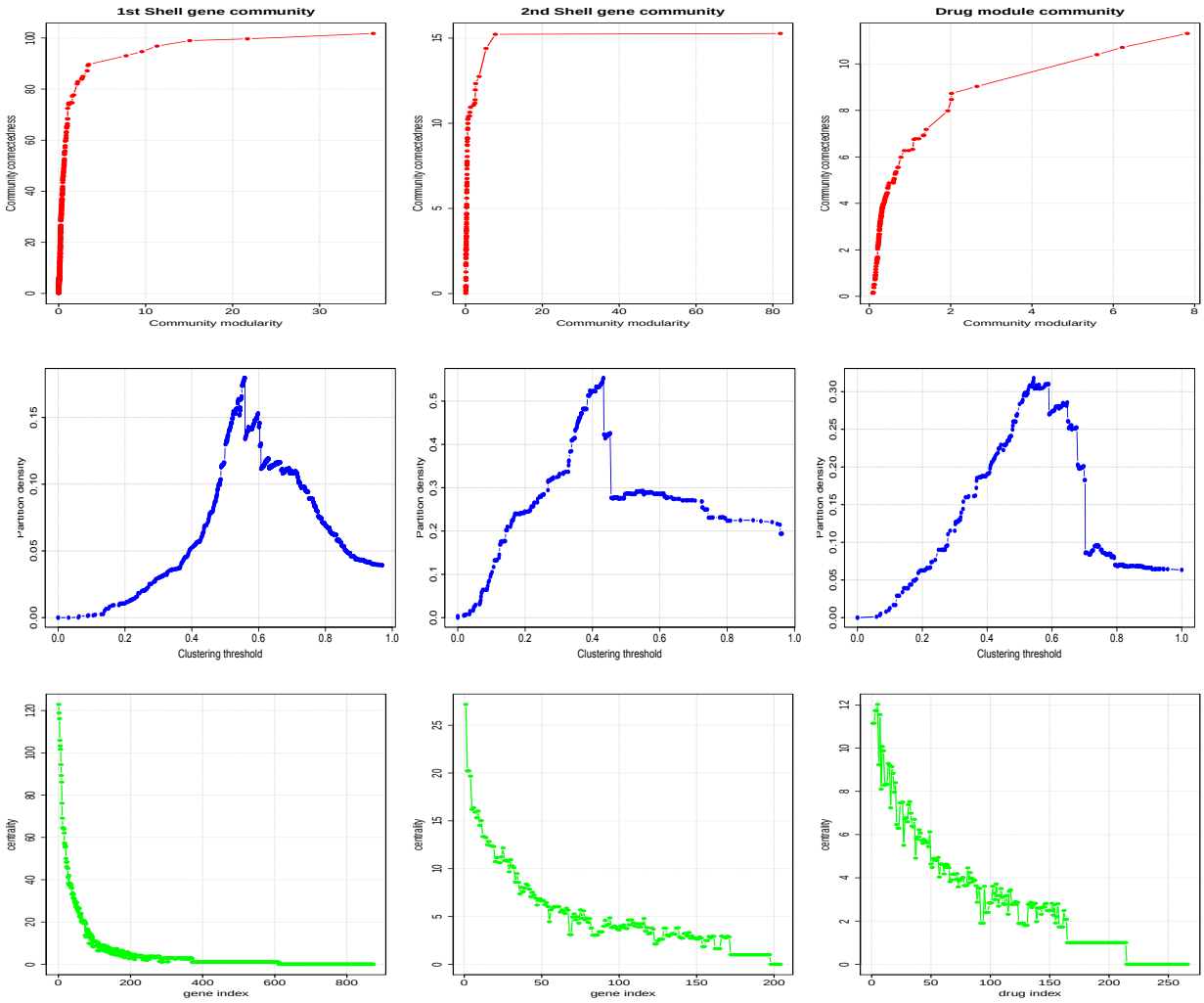
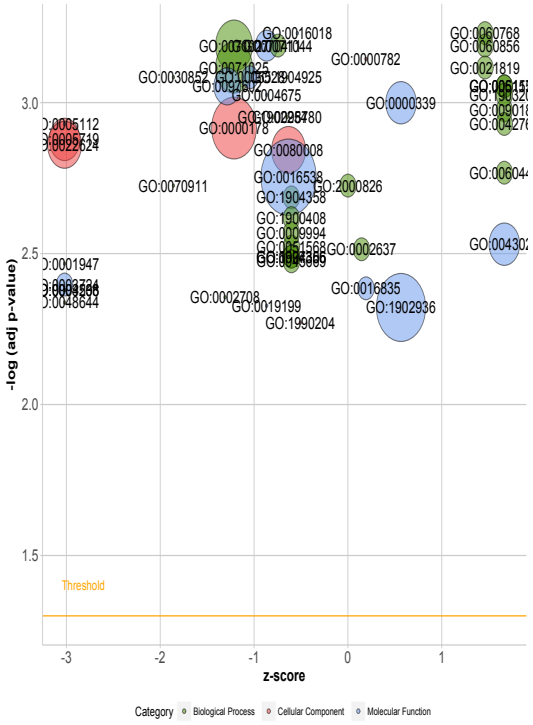


Fig. 4: Validation statistics for the three disease modules: 1st column is the 1st shell genes, 2nd column is the 2nd shell of genes, third column is the drug network. The top figure indicates the number of connections per disease module, the middle diagram shows the splitting criteria (partition density) for generating the modules. The bottom diagram shows the centrality statistic for each gene and drug.

The bubble plot in figure 5 represents the annotation of a specific module located in the 2nd shell disease network. The bubble plots become cluttered and difficult to read for more than one disease module however we can see that the three components of Gene Ontology the biological process, the cellular compartment and the molecular function (BP, CC and MF) are well represented and give important information visually, based on the size of the bubbles (i.e. terms that are highly represented) and color (green, red and blue respectively), The biological process terms are the most numerous indicating an active set of disease modules. Reducing the number of redundant terms, the readability of plots improves considerably but still maintaining the biological information. Any GO term with a gene overlap greater than or equal to a set threshold (0.75), the process keeps one term per group as a representative but does not take into account the GO hierarchy.

In table 6 the 25 top scoring disease modules are listed, ranked by their biological plausibility which is based on the gene ontology annotations and the other information based on equation 4. Those modules that link up with non C06 diseases are particularly interesting since any treatments may have potential for repurposing for C06 disorders. We performed a literature search and checked clinicaltrials.gov website to check the validity of our approach. We indicate in the appropriate column in table 6 where drugs have been identified by our system



GO:1902254	negative regulation of intrinsic apoptotic signaling pathway by p53 class mediator
GO:0080008	Cul4-RING E3 ubiquitin ligase complex
GO:0016538	cyclin-dependent protein serine/threonine kinase regulator activity
GO:0097306	cellular response to alcohol
GO:0030852	regulation of granulocyte differentiation
GO:0098780	response to mitochondrial depolarisation
GO:0060768	regulation of epithelial cell proliferation involved in prostate gland development
GO:1902936	phosphatidylinositol bisphosphate binding
GO:1990204	oxidoreductase complex
GO:0071025	RNA surveillance
GO:0004536	deoxyribonuclease activity
GO:0090185	negative regulation of kidney development
GO:0003724	RNA helicase activity
GO:0043022	ribosome binding
GO:0051568	histone H3-K4 methylation
GO:0005528	FK506 binding
GO:0061154	endothelial tube morphogenesis
GO:0070411	I-SMAD binding
GO:0042762	regulation of sulfur metabolic process
GO:0022624	proteasome accessory complex
GO:0000339	RNA cap binding
GO:0002708	positive regulation of lymphocyte mediated immunity
GO:1904356	regulation of telomere maintenance via telomere lengthening
GO:0051571	positive regulation of histone H3-K4 methylation
GO:0001947	heart looping
GO:1904925	positive regulation of mitophagy in response to mitochondrial depolarization
GO:0004675	transmembrane receptor protein serine/threonine kinase activity
GO:0019199	transmembrane receptor protein kinase activity
GO:0070911	global genome nucleotide-excision repair
GO:0060856	establishment of blood-brain barrier
GO:1900408	negative regulation of cellular response to oxidative stress
GO:0045669	positive regulation of osteoblast differentiation
GO:0003208	cardiac ventricle morphogenesis

Fig. 5: Bubble plot of gene ontology annotation of 2nd shell disease modules. The statistical significance threshold for the -log p-values is 1.5, hence all GO terms are significant

as potential candidates for repositioning with C06 disorders, where the drug has actually undergone/ongoing clinical trials is marked by a tick.

6 Discussion

We find that the C06 diseases were grouped at a high level (level 2) since the majority of the C06 disorders do not have genes associated with them. Out of the eight main sub divisions, because of our strict criteria, only four could provide viable disease modules: C06.130 (biliary tract diseases) with 17 disease modules, C06.198(digestive system abnormalities) with one disease module, C06.301 (gastrointestinal neoplasms) with six disease modules and finally C06.552(liver diseases) with 29 disease modules. Giving 53 disease modules in total for the C06 disorders. Linking the C06 modules through shared genes we identified 30 diseases that were related (by shared genes - their phenotypic manifestations can be quite different).

The highest scoring in terms of similarity and overlap were: Alzheimer's with two modules, Asthma with 11 modules, Autism with 21 modules, Diabetes with 17 modules, Hypertension with 17 modules, non-small-cell-carcinoma with one module, Obesity with 26 modules, Parkinson's with 13 modules, Rheumatoid Arthritis with 15 modules and Schizophrenia with 15 modules. The grand total is 194 disease modules, many with high degrees of similarity which need to be pruned using biclustering. Through the setting the parameters for the BiMax clustering algorithm we were able to remove very similar/redundant disease modules.

From table 6 it is apparent that diseases in the same module show significant comorbidity, the exact strength of comorbidity is difficult to express since we do not have individual patient data and cannot calculate relative risks or other statistical methods but can calculate a correlation coefficient based on the disease ontology mappings.

Using the Cosine statistic we are able to examine the modules in terms of disease module overlap (similarity), generally in terms of GO annotation the closer the overlap between modules then the biological similarity is greater. Recall, we are investigating C06 diseases in relation to their neighbors, the majority of our modules are overlapping with only a few that are isolated or separated. Taking the entire diseasome into account Menche found that only 7% of disease pairs had overlaps

Table 3: Examples of GO annotation for several disease modules. Each module consists of several proteins with many annotations from each of the three categories (CC, BP and MF). The data structure holding this information is 45,800 observations on five variables. Note throughout, Digestive system diseases and modules are labeled by their C06 designation while non-C06 diseases are named

Database ID	category	ID	term	gene id	DiseaseModule
5246	CC	GO:0005654	nucleoplasm	2656	Diabetes_4
35154	BP	GO:0042795	snRNA transcription from RNA polymerase II promoter	61	Diabetes_11
18138	BP	GO:0000187	activation of MAPK activity	596	Diabetes_5
24379	BP	GO:0007062	sister chromatid cohesion	1841	Diabetes_6
27855	CC	GO:0070062	extracellular exosome	MSN18	Diabetes_8
20815	BP	GO:0019932	second-messenger-mediated signaling	3273	Diabetes_5
142	CC	GO:0005720	nuclear heterochromatin	TCP14	C06.301_1
14312	CC	GO:0005794	Golgi apparatus	TCP15	C06.301_1
14410	CC	GO:0005813	centrosome	TCP16	C06.301_1
145	CC	GO:0005829	cytosol	TCP17	C06.301_1
14615	CC	GO:0005832	chaperonin-containing T-complex	TCP18	C06.301_1
147102	CC	GO:0005874	microtubule	TCP19	C06.301_1
66313	CC	GO:0043231	intracellular membrane-bounded organelle	SYT48	C06.301_4
161010	BP	GO:0070207	protein homotrimerization	500	C06.301_1
117511	BP	GO:0035584	calcium-mediated signaling using intracellular calcium source	BCAP318	C06.301_5
11831	BP	GO:0006919	activation of cysteine-type endopeptidase activity	73	C06.301_1
128211	BP	GO:0090263	positive regulation of canonical Wnt signaling pathway	SOX438	C06.301_6
430611	CC	GO:0009986	cell surface	THBS16	C06.301_2

Table 4: KEGG pathway annotation for disease module X. The larger number in the GeneRatio variable refers to genes present in the disease module, the smaller number is the number genes present in that particular pathway. The BgRatio is the ratio of the number of genes in network neighborhood.

ID	Description	GeneRatio	BgRatio	pvalue
hsa04912	GnRH signaling pathway	6/147	92/7301	0.01
hsa04062	Chemokine signaling pathway	9/147	185/7301	0.012
hsa04930	Type II diabetes mellitus	4/147	46/7301	0.013
hsa04915	Estrogen signaling pathway	6/147	98/7301	0.014
hsa05160	Hepatitis C	7/147	131/7301	0.016
hsa04141	Protein processing in endoplasmic reticulum	8/147	165/7301	0.018

Table 5: conventional drug treatments.

Disease	Genes	Drug treatment
Mammary neoplasms	83	goserelin, ziprasidon, pamidronate, chlorambucil, raltitrexed, raloxifene
Prostatic Neoplasms	73	goserelin, zoledronate, epirubicin, flutamide, cisplatin, hydrocortisone
Lung Neoplasms	42	erlotinib, afatinib, getfitinib, bevacizumab, crizotinib, cerintinib
Obesity	37	orlistat, lorcaserin, sibutramine, ribonabant, metformin
Rheumatoid Arthritis	34	methotrexate, leflunomide, hydroxychloroquine, sulfasalazine
Autistic disorder	34	risperidone, aripiprazole, clozapine, haloperidol, sertraline
Hypertensive disease	31	doxazosin, atenolol, ramipril, irbesartan
Diabetes Mellitus, Non-Insulin-Dependent	24	metformin, glibenclamide , gliclazide, repaglinide, sitagliptin
Non-small cell lung carcinoma	24	bevacizumab, ramucirumab, erlotinib, afatinib, gefitinib
Schizophrenia	20	chlorpromazine, fluphenazine, haloperidol, perphenazine, thiothixene
Parkinson disease	19	levodopa, carbidopa, safinsmide, ropinirole, pramipexol
Asthma	18	albuterol, metaproterenol, levalbuterol, pirbuterol, theophylline
Alzheimer's disease	16	donepezil, rivastigime, galantaime, memantine

the remaining 93% had topologically separated modules [30].

The C06 disorders in the 1st shell contains links to 1,329 non C06 diseases by means of 566 shared or common genes. We examine the most similar non-C06 disease modules with the greatest overlaps of genes and GO annotations and determine if they are in fact part of the same disease causing mechanisms. It is important to note that several drugs do not actually target the disease implicated proteins but instead affect their neighboring proteins in the disease module [7].

In table 6 those drugs marked with a tick in the reposition column have actually undergone clinical trials for repurposing. The ClinicalTrials.Gov identifier is presented where at least one such study has taken place. The diabetes drug Metformin has undergone trials for Cancer therapy. The multiple myeloma drug Bortezomib, is is now repurposed for Advanced Non-Small Cell Lung Cancer. Pazopanib is a multi-targeted receptor tyrosine kinase inhibitor that blocks tumour growth and is now aimed at Alzheimer's disease. Hydralazine-valproate was originally an anti-hypertensive drug but is now used to treat cervical and breast Cancers.

Table 6: Overlapping, non-redundant disease modules for C06 and non-C06 disorders that are linked by shared genes are now combined, shown here 10 out of 21. Any drugs that have been identified by our system as potential reposition candidates based on protein-to-drug pathways and chemical similarity are shown, those that have actually been repositioned are indicated by a tick mark. The number refers to the unique ID code for ClinicalTrials.Gov. The biological score is the combination of KEGG pathway and Gene Ontology ranks for each individual disease module and summed for the overall module.

ID	Diseases	Hub Genes	Reposition	Score
1	Small cell carcinoma-1, Diabetes-7,Hypertension-7	TP5342, YBX13, IKBKB12 MAPK37, FAS8, BAX6, PARP127	Metformin ✓ NCT01864096	15.3
2	Gastrointestinal Stromal Tumor-5, Small cell carcinoma-1, Diabetes-5	EZR33, EZR36, XPO110, LIMA12, HSPA59, LIMA11, HSPA1A7	Bortezomib✓ NCT00714246	14.8
3	Small cell carcinoma-1, Cholelithiasis-7, Diabetes-7	TP53 APP1 CTNNB12 CTNNB111 MAPK33	Lenalidomid	15.5
4	Small cell carcinoma-1, Diabetes-7, Adenomatous Polyposis Coli-2,	MAPK112, MTOR8, PSEN11 APP30, AR6, CTNNB13, ILK12	Pazopanib✓ NCT00367679	15.3
5	Small cell carcinoma-1, Adenomatous Polyposis Coli-3, Diabetes-7	CEBPB1, TP5342, PARP14, PARP15, MAPK130	N/A	17.5
6	Small cell carcinoma-1, Diabetes-10, Hypertension-10	CCAR22, HEY219, FBXO72, DVL120 FGF1039, SNW115, FOXO42, SUPT7L2, KSR13	N/A	14.8
7	Parkinson-2, Barrett Esophagus-17, Diabetes-5	APP, HSP90AA1, YWHA, GAPDH, HSP90AA1.	Ropinirole	8.41
8	Small cell carcinoma-1, Diabetes-12, Schizophrenia-2	EP3003, IKBKB7, PDPK129 PDPK18, IKBKG17	Glitazone	18.8
9	Diabetes-9, Small cell carcinoma-1, Hypertension-9	VHL2, MAPK81, MTOR8, EEF1G2, MTOR63, PML1, ACACA1, AKT195, EZH221, CTNNB1111,	N/A	11.5
10	Small cell carcinoma-1, Diabetes-5, Hypertension-5	MAPK18, CDH13, PLCG15, PLEC5, PLCG22	Hydralazine-valproate✓ NCT00532818	13.2

6.1 Comparison with competing systems

Where practical we compared and contrasted the results of our system with several of the competing community detection systems. In table 7 we explore (based on the literature) the key criteria. The first criteria is to see if the system can detect overlapping communities, this is important as it gives insights into shared proteins and their relationships in potentially several diseases. The next criteria is to determine if the systems can predict disease implicated proteins (usually as part of a test/validation set), if a high accuracy can be obtained, potentially novel and interesting medical insights could be gained. The next criteria determines if some sort indication of the biological plausibility of the discovered communities is possible. The final criteria checks if the system has the capacity to suggest alternative therapies for known drugs within overlapping communities of disease.

Table 7: Community detection system comparison

System	Overlap Comm	Disease prediction	Biological plausibility	Drug reposition
McGarry	Y	N	Y	Y
DIAMOnD	N	Y	N	N
Yu	N	N	Y	Y
MBiRW	N	Y	N	Y
CommWalker	N	Y	Y	N

Referring to table 7 we compare our system with the others. The DIAMOnD system developed by Ghiassian can detect disease modules and predict disease associated proteins but not overlapping modules [12]. There is little attempt by DIAMOnD to assess biological plausibility of the discovered communities which would provide a degree of validation. Nor is there any attempt to use the knowledge gained from the discovered communities to suggest alternative or tentative therapies by drug repositioning. The system developed by Yu actively seeks to form disease modules with the goal of repurposing drugs [48]. The modules formed are not overlapping and the data was only for Cancer related diseases. There was no attempt to link biological plausibility into Yu's system, however by cross-checking with clinicaltrails.gov they were able to confirm that many of the drugs their system identified were undergoing clinical trials. The MBiRW algorithm can predict candidate diseases for many different drugs and is able to predict novel disease associations for drugs without any known associated diseases information [25]. However, the authors appreciate there are limitations and that improvements to the algorithm could be made if they used prior biological knowledge. The CommWalker algorithm combines multiple semantic views of the protein interaction networks to deduce community organisation [24]. It is able to take advantage of prior biological knowledge using gene ontology and is the most similar to our work with the exception that ours is able to detect overlapping communities. Nor does CommWalker make predictions for candidate drugs for repurposing.

6.2 Medical evaluation of disease proteins and drugs

We have validated where we can the communities of disease modules generated by our system. We are reasonably certain that they actually represent biologically realistic and plausible entities when we use KEGG pathways and GO annotations. For example, the thiazolidinediones (glitazones) are agents (e.g. pioglitazone) which are currently in clinical use for the management of type 2 diabetes and were identified by our system in connection to repurposing for use in treatment of small cell carcinoma, based on a series of hub genes; EP3003, IKBKB7, PDPK129, PDPK18 and IKBKG17.

The thiazolidinediones all have a slow onset of action requiring 2 months to achieve a reduction in blood glucose levels by reducing hepatic glucose output with a concomitant increase in glucose uptake by muscles. Their mechanism of action is complex and mediated via a class of nuclear receptors known as peroxisome proliferator-activated receptors1 (PPARλ is targeted by the thiazolidinediones). PPARλ is complexed with the retinoid X receptor (RXR) and the thiazolidinediones act as agonists that bind to the PPARλ-RXR complex (transcription factor) which in turn initiates a change in the transcription of a number of genes; some of the genes controlling lipid and glucose metabolism are under the control of the PPARλ-RXR transcription factor as are genes controlling cell proliferation and differentiation suggesting a possible role for the thiazolidinediones in the treatment of certain cancers. The possibility of repurposing existing drugs is an attractive one due to the significant cost saving versus new drug development and in the case of cancer, and in particular small cell carcinoma, important since it is known that tumours can become refractory to drug treatment.

Furthermore, within a number of our modules we see diabetes and hypertension listed together in table 6. The fact that our model has grouped these diseases together is biologically plausible since it is known that they share aspects of pathophysiology, particularly in connection to their use of cyclic guanosine monophosphate (cGMP) as a signalling system and indeed, several drugs on the market used in the management of both conditions target cGMP (they target either cGMP forming or cGMP degrading enzymes) based signalling pathway.

An obvious limitation to this current work is that we have not tested the clinical validity of our hypothesis but with the relevant in vitro/in vivo work it may be possible to confirm that the agents identified could be suitable for repurposing. Ultimately, approaches like the one outlined here if coupled with the appropriate biological testing, may stimulate a reappraisal of agents based on the grouping of diseases within modules. The concept of the disease network needs to be accepted by the medical practitioners if we are to understand and combat diseases more effectively, rather than partition diseases into neat categories with the notion that they are free standing and independent.

7 Conclusions

In this current work we have developed a new approach using link clustering that has indicated overlap between disease modules and hub genes, which in turn may suggest new molecular mechanisms that underpin apparently unrelated diseases. The usefulness of our approach is to identify diseases that are hereditary or have a genetic component to them. A defective protein may eventually fail to provide its function in the cell or may operate with reduced efficiency, the possible effects on its interaction partners (proteins or drugs) will determine if other health problems may occur. Many proteins have been identified with some sort of condition or disease, the key aspect is to predict the effects on other proteins. The idea of characterizing disease at the molecular level rather than at an organ and or symptom-based phenotype is bound to lead to a greatly improved understanding of complex pathophysiology's and also suggest new diagnostic tests and mechanism-based therapeutic interventions (this last point is particularly relevant if precision medicine is to fulfill its full promise of changing the existing medical paradigm). The work discussed in this paper confirms large levels of functional overlap or modularity occur in protein networks giving advantages of multiple functionality. Indeed this is undoubtedly the result of evolutionary processes since it confers the benefits of robustness, redundancy of elements and increasing the repertoire of cellular functionality or abilities with the same genes and components. However, disease modules are a corollary of this phenomena as they span the topological and functional boundaries of gene/protein networks. The method we use, namely link clustering enables proteins to be allocated more realistically across several communities (implying multiple biological functions and roles) than non-link community approaches which would allocate each protein to a single community. For future work we are currently investing the role of protein interactions that lead to drug side-effects as a means of predicting candidate drugs for repositioning. To achieve this we are investigating the modification of a random walk algorithm that can move between the overlapping disease modules.

Compliance with ethical standards

Conflict of interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgements We would like to thank the anonymous reviewers for their helpful comments which have improved the paper.

References

1. Ahn Y, Bagrow J, Lehmann S (2010) Link communities reveal multiscale complexity in networks. Nature 466:761–764
2. Albert R, Barabasi A (2002) Statistical mechanics of complex networks. Rev Mod Physics 74(1):450–461
3. Ashburn T, Thorl KB (2004) Drug repositioning: identifying and developing new uses for existing drugs. Nature Reviews Drug Discovery 3:673–683
4. Atias N, Sharan R (2011) An algorithmic framework for predicting side effects of drugs. Journal of Computational Biology 18(3):207–218
5. Barabasi A (2016) Network Science, 1st edn. Cambridge University Press
6. Barabasi A, Oltvai Z (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101–113
7. Barabasi A, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nat Rev Genet 12:56–68, DOI 10.1038/nrg2918
8. Barrenas F, Chavali S, Holme P, Mobini R, Benson M (2009) Network properties of complex human disease genes identified through genome-wide association studies. PLoS ONE 4(11):e8090, DOI 10.1371/journal.pone.0008090
9. Bauer-Mehren A, Bundschus M, Rautschka M, Mayer M, Sanz F, Furlong L (2011) Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. PLoS ONE 6(6):e20,284, DOI 10.1371/journal.pone.0020284
10. Dissez G, Ceddia G, Pinoli P, Ceri S, Masseroli M (2019) Drug repositioning predictions by non-negative matrix tri-factorization of integrated association data. In: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Association for Computing Machinery, New York, NY, USA, BCB '19, p 25–33, DOI 10.1145/3307339.3342154
11. Gamalielsson J, Nilsson P, Olsson B (2006) A GO-Based method for assessing the biological plausibility of regulatory hypotheses. In: ICCS 2006. Lecture Notes in Computer Science, Springer, Reading, UK, vol 3992, pp 879–886
12. Ghiassian S, Menche J, Barabasi A (2015) A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. PLoS Computational Biology 11(4), DOI 10.1371/journal.pcbi.1004120
13. Goh K, Choi I (2012) Exploring the human diseasome: the human disease network. Brief Funct Genomics 11(6):533–42, DOI 10.1093/bfgp/els032. Epub 2012 Oct 12
14. Goh K, Cusick M, Valle D, Childs B, Vidal M, Barabasi A (2007) The human disease network. Proceedings of the National Academy of Sciences 104(21):8685–8690
15. He D, Liu Z, Chen L (2011) Identification of dysfunctional modules and disease genes in congenital heart disease by a network-based approach. BMC Genomics 12, DOI 10.1186/1471-2164-12-592.
16. Hoff P (2009) Multiplicative latent factor models for description and prediction of social networks. Computational and Mathematical Organization Theory 15(4):207–218
17. Hric D, Darst R, Fortunato S (2014) Community detection in networks: Structural communities versus ground truth. Phys Rev E 90:062,805
18. Kalinka A, Tomancak P (2011) Linkcomm: an R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type. Bioinformatics 27(14):2011 – 2012
19. Klamt S, Saez-Rodriguez J, Lindquist J, Simoeni L, Gilles E (2006) A methodology for the structural and functional analysis of signalling and regulatory networks. BMC Bioinformatics 7(56):
20. Kolaczyk E, Csardi G (2014) Statistical Analysis of Network Data with R. Springer
21. Lechner M, Hohn V, Brauner B, Dunger I, Fobo G, Frishman G (2012) Cider: multifactorial interaction networks in human diseases. Genome Biology 13(7):R62, DOI 10.1186/gb-2012-13-7-r62, URL http://genomebiology.com/2012/13/7/R62
22. Lee A, Ming-Chih L, Hsu C (2011) Mining dense overlapping subgraphs in weighted protein–protein interaction networks. BioSystems 103:392 – 399
23. Li W, Xie J, Mo J (2018) An overlapping network community partition algorithm based on semi-supervised matrix factorization and random walk. Expert Systems with Applications 91:277–285

24. Luecken N, Page M, Crosby A, Mason S, Reinert G (2017) CommWalker: correctly evaluating modules in molecular networks in light of annotation bias. Bioinformatics 1-7, DOI 10.1093/bioinformatics/btx706
25. Luo H, Wang J, Li M, Luo J, Peng X, Wu F, Pan Y (2016) Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. Bioinformatics 32(17):2664–2671, DOI DOI:https://doi.org/10.1093/bioinformatics/btw228
26. Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, Kuang W, Peng J, Chen L, Zeng J (2017) A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. bioRxiv DOI 10.1101/100305
27. McGarry K (2013) Discovery of functional protein groups by clustering community links and integration of ontological knowledge. Expert Systems with Applications 40(13):5101–5112
28. McGarry K, Chambers J, Oatley G (2007) Graph based analysis of protein interaction for diabetes research. Artificial Intelligence in Medicine 41(2):129–144
29. McGarry K, Graham Y, McDonald S, Rashid A (2018) RESKO: Repositioning drugs by using side effects and knowledge from ontologies. Knowledge Based Systems 160:34–48, DOI 10.1016/j.knosys.2018.06.017
30. Menche J, Sharma A, Kitsak M, Ghiassian S, Vidal M, Loscalzo J, Barabasi A (2015) Uncovering disease-disease relationships through the incomplete human interactome. Science 347(6224), DOI 10.1126/science.1257601
31. Meng Y, Liub Q, Chen D, Meng Y (2017) Pathway cross-talk network analysis identifies critical pathways in neonatal sepsis. Computational Biology and Chemistry 68(3):101–106
32. Michael K, Szklarczyk D, Franceschini A, von Mering C, Jensen L, Juhl L, Bork P (2012) Stitch 3: zooming in on protein–chemical interactions. Nucleic Acids Research 40(D1):D876–D880, DOI 10.1093/nar/gkr1011
33. Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M (2005) Whole-proteome prediction of protein function via graph theoretic analysis of interaction maps. Bioinformatics 21(1):302–310
34. Newman M, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026,113
35. Palla G, Derényi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043):814–818
36. Peel L, Larremore D, Clauset A (2017) The ground truth about metadata and community detection in networks. Science Advances 3(5), DOI 10.1126/sciadv.1602548
37. Peyvandipour A, Saberian N, Shafi A, Donato M, Draghici S (2018) A novel computational approach for drug repurposing using systems biology. Bioinformatics pp 1–9, DOI 10.1093/bioinformatics/bty133
38. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL (2002) Hierarchical organization of modularity in metabolic networks. Science 297(5586):1551–1555
39. Schaub M, Delvine J, Rosvall M, Lambiotte R (2017) The many facets of community detection in complex networks. Applied Network Science 1(2), DOI 10.1007/s41109-017-0023-6
40. Sun K, Buchan N, Larminine C, Przulj N (2014) The integrated disease network. Integrative Biology 6(6):1069–1079, DOI 10.1039/C4IB00122B
41. Vidal M, Cusick M, Barabasi A (2011) Interactome networks and human disease. Cell 144(6):986–998, DOI doi:10.1016/j.cell.2011.02.016
42. Walhout A (2009) Getting an edge on human disease. Molecular Systems Biology 322:1–2
43. Wanders R (2004) Metabolic and molecular basis of peroxisomal disorders: a review. American Journal of Medical Genetics 126A:355–75
44. Wang J, Du Z, Payattakool R, Yu P, Chen C (2007) A new method to measure the semantic similarity of GO terms. Bioinformatics 23(10):1274–1281
45. Weininger D (1988) Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. Journal of Chemical Information and Modeling 28(1):31–6
46. Yu G, He Q (2016) ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol BioSyst 12:477–479, DOI 10.1039/C5MB00663E
47. Yu G, Wang L, Han Y, He Q (2012) clusterprofiler: an r package for comparing biological themes among gene clusters. Journal of Integrative Biology 16(5):284–287, DOI 10.1089/omi.2011.0118
48. Yu L, Wang B, Gao L (2016) The extraction of drug-disease correlations based on module-distance in incomplete human interactome. BMC Systems biology 10(111), DOI 10.1186/s12918-016-0364
49. Zitnik M, Zupan B (2015) Data fusion by matrix factorization. IEEE Transactions on pattern analysis and machine intelligence 37(1):41–53