Molecular and Cellular Probes 30 (2016) 178–181

Contents lists available at ScienceDirect

Molecular and Cellular Probes

journal homepage: www.elsevier.com/locate/ymcpr

Short communication

Rare disease relations through common genes and protein interactions

Sara Fernandez-Novo [a], Florencio Pazos [b], Monica Chagoyen [b, *]

[a] Escuela Politecnica Superior, Universidad Autonoma de Madrid, Madrid 28049, Spain
[b] Computational Systems Biology Group, National Centre for Biotechnology (CNB-CSIC), Madrid 28049, Spain

ARTICLE INFO

Article history:
Received 18 February 2016
Received in revised form
15 March 2016
Accepted 15 March 2016
Available online 16 March 2016

Keywords:
Rare diseases
Biomedical databases
Disease relations
Protein interactions

ABSTRACT

ODCs (Orphan Disease Connections), available at http://csbg.cnb.csic.es/odcs, is a novel resource to explore potential molecular relations between rare diseases. These molecular relations have been established through the integration of disease susceptibility genes and human protein-protein interactions. The database currently contains 54,941 relations between 3032 diseases.

Rare diseases are those diseases that affect a relatively limited number of individuals (no more than 5 out of 10,000 in the European Union). Due to their low prevalence, it is important to gather all the information available on these diseases and to look for relations between them. This could open new opportunities for transferring knowledge from one disease to the other (markers, targets, drugs, etc.) and for exploiting synergies between independent research lines.

Previous studies have used different approaches to establish molecular relations between diseases: common genes [1,2], microRNAs [3], functional linkages [4], protein localization [5], protein-protein interactions [6] or consecutive metabolic reactions [7]. In spite of the value of these molecular relations for a wide range of biomedical researchers and clinicians, only a few resources offer friendly graphical interfaces for users to search, retrieve or inspect a given disease-disease relation in detail. This is the case of MalaCards [8], a disease database that allows users to navigate from a disease to related diseases based on a variety of annotations (like genes, pathways, phenotypes, compounds and Gene Ontology terms). And DiseaseConnect [9], a resource focused on disease relations built by shared susceptibility genes and differential expression data.

Focusing on rare diseases, Zhang et al. constructed and studied global properties of several disease networks [2]. In the most populated network, two rare diseases were connected if they shared at least one susceptibility gene thus analysing at that time 2259 relations among 1170 diseases. But by using this shared-gene formalism they recognised that many rare disease relations could not be discovered. To overcome this limitation, they additionally constructed networks on the basis of enriched annotations: biological processes, cellular components, phenotypes and pathways. These networks were constructed for the subset of diseases with at least four known susceptibility genes. Thus they contained only a small fraction of the diseases in the gene-based network (up to 15%), as most rare diseases are monogenic.

In order to find additional potential molecular relations for the largest number of rare diseases, an alternative strategy is to link two rare diseases not only because they share genes, but because their associated gene products are interacting as well. This strategy was applied to study the topology and function of a global human disease network [6]. It is also sustained by previous studies that have reported a higher similarity of both symptoms [10] and comorbidities [11] among those diseases associated to interacting proteins than among those associated to proteins that do not interact.

In this work we present Orphan Disease Connections (ODCs), a resource of potential relations between rare diseases established by

* Corresponding author. Centro Nacional de Biotecnologia (CNB-CSIC), Darwin 3, 28049 Madrid, Spain.
E-mail address: monica.chagoyen@cnb.csic.es (M. Chagoyen).

http://dx.doi.org/10.1016/j.mcp.2016.03.004
0890-8508/

shared susceptibility genes and protein interactions of the corresponding gene products. To the best of our knowledge there are no searchable public resources providing potential rare disease connections on the basis of protein-protein interactions. ODCs has a web interface, available at http://csbg.cnb.csic.es/odcs, to search for the diseases connected to one of interest, to explore in the detail the connections between two rare diseases, or to search for rare diseases associated with a given gene. This interface was designed with simplicity in mind, in a Google-like style so that filling a single text-box is enough to start navigating. Search results are presented both in text and in a graphical interactive way (network visualization), allow users to explore in detail the molecular connections and include links to external resources.

The first version of OCDs is built upon two main sources of data: rare disease susceptibility genes from Orphadata version 1.0.20. (Orphadata: Free access data from Orphanet© INSERM 1997. Available on http://www.orphadata.org), and human protein-protein interactions from HIPPIE [12]. HIPPIE is a resource of human protein interactions compiled from a large number of databases. In HIPPIE, each interaction is associated to a numeric score that reflects its reliability. Therefore, a HIPPIE score is a good estimator of the amount and quality of the experimental evidence for a given interaction.

Additionally OCDs contains link-outs to conveniently navigate to other important biomedical resources like the International Classification of Diseases (ICD-10), OMIM human genes and reported clinical cases [13], UniProt protein sequences and functions [14] and MeSH controlled vocabulary (www.nlm.nih.gov/mesh).

Two rare diseases are connected in OCDs if they share a susceptibility gene or if there is at least an interaction between the proteins encoded by their susceptibility genes. With this approach, a much larger number of connections than those based only in shared genes could be established, as most rare diseases (73%) are associated with only one gene. The current version contains 54,941 relations among 3032 diseases, of which 5263 relations correspond to shared-genes.

To assess the value of our approach, we have calculated the phenotypic similarity of all pairs of rare diseases for the subset of diseases where phenotypic information as well as susceptibility genes are available. The phenotypic similarity between two diseases was calculated as follows. First, phenotypes associated to rare diseases were compiled from the Human Phenotype Ontology (HPO) [15] and direct disease-phenotype associations were expanded to include the parent terms of a given phenotype in the HPO hierarchy. Second, we calculated the probability of each phenotype term in the ontology $p(c)$, as the number of diseases associated with it, divided by the total number of diseases. Finally the phenotypic similarity between two diseases was defined based on the probability of the most specific common phenotype [16]:

$$sim(d1, d2) = \max_{c \in S(c1, c2)} [-\log p(c)]$$

where $c1$ and $c2$ are all the phenotype terms associated to diseases $d1$ and $d2$ respectively.

Then we classified disease pairs in five groups, those that share: (i) genes, (ii) protein interactions, (iii) molecular pathways, (iv) protein complexes, and (v) the remaining pairs (not sharing any of the previous). In order to build this classification human protein interactions were compiled from HIPPIE [12], pathways from Reactome [17] and protein complexes from CORUM [18].

Fig. 1 shows the distribution of phenotypic similarity of each of the five groups. As expected, diseases sharing genes have more similar phenotypes (mean value 1.38), followed by those sharing interactions (0.93), pathways (0.84), complexes (0.64) and the remaining pairs (0.39).

According to these results, apart from common genes, the approach that yields the highest proportion of significant disease relations for the largest set of diseases is the one based on protein interactions. This justifies the development of resources providing potential rare disease associations through protein interactions such as this one.

Three types of searches can be performed in ODCs: retrieving all the relations of a rare disease (Rare disease search), exploring in detail the potential relation between two rare diseases (Diseases connection search), or retrieving the rare diseases associated with a given gene (Gene search). To facilitate searches, the input text-boxes include pull down menus that show a list of matching diseases or genes as the user types 3 characters or more. Each of the three searches provides an alternative view on the relations: disease-centric, disease-disease connection centric and gene-centric. They all offer a graphical visualization of the relations found (disease-gene, disease-disease and gene-gene) in the form of networks by means of Cytoscape Web [19]. Views provide as well detailed textual information and links to external resources for navigation. Finally, users can download the relations found with an export utility.

A rare disease search takes to a disease-centric view. In this view, the query disease is shown together with all the related diseases, both graphically in an interactive network, as well as in a list. In this list ODCs first ranks relations based on shared genes (by number), followed by those based on protein interactions (by the sum of HIPPIE scores). The type of relation (by common genes or by interacting proteins) is also differently represented in the graph. Additional disease information is also provided: disease classification (according to Orphanet and ICD-10), epidemiological data, signs and symptoms, and associated genes (with links to UniProt and OMIM databases). From this page, the user can explore the details of a particular disease-disease relation.

When a user searches for a potential relation between two diseases ODCs shows a graph with all the genes associated to the two diseases, as well as their protein product interactions (if any). Shared genes are marked (both in the network and in the text) and HIPPIE interactions scores are reported. This view also shows, in textual format, other information for both diseases: their nomenclature, classification and epidemiological data, as well as their common and distinct signs and symptoms. ODCs offers two additional utilities to help users in judging the relevance of a disease-disease relation. The first consists of link-outs to the HIPPIE web site that show the experimental evidences supporting each protein-protein interaction. The second is a PubMed search builder that allows users to verify whether different associations among diseases and interacting gene products have been reported in the literature (by combining disease and gene terms through the provided checkboxes).

As an example, a detailed graphical view on the connections among Noonan and Distal 22q11.2 microdeletion syndromes is shown in Fig. 2. The relation between these two syndromes is supported by the fact that the ERK MAP kinase signalling pathway is disrupted in both developmental syndromes [20]. This relation, established in ODCs through protein interactions, could not be found based solely on shared genes.

Finally, a gene search takes to a gene-centric view. In this case, all the diseases associated to this gene are shown, together with the genes whose protein products are known to interact with its product. From here, the user can navigate to a given disease ('rare disease view') or to another gene view.

Establishing and studying molecular relations between diseases is a very active area of research in systems medicine. Looking for potential relations between diseases is especially important in the