Journal of Proteomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Proteomics

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

Integrating multi-source information on a single network to detect disease-related clusters of molecular mechanisms

Margarita Zachariou, George Minadakis, Anastasis Oulas, Sotiroula Afxenti, George M. Spyrou*

The Cyprus Institute of Neurology and Genetics, 6 International Airport Avenue, P.O.Box 23462, 2370 Nicosia, Cyprus

ARTICLE INFO

Keywords:
Network science
Alzheimer's disease
Multi-source integration
Predictive
Preventive and personalised medicine

ABSTRACT

The abundance of available information for each disease from multiple sources (e.g. as genetic, regulatory, metabolic, and protein–protein interaction) constitutes both an advantage and a challenge in identifying disease-specific underlying mechanisms. Integration of multi-source data is a rising topic and a great challenge in precision medicine and is crucial in enhancing disease understanding, identifying meaningful clusters of molecular mechanisms and increasing precision and personalisation towards the goal of Predictive, Preventive and Personalised Medicine (PPPM). The overall aim of this work was to develop a novel network-based integration methodology with the following characteristics: (i) maximise the number of data sources, (ii) utilise holistic approaches to integrate these sources (iii) be simple, flexible and extendable, (iv) be conclusive. Here, we present the case of Alzheimer's disease as a paradigm for illustrating our novel approach.

Significance: In this work we present an integration methodology, which aggregates a large number of the available data sources and types by exploiting the holistic nature of network approaches. It is simple, flexible and extendable generating solid conclusions regarding the molecular mechanisms that underlie the input data. We have illustrated the strength of our proposed methodology using Alzheimer's disease as a paradigm. This method is expected to serve as a stepping-stone for further development of integration methods of multi-source omic-data and to contribute to progress towards the goal of Predictive, Preventive and Personalised Medicine (PPPM). The output of this methodology may act as a reference map of implicated pathways in the disease under investigation, where pathways related to additional omics data from any kind of experiment may be projected. This will increase the precision in the understanding of the disease and may contribute to personalised approaches for patients with different disease-related pathway profile, leading to a more precise, personalised and ideally preventive management of the disease.

1. Introduction

Most major classes of human diseases are rarely caused by a single gene and are often the result of the interaction among multiple genes and other factors such as epigenetics and the environment. Interestingly, most human diseases are not independent and share common genetic profiles. Moreover, the abundance of available information for each disease from multiple sources (e.g. as genetic, regulatory, metabolic, and protein–protein interaction) constitutes both an advantage and a challenge in identifying disease-specific underlying mechanisms. Hence, a systematic approach which considers a disease in the framework of a complex network emerging from the integration of information from various levels can be crucial in understanding both the pathophysiology of single diseases and the common biological basis among different diseases [1]. This systematic approach falls within the realms of *systems bioinformatics*, a newly emerging discipline, which lies in the intersection of systems biology and classical bioinformatics (for a comprehensive review see [2]). Systems bioinformatics harnesses the powerful methods of network science to both integrate information across different levels/sources in the form of *networks* as well as to extract information from the features of complex biological networks which were previously unavailable with typical bioinformatics analysis of “omic” data. Network science provides a powerful mathematical framework for the investigation of large and complex systems by representing them in the form of networks [3,4]. A network (or graph) is a collection of nodes (or vertices) connected by edges (or links) and is used to model the pairwise relations among certain objects. In biological networks nodes can for example represent genes, proteins and

Abbreviations: AD, Alzheimer's disease; CNVs, Copy number variations; DEGs, Differentially expressed genes; PPPM, Predictive preventive and personalised medicine; PPIs, Protein–protein interactions; SNPs, Single nucleotide polymorphisms

* Corresponding author.
E-mail address: georges@cing.ac.cy (G.M. Spyrou).

https://doi.org/10.1016/j.jprot.2018.03.009
Received 22 November 2017; Received in revised form 27 February 2018; Accepted 5 March 2018

drugs and edges can represent physical, chemical or functional pairwise relationships [5]. Network approaches in diseases are key in representing the knowledge per layer and have shown great promise in various directions such as in disease related pathway discovery network-based biomarker discovery [6] and network-based drug-discovery [7]. Network methods can be applied beyond the representation of information of data per level to formulate integrated representations of a set of networks to a single network. For more information on both network approaches as well as other methods for biological data integration the reader is referred to relevant reviews [8,9].

For many chronic and severe diseases such as cancer, diabetes 2 and neurodegenerative diseases treatment is initiated after the onset of the disease and often at a non-reversible stage. Thus, the lifelong management of such diseases is expected to have an enormous economic burden to healthcare systems. This realisation has led to a shift of paradigm from medicine as a delayed intervention, overdosed and untargeted medication to Predictive, Preventive and Personalised Medicine (PPPM) [10]. The European Association for Predictive, Preventive and Personalised Medicine (EPMA) has recently released expert recommendations and guidelines on the importance of shifting medicine and healthcare towards PPPM both at European and global scale [10]. PPPM is emerging as the integration of multi-disciplinary efforts in healthcare aimed at reducing the prevalence of diseases through (1) predicting onset of disease, (2) providing preventive measures and focusing on both healthy individuals and patients (3) developing personalised treatments, and, (4) providing cost-effective solutions for healthcare systems. This can be achieved through the integration of approaches such as bioinformatics, disease modeling, pharmacogenetics, individual patient profiling, etc. (for recent studies see [11–14]). In the heart of the PPPM efforts lies the need for deeper understanding of the molecular mechanisms implicated in a disease in order to (i) target the disease treatment more precisely (ii) facilitate biomarker discovery, and, (iii) identify the healthy individuals for which preventive measures are required. Towards this direction, we propose a method for integrating multi-source data to uncover the key molecular mechanisms involved in a disease. These insights can lead to a more effective, early and predictive diagnosis in patients as well as targeted prevention in healthy individuals at-risk.

Integration of multi-source data is a rising topic and a great challenge in PPPM and is crucial in enhancing disease understanding and identifying meaningful clusters of molecular mechanisms, thus, increasing treatment precision and personalisation. The overall aim of this work was to develop a novel network-based integration methodology which can (i) maximise the number of data sources, (ii) utilise holistic approaches to integrate these sources (iii) be simple, adjustable and expandable, (iv) be conclusive. This systems level approach has the potential to provide holistic insights to research questions related to disease-risk estimation, prediction of disease onset and progression, effective treatment, identification of putative drug targets and computational drug discovery and repurposing. The case of Alzheimer's disease (AD) was chosen as a paradigm for illustrating our novel approach. AD is a chronic neurodegenerative disease of complex, and as yet, uncertain aetiology despite the vast availability and the multisource nature of omic data [15].

Our work was structured in three main steps (further detailed in Methods and Results) as shown in Fig. 1. The first step was the retrieval of AD-specific data from multiple sources and the representation of each data set in network form. The second step was the construction of a super network based on the different sources of information and the aggregation into a gene-specific score based both on gene characteristic information and on gene-gene integrated inter-relation (i.e. the topology of the resultant super network). The third step was the enrichment analysis of the selected top-ranked nodes/genes followed by pathway mapping on a novel reference pathway network map (created by parsing the KEGG database for functional links among pathways). Finally, by setting a minimum connectivity requirement among the initial pathways, we derived an enriched set of networked pathways that was clustered to a few comprehensive functional pathway groups for further investigation.

2. Methods

2.1. Data resources

Data was downloaded in August 2017 from the human disease database Malacards (http://www.malacards.org/) [16] using the “Alzheimer's Disease” parent term and (MalaCards ID ALZ034). This provided multiple sources of information, of which we extracted specific genetic, molecular and therapeutic information related to AD. The sources of information were: (i) Genes involved in AD, (ii) Pathways related to AD (MalaCard uses PathCards as a source), (iii) Variants – Single Nucleotide Polymorphisms (SNPs) and Copy number variations (CNVs) related to AD, (iv) Drugs related to AD, (v) Differentially expressed genes (DEGs) in AD (blood and brain).

In order to extract the known gene targets from the drugs available via MalaCards, we downloaded a local copy of DrugBank (https://www.drugbank.ca/) [17] and parsed it for all the known gene targets of the Alzheimer's-related drugs.

In addition, we obtained non-coding RNA, specifically microRNA (miRNA) regulation information, using three different miRNA-gene target databases, namely (mirtarbase -6.1 [18], microcosm-2012-12-05 [19], targetscan-2012-12-05 [20]). We retained only the miRNA gene targets that had previously been associated with Alzheimer's according to mir2Disease (http://www.mir2disease.org/) database [21]. This was done by downloading all entries in mir2Disease and parsing the data for miRNAs implicated in AD.

Finally, we extracted information for protein-protein interactions and genetic interactions using the Alzheimer's-related genes obtained from MalaCards. This was achieved using the GeneMANIA (http://genemania.org/) [22] standalone tool that in turn makes use of databases BioGRID (https://thebiogrid.org/) [23] and PathwayCommons (http://www.pathwaycommons.org/) [24] to retrieve the above-mentioned information.

2.2. Network construction

In total we obtained 6 different sources of information and constructed the corresponding networks. Specifically, utilizing theses different types of sources we generated 6 types of networks related to AD, namely a Gene-Drug network, a Gene-Pathway network, a Gene-Variant (SNP and CNVs) network, a Gene-microRNA network, a Gene-Gene network (based on protein-protein interactions) and a Gene-Gene network (based on genetic interactions) edge lists.

Networks were generated for all of these cases using Cytoscape (http://www.cytoscape.org/) [25]. For an illustration of an aggregated version of all the above networks see Fig. 2 and for the individual networks see Supplementary Figs. S1–S6.

In order to obtain gene-specific network related information, we transformed all the networks into a gene-gene network format. This was achieved by looking at the commonalities shared by all pairs of genes. Namely, three separate new networks were constructed whereby the genes were associated based on their common drugs, pathways and miRNAs. Thus, a total of three new networks with weighted edges were constructed using these methods. These were added to the remaining two gene-gene networks from the protein-protein and genetic interaction data. A final table was generated for all pairwise combinations of genes involved in AD and the edge weight using the above-mentioned five gene-gene networks. In case that no information was available for a given pair of genes the respective weight was given the value of 0.

In addition to the gene-gene networks we also extracted gene-specific information where possible, i.e. for (i) drugs (how many target each gene), (ii) miRNAs (how many target each gene), (iii) pathways