Software

MODalyseR—a novel software for inference of disease module hub regulators identified a putative multiple sclerosis regulator supported by independent eQTL data

Hendrik A. de Weerd 1,2,†, Julia Åkesson 1,2,†, Dimitri Guala 3,4, Mika Gustafsson 2,*,‡ and Zelmina Lubovac-Pilav 1,*,‡

1School of Bioscience, Systems Biology Research Center, University of Skövde, Skövde 541 45, Sweden, 2Department of Physics, Chemistry and Biology, Linköping University, Linköping 581 83, Sweden, 3Department of Biochemistry and Biophysics, Stockholm University, Solna 17121, Sweden and 4Merck AB, Solna 16970, Sweden

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
‡Authors share senior authorship.

Associate Editor: Zhang Zhang

Received on October 14, 2021; revised on January 4, 2022; editorial decision on January 11, 2022; accepted on January 24, 2022

Abstract

Motivation: Network-based disease modules have proven to be a powerful concept for extracting knowledge about disease mechanisms, predicting for example disease risk factors and side effects of treatments. Plenty of tools exist for the purpose of module inference, but less effort has been put on simultaneously utilizing knowledge about regulatory mechanisms for predicting disease module hub regulators.

Results: We developed MODalyseR, a novel software for identifying disease module regulators and reducing modules to the most disease-associated genes. This pipeline integrates and extends previously published software packages MODifieR and ComHub and hereby provides a user-friendly network medicine framework combining the concepts of disease modules and hub regulators for precise disease gene identification from transcriptomics data. To demonstrate the usability of the tool, we designed a case study for multiple sclerosis that revealed IKZF1 as a promising hub regulator, which was supported by independent ChIP-seq data.

Availability and implementation: MODalyseR is available as a Docker image at https://hub.docker.com/r/ddeweerd/modalyser with user guide and installation instructions found at https://gustafsson-lab.gitlab.io/MODalyseR/.

Contact: zelmina.lubovac@his.se or mika.gustafsson@liu.se

Supplementary information: Supplementary data are available at Bioinformatics Advances online.

1 Introduction

Research in network medicine has to a high extend focused on developing new algorithms for inference of disease modules (Bader and Hogue, 2003; Bruhn et al., 2014; Ghiassian et al., 2015; Gustafsson et al., 2014a; Langfelder and Horvath, 2008, 2012; Vlaic et al., 2018) and benchmarking these (Badam et al., 2021; Choobdar et al., 2019; de Weerd et al., 2020) to come up with ‘gold standard’ approaches in this field. We recently developed MODifieR (de Weerd et al., 2020) to integrate the result of eight different module detection methods and evaluate their performance. However, it remains challenging to translate these modules into clinical applications by for example utilizing them to uncover biomarkers associated with diseases. A potential reason for this include that modules often are large and non-specific, often covering plenty of genes, which could not be associated directly using the omic that was used for detection. We and others have used modules as a first coarse-graining basis for extracting most relevant disease-associated genes by integrating knowledge about regulatory mechanisms inside the modules that could potentially explain the driving force behind the disease processes or the effect of a certain treatment. The concept of forming such modules by integrating variables of importance for pathogenesis could also serve as a ‘digital twin’ model to individualize patient treatment bringing us closer to precision medicine (Björnsson et al., 2019; Cirillo and Valencia, 2019; Gligorijević et al., 2016; Malod-Dognin et al., 2018).

Another type of popular network tool that has been used successfully in systems medicine is hub analysis, which aims at identifying a minimal set of upstream core regulators (often <10) that are significantly enriched regulating a gene set (Gustafsson et al., 2015;
Lefebvre et al., 2010; Liu et al., 2019). Hub analysis has been hampered by similar problems as modules, i.e. lack of agreement in gold-standard data that has hampered model comparisons and development. This led us to recently develop ComHub for robust prediction of hubs in gene regulatory networks combining the prediction of six different methods (Åkesson et al., 2021).

Modules and upstream hubs are rarely combined except for some individual methods like WGCNA (Langfelder and Horvath, 2008, 2012). We therefore developed MODalyseR, which combines MODifieR and ComHub in a systems medicine pipeline tool. To increase its usability for a great variety of researchers, we have included it as a R shiny web application in a docker image to run on all operating systems, with the option to export computationally demanding methods as an automated nextflow (Di Tommaso et al., 2017) pipeline to efficiently run on computer clusters. As a showcase, we applied MODalyseR to white matter lesions from patients with multiple sclerosis (MS).

In MS parts of the white matter of the central nervous system is damaged by immune cells. MS is one of the most common causes of disability in young adults (MSIF, 2020), and is a highly multifactorial autoimmune disease with more than 200 associated interacting single-nucleotide polymorphisms (SNP) and environmentally associated risk factors (Olsson et al., 2017). These factors have in several studies shown to co-localize within MS modules of the protein interaction network (Gustafsson et al., 2014b; Hellberg et al., 2016; Menche et al., 2015), which further also are upstream regulated through hub transcription factors (Gustafsson et al., 2015). MODalyseR first identified modules of 1200 proteins that were independently confirmed MS enriched. Through our hub enrichment analysis the modules were narrowed down to 40 proteins and IKZF1 identified as an upstream hub TF, which was again supported by independent data. We therefore believe MODalyseR to be an important contribution to reproducible research and standardized network analysis for clinicians and bioinformatic support services.

2 Methods
2.1 Implementation of application
MODalyseR is a standalone interactive R Shiny web application implemented in the golem framework (Guyader et al., 2020) and shipped as a Docker container that can be run on a local machine. The frontend of the application is an interactive reactive website accessible in a web browser on the localhost. The backend is implemented in the R programming language and consists of the functions implemented in the MODifieR R package and ComHub Python library with in addition functions to integrate the result of each of the packages into a single module. Persistent storage of input data and results between user sessions is provided by an SQLite database integrated in the application. To interface with the database, the SQLite R package (Müller et al., 2020) is used. Cytoscape 3.8.2 (Shannon et al., 2003) is included in the Docker image and the RcY3 (Ono et al., 2015) package is used to communicate between MODalyseR and Cytoscape. For computationally demanding methods and datasets a self-contained nextflow script that automatically downloads a singularity image (Kurtzer et al., 2017) is provided to execute the methods on an external computing source, for example a High Performance Computer cluster. The output of this nextflow analysis can then be uploaded to MODalyseR.

2.2 Pipeline
MODalyseR implements a pipeline for disease module analysis (Fig. 1). First, disease modules are inferred from transcriptomics data. Second, the inferred modules are filtered based on hub regulation resulting in a core set of genes with strong disease association. The pipeline is built on the previously developed tools MODifieR and ComHub. MODifieR includes eight commonly used network inference methods for disease module inference, all which are integrated in MODalyseR. These methods often predict modules containing hundreds or thousands of genes. To further limit these modules to the genes with the strongest disease association, hubs inferred with ComHub are used to filter the modules based on hub regulation.

ComHub robustly predicts hub regulators by averaging over the predictions by a compendium of gene regulatory network inference methods. MODalyseR includes seven commonly used network inference methods from the ComHub package, with the possibility to also include in-house network predictions. The output of ComHub is the average outdegree of each regulator selecting the TFs with highest degree as hubs. The default value in ComHub is set to 10%, which previously resulted in highest performance (Åkesson et al., 2021), but can be changed in the software by the user.

Interactions between hubs and module genes are determined by overlaying the hubs and module genes on a gene regulatory network. In MODalyseR, a consensus network inferred by averaging over the predictions of the available gene regulatory network inference methods from ComHub is used by default. Module genes, which are regulated by hubs, are kept in the module. A hub should have a significant overrepresentation of interactions with module genes compared to the remaining network for the hub to be used for filtering the module genes, assessed with Fisher’s exact test (odds ratio >2 and P < 0.05). This approach results in a smaller module containing more disease specific genes.

MODalyseR includes several options for performing functional and disease analysis of inferred modules. First, gene set enrichment analysis of GO terms, KEGG pathways and disease ontology terms. Second, genes can be labeled with user-chosen terms, such as known biomarkers or protein secretion. Third, modules can be easily exported to Cytoscape for visualization.

2.3 Use case
We applied MODalyseR on RNA-seq data of MS white matter lesions with Gene Expression Omnibus (GEO) accession number GSE138614 (Eikjær et al., 2019). We chose this dataset as we could directly compare the success of our pipeline using active and non-active lesions as positive and negative controls, respectively. The dataset consisted of 71 white matter samples from 10 patients with progressive MS and 25 control white matter samples from 5 donors. The MS samples were from five different white matter areas: 21 Normal appearing white matter (NAWM), 3 remyelinating lesions, 15 active lesions, 14 inactive lesions and 16 chronic active lesions. Differential gene expression analysis for each contrast was acquired from Supplementary Material in GEO, in addition to the raw RNA-seq count data. The count data were normalized with the function varianceStabilizingTransformation from the R package DESeq2 (Love et al., 2014) and subsequently quantile normalized using normalize.quantiles from the R package preprocessCore (Bolstad, 2019). The R script used for preprocessing can be found in Supplementary File S1.

2.4 Chip-seq confirmation of MS-associated targets of IKZF1
We validated the TF IKZF1 as a regulator of MS by calculating the enrichment of SNPs associated with MS in IKZF1 binding regions. The enrichment of MS SNPs were calculated using the function permTest from the R package regioneR (Gel et al., 2018), which performs a permutation test. SNPs significantly associated with MS, using a nominal P < 1 × 10−5 as we previously did (Hellberg et al., 2016), were obtained from Sawcer et al. (2011). We used all SNVs in linkage disequilibrium with the MS-associated SNPs computed using SNiPA (Arnold et al., 2015) (r2 > 0.8, 1000 Genomes Phase 3 v5 variant set, European population). There were 11 249 SNVs in linkage disequilibrium with the 710 MS SNPs. IKZF1-binding sites from IKZF1 ChIP-seq on human GM12878 were downloaded from the ENCODE portal (Davis et al., 2018) (https://www.encodeproject.org/). We used conservative idr thresholded peaks with identifier ENCF448JWW. Common SNVs in NCBI dbSNP Build 153 with RegulomeDB rank scores were downloaded from RegulomeDB (Boyle et al., 2012) (https://www.regulomedb.org/). SNVs and IKZF1 binding regions were annotated with promoter regions using the function annotatePeak from the R package ChIPseeker (Yu et al., 2015).
3 Results

3.1 MODalyseR

Tools for network medicine has so far most often focused on individual network-based concepts alone and often require the installation of several dependencies making the combinations of multiple different concepts complicated for medicine-oriented labs and others. In addition, many of the state-of-the-art methods take substantial computational power limiting the efficiency of online services for this service. We therefore created MODalyseR as a standalone interactive R Shiny (Chang et al., 2020) web application implemented in the golem (Guyader et al., 2020) framework and shipped as a Docker image that can easily be run on a local machine. The MODalyseR docker image is available at https://hub.docker.com/r/ddeweerd/modalyser and a detailed user guide and installation instructions can be found at https://gustafsson-lab.gitlab.io/MODalyseR/. MODalyseR incorporates eight disease module inference methods previously described in the MODifieR package. In addition, the software includes hub detection from the python package ComHub to infer hub regulators of the disease modules using seven different methods. Both types of methods can easily be run after each other using mRNA expression data and an external network of interest as input. We implemented enrichment analysis for the user to see progress and being able to adjust hyperparameters at the different steps based on functional data.

The interface of MODalyseR is divided into three main tabs, namely tools, visualization and database. The tools tab (Fig. 2a) is further divided into three columns of related functions following the general flow (Fig. 1) of the analysis. The first block, 1. Upload data deals with the uploading of all input data. Disease module inference and subsequent hub detection is performed in the second block 2. Inference methods. In the third block 3. Analyze predictions the user can set and rerun enrichment analysis for finetuning in an iterative procedure.

The input for MODalyseR is a set of differentially expressed genes, a normalized RNA-seq count matrix and optional annotation. The input data are wrapped in an input object that is used for all disease module inference methods. Additionally, a protein–protein interaction (PPI) network is required for certain disease module inference methods. MODalyseR offers a built-in function to obtain a STRING PPI (Szklarczyk et al., 2019) network, which can be set using a custom confidence score cutoff. The MODifieR input object can easily be converted to a ComHub object to work with the gene regulatory inference methods. After processing the data in the general workflow as described in Figure 1, the result consists of module objects containing a hub-filtered list of module genes that can be used for enrichment analysis or visualized in Cytoscape with optional annotation added.

3.2 Hub-based module filtering revealed IKZF1 as a putative MS-associated TF regulator, which was justified by independent ChIP-seq data

To demonstrate the applicability of MODalyseR, we applied it to a recent dataset of MS including five types of areas of the white matter lesions, hereafter called active lesion, chronic active lesion, remyelinating lesion, inactive lesion and NAWM (Table 1). The three first represent inflammatory areas and the two last are expected to be more similar to controls. First, we computed modules using default MODalyseR parameters for each of the eight module inference methods across lesions and calculated disease significance (Step 4: Fig. 1). This analysis showed the CliqueSum modules to yield highest significance for MS according to Disease Gene Ontology (DGO), which also is consistent with our previous benchmark study (Badam et al., 2021). The modules covered 1193–1498 genes each and three were enriched for MS, whereof the non-inflammatory NAWM was one. Next, we proceeded with hub filtering, which resulted in empty modules for the inactive lesion and NAWM and modules of size 66, 32 and 38 genes for the active lesions (Fig. 3), chronic active lesions (Supplementary Fig. S1) and remyelinating lesions (Supplementary Fig. S2), respectively. DGO analysis showed these modules to be significantly enriched of MS and three hub TFs were enriched for bindings in these filtered modules, namely IKZF1 (all three lesions), IRF8 (two lesions) and TFEC (one lesion) (Table 1). MS were among the two most enriched diseases for all three modules (Supplementary Tables S1–S3). In addition, these modules were enriched for KEGG pathways, such as Th17 cell differentiation and Th1 and Th2 cell differentiation, which indicates inflammatory activity (Supplementary Tables S4–S6).

As our interactions were derived from mRNA correlations, we next aimed to test if they could act directly between the TF and target promoters. As IKZF1 was consistently identified in all active lesions, we reasoned that it plays a central role in lesions and therefore mapped DNA binding regions from ChIP-seq data of IKZF1 using data from ENCODE and RegulomeDB. Indeed, we found that all binding regions of IKZF1 were generally enriched for MS-associations, we reasoned that it plays a central role in lesions and therefore mapped DNA binding regions from ChIP-seq data of IKZF1 using data from ENCODE and RegulomeDB. Indeed, we found that all binding regions of IKZF1 were generally enriched for MS-associated SNPs both at distal and promoter positions (permutation test $P < 10^{-6}$) (Fig. 4a). Lastly, we tested if those SNPs were annotated (using RegulomeDB) to have a downstream effect of the target gene expression (so-called eQTLs) and found a highly significant 10- and 7-fold enrichment versus intergenic ($P = 10^{-50}$ and promoter regions ($P = 2 \times 10^{-7}$), respectively (Fig. 4b). Taken together, our use case shows that our implemented two network-based methods
can be combined for increased sensitivity and detect direct upstream important regulators.

4 Discussion

Disease modules are often used to model complex diseases with the aim to uncover disease biomarkers, which could be used for early diagnosis and personalized treatments. However, existing methods often include many false positive genes in the disease modules, making it difficult to utilize the modules in clinical applications. Studies in the field of network-based disease module identification have been restricted to present new module inference methods and apply modules to explore disease mechanisms with the aim to identify gene candidates associated with the studied disease. While these studies are often delimited to only use PPI networks in combination with some omics data, we propose a novel approach that adds an additional layer of knowledge in terms of regulatory networks, and more specifically regulatory hubs. To our knowledge, it is a unique way to enhance disease modules approach with a regulatory hubs layer, which can accelerate future clinical use, as hubs have gained huge importance in systems medicine and are commonly recognized as potential disease biomarker candidates. With MODalyseR, we do not only introduce a novel scientific approach that integrates
regulatory mechanisms within sub-modules of highly interacting genes with shared upstream regulators to identify a core set of genes with strong disease association. We also provide user-friendly software for network-based disease analysis that enables biomedical researchers without computational background to explore disease-related mechanisms, and select interesting biomarker candidates.

MODalyseR implements a pipeline, which reduces modules to the most important genes, which could serve well as potential robust biomarkers. With MODalyseR, we predicted highly MS enriched modules and found IKZF1 as a possible master regulator of MS in white matter lesions. IKZF1 is an important TF for the development of lymphocytes and has previously been connected to leukemia and autoimmune diseases (Hoshino et al., 2017; Marke et al., 2018). Validation of IKZF1 using independent Chip-seq analysis showed a strong significant enrichment of MS-associated SNPs in IKZF1 binding regions, which also were enriched for eQTLs. Hence, the regulatory mechanism of IKZF1 is likely altered in MS affecting downstream genes. This case study shows how MODalyseR can be used to elucidate the regulatory mechanism of complex diseases by both identifying a limited set of disease-associated genes and the main regulators of those genes. There are also other pipelines, such as NetFinder (Ahsen et al., 2019), which are aimed at predicting regulators of gene expression-based biomarkers. However, NetFinder requires a set of biomarkers to already exist while MODalyseR is provided as a standalone web application and not as a web server tool. While this demands an installation of docker to use the tool, it was a necessary decision to provide a tool, which can run computationally demanding inference methods. The option to download and run module and network inference methods as a nextflow pipeline, locally or on a cluster, we believe increases the applicability of the software by making it possible to use the tool on large datasets.

In summary, most complex diseases have been analyzed extensively and studied using RNA-seq data. This makes MODalyseR a generally applicable multi-purpose systems medicine tool to be applied on existing disease to generate prior hypothesis testing through easy access to many of the most popular network-based tools using state-of-the-art default parameters. The pipeline also allows simultaneous testing of many methods and datasets in parallel and automatically evaluates the methods best suitable for a particular disease, which we previously saw can vary greatly.

**Fig. 3.** The hubs IKZF1, IRF8 and TFEC (red) were identified as significant regulators of the active lesion CliqueSum module. The module was reduced to 66 genes (blue) with strong enrichment of MS

**Fig. 4.** IKZF1-binding regions were enriched for MS SNPs. (a) Proportion of MS SNPs among common SNVs in intergenic regions compared to IKZF1-binding sites, both for the genome and promoter regions. IKZF1-binding sites were significantly enriched for MS SNPs evaluated with a permutation test. (b) Proportion of eQTLs among SNVs in intergenic regions compared to proportion of eQTLs among MS SNPs in IKZF1-binding sites, both for the genome and promoter regions. MS SNPs in IKZF1-binding sites were significantly enriched for eQTLs evaluated with Fisher’s exact test.
Moreover, the advanced user can refine the networks further by integrating other data-sources and further combining the best methods and datasets as we previously shown (Badam et al., 2021).

Lastly, the use case stresses how this can be done in practice for in principle any complex disease where RNA-seq data are available.

Acknowledgements

The authors thank the iGEM Linköping 2020 team for their work on the iGEM 2020 competition software submission, which was extended into MODalyseR. National Supercomputer Centre (NSC; LiU-compute-2020-21) is acknowledged for computing resources.

Author contributions

H.A.d.W. and J.Å. developed and implemented the application, performed the use case, wrote the software documentation and built the Docker and Singularity images. H.A.d.W., J.Å., Z.L.-P. and M.G. designed the study and methodology, Z.L.-P. and M.G. were responsible for project administration. H.A.d.W., J.Å., D.G., Z.L.-P. and M.G. wrote, revised and approved the final manuscript.

Funding

This work was supported by the Knowledge Foundation (dnr HSK219/26); Swedish Foundation for Strategic Research (SB16-0011); and Swedish Research Council [grant 2019-04193].

Conflict of Interest: none declared.

References


