Network diffusion on multiple-layers: current approaches and integrative analysis of Rheumatoid Arthritis data.

Noemi Di Nanni[1,2], Matteo Gnocchi[1], Marco Moscatelli[1], Luciano Milanesi[1] and Ettore Mosca[1]

[1]Institute of Biomedical Technologies - CNR, Segrate (MI), Italy
[2]Department of Industrial and Information Engineering – University of Pavia, Pavia, Italy
Contact: noemi.dinanni@itb.cnr.it, ettore.mosca@itb.cnr.it

Introduction

The principle of spreading information throughout a network (Network Diffusion, shortly ND) has been applied to solve several problems in biological data analysis[1]. In fact, the known and/or predicted interactions among gene products can be used to analyze omics-derived quantities at a systems-level. More recently, the success of ND on a "single layer" prompted the development of the first approaches on multiple layers. Here, we review current ND-based methods for the analysis of multi-omics data and show preliminary results on the application of ND to genetic variation, DNA methylation and gene expression data collected in a study on Rheumatoid Arthritis (RA).

Methods

RA data were obtained from a previous study[2], in which a total of 88 genes were found significant in a genome-wide association analysis, 2291 genes were differentially methylated and 2831 genes were differentially expressed between RA fibroblast-like synoviocytes and control cells (the reported numbers were obtained after the exclusion of a few deprecated genes). The genome-scale network containing direct and indirect interactions between gene products was defined using NCBI interaction data[3]. For each gene g_i, we used the steady state values (x_i, y_i, z_i) found by means of the network propagation algorithm described in Bersanelli et al.[4] for quantifying the network proximity of g_i to altered genes of the three lists. In each list, we selected a number of top ranking genes with the highest steady state values equal to the number of genes occurring in the initial lists. OMIM[5] and KEGG[6] pathway databases were used as source of genes known to be associated with RA, while PubMatrix[7] was used to collect the number of articles mentioning RA genes.

Results

ND was applied for the integrative analysis of multiple omics datasets with several goals (Table 1).

Method	Implementation	Goal	Network type
CATAPULT[8]	Matlab	Gene prioritization	Heterogeneous network
EMDN[9]	R	Module detection	Co-expression and co-methylation networks
Mashup[10]	Matlab	Function prediction	Multiple networks
M – module[11]	R	Module detection	Multiple co-expression networks
RegNet[12]	R	Impact of gene expression on user-defined target genes	2 omics, 1 network
Ruffalo et al.[13]	NA	Detection cancer driver genes	2 omics, 1 network
SNF[14]	Matlab, R	Disease subtype and patient stratification	Fusion of multiple networks into one
TieDie[15]	SciPy, Matlab	Module detection	2 omics, 1 network

Table 1 | Network diffusion-based methods for the integrative analysis of multiple "-omics".

CATAPULT[8] is a method that uses ND on a heterogeneous gene-trait network to derive the features of a biased support vector machine for inferring gene-phenotype associations with the combination of data across species. Mashup[10] applies ND on each of several networks to characterize the topological context of each node; then, low-dimensional feature vectors, describing the topological properties of each node, are calculated by jointly analyzing the diffusion states of all networks to predict genetic interactions and protein and gene function. Ma et al.[11] proposed the M-Module framework to identify gene modules from multiple co-expression networks, where ND is used to incorporate prior (gene mutation) information in each network. EMDN[9] algorithm is based on M-Module, but is specifically applied for the analysis of gene co-expression and co-methylation. TieDie[15] applies ND to identify a subnetwork that links a source gene set (genomic alterations) to a target gene set (gene expression changes) on the same a priori network. SNF[14] uses ND to fuse multiple patient similarity networks, defined using different omics, into a unique patient similarity network that indicates disease subtypes. Ruffalo et al.[13] uses ND to jointly analyse mutated genes and differentially expressed genes on the same network; the authors combine the diffusion scores in different ways to derive input values for a logistic regression model that predicts gene association with the pathology of interest. RegNet[12] is an R package that uses ND to quantify the impact of gene expression changes on user-defined target genes in a network inferred from gene expression and copy number data.

We applied ND to jointly analyse genetic alterations, differentially methylated genes and differentially expressed genes found in RA fibroblast-like synoviocytes from a previous study[2].

For each gene list, ND induces a gene ranking that takes into account also the a priori known interactions between all human genes. ND-based gene rankings showed higher overlaps between the two pairs of lists and among all three lists. Further, ND-based gene rankings were richer of known RA genes than the initial lists: two examples are the genes HLA-DRB1 and CCL2. The majority of the genes not occurring in initial lists, but highly ranked by ND in relation to at least 2 types of alteration, are mentioned in the literature of RA.

Our preliminary results are in line with previous evidences indicating that ND contributes to highlight interesting patterns in multi-omics data analysis. Further studies are requested to improve current methods in relation to different research questions, types of omics, approaches to define networks (inference from data or a priori knowledge or both), adjustment of ND values (e.g. overestimation of hubs), just to mention a few.

References

1. Cowen et al. Network propagation: a universal amplifier of genetic associations. Nat Reviews Genetics, 2017, 18(9): 551-562. doi: 10.1038/nrg.2017.38
2. Whitaker et al. Integrative Omics Analysis of Rheumatoid Arthritis Identifies Non-Obvious Therapeutic Targets. PLoS ONE, 2015, 10(4): e0124254. doi: 10.1371/journal.pone.0124254
3. Brown et al. Gene: a gene-centered information resource at NCBI. Nucleic Acids Research, 2015, 43 (Database issue): D36–D42. doi: 10.1093/nar/gku1055
4. Bersanelli et al. Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules. Scientific Reports, 2016, 6: 34841. doi: 10.1038/srep34841
5. Hamosh et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research, 2005, 33 (Database Issue): D514–D517. doi: 10.1093/nar/gki033
6. Kaneisa et Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research, 2000, 28(1): 27–30. doi: https://doi.org/10.1093/nar/28.1.27
7. G Becker et al. PubMatrix: a tool for multiplex literature mining. BMC Bioinf., 2003, 4:61. doi: 10.1186/1471-2105-4-61
8. Singh-Blom et al. Prediction and validation of gene-disease associations using methods inspired by social network analyses. PLoS One, 2013, 8(5):e58977. doi: 10.1371/journal.pone.0058977
9. Ma et al. Multiple network algorithm for epigenetic modules via the integration of genome-wide DNA methylation and gene expression data. BMC Bioinf., 2017, 18:72. doi: 10.1186/s12859-017-1490-6
10. Cho et al. Compact Integration of Multi-Network Topology for Functional Analysis of Genes. Cell Systems, 2016, 3: 540–548. doi: https://doi.org/10.1016/j.cels.2016.10.017
11. Ma et al. Modeling disease progression using dynamics of pathway connectivity. Bioinformatics, 2014, 30(16):2343–50. doi: 10.1093/bioinformatics/btu298

12. Seifert et al. regNet: An R package for network-based propagation of gene expression alterations. Bioinformatics, 2017, btx544, 1-4, https://doi.org/10.1093/bioinformatics/btx544
13. Ruffalo et al. Network-based integration of disparate omic data to identify 'silent players' in cancer. PLoS Comput. Biol, 2015, 11(12): e1004595. doi:10.1371/journal. pcbi.1004595
14. Wang et al. Similarity network fusion for aggregating data types on a genomic scale. Nat. Methods, 2014, 11(3): 333–337. doi:10.1038/NMETH.2810
15. Paull et al. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). Bioinformatics, 2013, 29(21), 2757–2764. doi:10.1093/bioinformatics/btt471