EVALUATING SOURCES OF VARIABILITY IN PATHWAY PROFILING

Annalisa Barla\(^{(1)}\), Samantha Riccadonna\(^{(2)}\)\(^{∗}\), Salvatore Masecchia\(^{(1)}\), Margherita Squillario\(^{(1)}\), Michele Filosi\(^{(2,3)}\), Giuseppe Jurman\(^{(2)}\) and Cesare Furlanello\(^{(2)}\)\(^{†}\)

\(^{(1)}\)DISI, University of Genoa, via Dodecaneso 35, I-16146 Genova, Italy.
\(^{(2)}\)FBK, via Sommarive 18, I-38123 Povo (Trento), Italy.
\(^{(3)}\)CIBIO, University of Trento, via Delle Regole 101, I-38123 Mattarello (Trento), Italy

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: A bioinformatics platform is introduced aimed at identifying models of disease-specific pathways, as well as a set of network measures that can quantify changes in terms of global structure or single link disruptions. The approach integrates a network comparison framework with machine learning molecular profiling. The platform includes different tools combined in one Open Source pipeline, supporting reproducibility of the analysis. We describe here the computational pipeline and explore the main sources of variability that can affect the results, namely the classifier, the feature ranking/selection algorithm, the enrichment procedure, the inference method and the networks comparison function.

Results: The proposed pipeline is tested on a microarray dataset of late stage Parkinsons’ Disease patients together with healthy controls. Choosing different machine learning approaches we get low pathway profiling overlapping in terms of common enriched elements. Nevertheless, they identify different but equally meaningful biological aspects of the same process, suggesting the integration of information across different methods as the best overall strategy.

Availability: All the elements of the proposed pipeline are available as Open Source Software: availability details are provided in the main text.

Contact: annalisa.barla@unige.it

1 INTRODUCTION

We present a computational framework for the study of reproducibility in network medicine studies (Barabasi et al., 2011). Networks, molecular pathways in particular, are increasingly looked at as a better organized and more rich version of gene signatures. However, high variability can be injected by the different methods that are typically used in system biology to define a cellular wiring diagram at diverse levels of organization, from transcriptomics to signalling, of the functional design. For example, to identify the link between changes in graph structures and disease, we choose and combine in a workflow a classifier, the feature ranking/selection algorithm, the enrichment procedure, the inference method and the networks comparison function. Each of these components is a potential source of variability, as shown in the case of biomarkers from microarrays (The MAQC Consortium, 2010). Considerable efforts have been directed to tackle the problem of poor reproducibility of biomarker signatures derived from high-throughput -omics data (The MAQC Consortium, 2010), addressing the issues of selection bias (Ambroise and McLachlan, 2002; Furlanello et al., 2003) and more recently of pervasive batch effects (Leek et al., 2010). We argue that it is now urgent to adopt a similar approach for network medicine studies. Stability (and thus reproducibility) in this class of studies is still an open problem (Baralla et al., 2008). Underdeterminacy is a major issue (De Smet and Marchal, 2010), as the ratio between network dimension (number of nodes) and the number of available measurements to infer interactions plays a key role for the stability of the reconstructed structure. Furthermore, the most interesting applications are based on inferring networks topology and wiring from high-throughput noisy measurements (He et al., 2009).

Despite its common use even in biological contexts (Sharan and Ideker, 2006), the problem of quantitatively comparing networks (e.g., using a metric instead of evaluating network properties) is a still an open issue in many scientific disciplines. The central problem is of course which network metrics should be used to evaluate stability, whether focusing on local changes or global structural changes. As discussed in Jurman et al. (2011), the classic distances in the edit family focus only on the portions of the network interested by the differences in the presence/absence of matching links. Spectral distances - based on the list of eigenvalues of the Laplacian matrix of the underlying graph - are instead particularly effective for studying global structures. In particular, the Ipsen-Mikhailov (Ipsen and Mikhailov, 2002) distance was found robust in a wide range of situations (Jurman et al., 2011). However, global distances can be tricked by isomorph or close to isomorph graphs. In Jurman et al. (2012), both approaches are improved by proposing a glocal measure which combines a spectral distance with a typical Hamming local editing component. In this paper we use this new tool to quantify how stability of network reconstruction is modified in practice by the different inference and enrichment methods.
Pathway enrichment methods are widely used in bioinformatics analysis, for example to assess the relevance of biomarker lists or as a first step in network analysis. The enrichment step is performed using the functional information stored in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa and Goto, 2000) and in the Gene Ontology (GO) database (Ashburner et al., 2000). The reconstruction of molecular pathways from high-throughput data is then based on the theory of complex networks (e.g., Barabasi et al., 2005). The functional information stored in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa and Goto, 2000) is used to feed the profiling part of the pipeline within a proper Data Analysis Protocol, which will ensure accurate and reproducible results (The MAQC Consortium, 2010). The prediction model $\mathcal{M}$ is built by using two different algorithms for classification and feature ranking. The more recent one is the $\ell_1/\ell_2$ regularization with double optimization, capable of selecting subsets of discriminative genes. The algorithm can be tuned to give a minimal set of discriminative genes or larger sets including correlated genes and it is based on the optimization principle presented in Zou and Hastie (2005). The implementation used consists of two stages (De Mol et al., 2008) and it is cast in nested loops of 10-fold cross-validation (Barla et al., 2009). The first stage identifies the minimal set of relevant variables (in terms of prediction error), while, starting from the minimal list, the second one selects the family of (almost completely) nested lists of relevant variables for increasing values of linear correlation. As alternative choice we consider Liblinear, a linear Support Vector Machine (SVM) classifier specifically designed for large datasets (millions of instances and features) (Fan et al., 2008). In particular, the classical dual optimization problem with L2-SVM loss function is solved with a coordinate descent method. For our experiment we adopt the $\ell_2$-regularized penalty term and the module of the weights for ranking purposes within a $100 \times 3$-fold cross validation schema. We build a model for increasing feature subsists where the feature ranking is defined according to the importance for the classifier. We choose the model, and thus the top ranked features, providing a balance between the accuracy of the classifier and the stability of the signature (The MAQC Consortium, 2010). Thus, the output of this first step is a gene signature $g_1, \ldots, g_k$ (one for each model $\mathcal{M}$) containing the $k$ most discriminant features, ranked according their frequency score.

### Pathway Enrichment

The successive enrichment phase derives a list of relevant pathways from the discriminant features, moving the focus of the analysis from single genes to functionally related pathways. As outlined in the review by Huang et al. (2009), in the last 10 years the gene-annotation enrichment analysis field has been growing rapidly and several bioinformatics tools have been designed for this task. Huang et al. (2009) provide a unique categorization of these enrichment tools in three major categories based on the underlying algorithm: singular enrichment analysis (SEA), gene set enrichment analysis (GSEA), and modular enrichment analysis (MEA). We choose one representative $\mathcal{E}$ for each class for our comparison referring as sources of information $\mathcal{D}$ both to the KEGG, to explore known information on molecular interaction networks, and GO, to explore functional characterization and biological annotation. In the first category we choose WebGestalt (WG), an online gene set analysis toolkit (Zhang et al., 2005a) taking as input a list of relevant genes/probesets. The enrichment analysis is performed in KEGG and GO identifying the most relevant pathways and ontologies in the signatures. WG adopts the hypergeometric test to evaluate functional category enrichment and performs a multiple test adjustment (the default method is the one from Benjamin and Hochberg [1995]). The user may choose different significance levels and the minimum number of genes belonging to the selected functional groups. GSEA (Subramanian et al., 2005) is our representative of the second class. It first performs a correlation analysis between the features and the phenotype obtaining a ranked list of features. Secondly it determines whether the members of given gene sets are randomly distributed in the ranked list of features obtained above, or primarily found at the top or bottom. We use the preranked analysis tool, feeding the ranked lists of genes produced by the profiling phase directly to the enrichment step of GSEA. To avoid a miscalculation of the enrichment score ES, we provide as input the complete list of variables (not just the selected ones), assigning to the not-selected a zero score. Note that GSEA calculates enrichment scores that reflect the degree to which a pathway is overrepresented at the top or the bottom of the ranked list. In our analysis we considered only pathways enriched with the top of the list. Finally, the tool in the MEA class is the Pathways and Literature Strainer (PaLS) (Aibes et al., 2008), which takes a list or a set of lists of genes (or protein identifiers) and shows which ones share the same GO terms or KEGG pathways, following a criterion based on a...
having more than problem, in the following experiments we limit the analysisto pathways intrinsic underdeterminacy of the task. As an additional caution against this sole genes belonging to the pathway
to the relevance networks class of algorithms and is employed for the processing inequality property (Cover and Thomas, 1991). CLR belongs
cells, it is able to address a wider range of network deconvolution problems. designed for handling the complexity of regulatory networks in mammalian
Fig. 1. General scheme of the analysis pipeline, with the indication of the algorithms and tools used in the PD dataset application (lower boxes).
threshold \( t \) set by the user. The tool provides as output those functional groups that are shared at least by the \% of the selected genes. PaLS is aimed at easing the biological interpretation of results from studies of differential expression and gene selection, without assigning any statistical significance to the final output. Applying the above mentioned pathway enrichment techniques, we retrieve for each gene \( g_i \) the corresponding whole pathway \( p_i = \{ i_1, \ldots, i_h \} \), where the genes \( i_j \neq g_i \) not necessarily belong to the original signature \( g_1, \ldots, g_p \). Extending the analysis to all the \( h_j \) genes of the pathway allows us to explore functional interactions that would otherwise get lost.

**Subnetwork Inference**

For each pathway, networks are inferred separately on data from the different classes. The subnetwork inference phase requires to reconstruct a network \( N_{p_i, y} \) on the pathway \( p_i \) by using the steady state expression data of the samples of each class \( y \). The network inference procedure is limited to the sole genes belonging to the pathway \( p_i \) in order to avoid the problem of intrinsic underdeterminacy of the task. As an additional caution against this problem, in the following experiments we limit the analysis to pathways having more than 4 nodes and less than 10000 nodes. The pipeline allows to run analysis in parallel with different methods and thus to evaluate the variability along the whole pipeline. We adopt four different subnetwork reconstruction algorithms \( \mathcal{V} \): the Weighted Gene Co-Expression Networks (WGCN) algorithm (Harvard, 2011), the Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) (Margolin et al., 2006), the Context Likelihood of Relatedness (CLR) approach (Faith et al., 2007), and the Reverse Engineering Gene Networks using Artificial Neural Networks (RegnANN) (Grimaldi et al., 2011). In this work, we applied WGCNA, CLR and ARACNE to analyze the pathway identified in the Pathway Enrichment step, while RegnANN was used, as an alternative algorithm, to reconstruct interesting disrupted pathways and to compare its results with results from methods mentioned above. WGCNA is based on the idea of using (a function of) the absolute correlation between the expression of a couple of genes across the samples to define a link between them. ARACNE is a method for inferring networks from the transcription level (Margolin et al., 2006) to the metabolic level (Nemenman et al., 2002). Beside it was originally designed for handling the complexity of regulatory networks in mammalian cells, it is able to address a wider range of network deconvolution problems. This information-theoretic algorithm removes the vast majority of indirect candidate interactions inferred by co-expression methods by using the data processing inequality property (Cover and Thomas, 1993). CLR belongs to the relevance networks class of algorithms and is employed for the identification of transcriptional regulatory interactions (Faith et al., 2007). In particular, interactions between transcription factors and gene targets are scored by using the mutual information between the corresponding gene expression levels coupled with an adaptive background correction step. Indeed the most probable regulator-target interactions are chosen comparing the mutual information score versus the “background” distribution of mutual information scores for all possible pairs within the corresponding network context (i.e. all the pairs including either the regulator or the target). RegnANN is a newly defined method for inferring gene regulatory networks based on an ensemble of feed-forward multilayer perceptrons. Correlation is used to define gene interactions. For each gene a one-to-many regressor is trained using the transcription data to learn the relationship between the gene and all the other genes of the network. The interaction among genes is estimated independently and the overall network is obtained by joining all the neighborhoods. Summarizing, we obtain a real-valued adjacency matrix as output of the subnetwork inference step for each dataset \( X \), for each class \( y \), for each model \( M \), for each enrichment tool \( E \), for each source of information \( D \), for each pathway \( p_i \), and for each subnetwork inference algorithm \( \mathcal{N} \). We thus need to quantitatively evaluate network differences, i.e. using a metric instead of evaluating network properties.

**Subnetwork Distance and Stability**

Among the possible choices already available in literature, we focus on two of the most common distance families: the set of edit-like distances and the spectral distances. The functions in the former family quantitatively evaluate the differences between two networks (with the same number of nodes) in terms of minimum number of edit operations (with possibly different costs) transforming one network into the other, i.e. deletion and insertion of links, while spectral measures relies on functions of the eigenvalues of one of the connectivity matrices of the underlying graph. As discussed in Jurman et al. (2011), the drawback of many classical distances (such as those of the edit family) is locality, that is focusing only on the portions of the network interested by the differences in the presence/absence of matching links. Spectral distances can overcome this problem considering the global structure of the compared topologies. Within them, we consider the Ipsen-Mikhailov \( \epsilon \) distance: originally introduced in Ipsen and Mikhailov (2002) as a tool for network reconstruction from its Laplacian spectrum, it has been proven to be the most robust in a wide range of situations by Jurman et al. (2011). We are also aware that spectral measures are not flawless: they cannot distinguish isomorphic or isospectral graphs, which can correspond to quite different conditions within the biological context. We thus introduce the global distance \( \phi \) as a possible solution against both issues: \( \phi \) is defined as the product metric of the Hamming distance \( H \) (as representative of the edit-family) and the \( \epsilon \) distance. Full mathematical details are available in Jurman et al. (2012).

Relying on the distances \( \epsilon \) and \( \phi \), we evaluate networks corresponding to the same pathway for different classes, *i.e.* all the pairs \((N_{p_i, +1}, N_{p_i, -1})\) and rank the pathways themselves from the most to the least changing across classes.

Moreover, we attached to each network a quantitative measure of stability with respect to data subsampling, in order to evaluate the reliability of inferred topologies. In particular, for each \( N_{p_i, y} \), we extracted a random subsampling (of a fraction \( r \) of \( X \) labelled as \( y \)) on which the corresponding \( N_{p_i, y} \) will be reconstructed. Repeating \( m \) times the subsampling/inferring
is available at http://slipguru.disi.unige.it/Software/L1L2Py. Liblinear was library (http://mlpy.fbk.eu). We adopt the

| $\mathcal{M}$ | $\mathcal{D}$ | $\mathcal{E}$ | WG | GSEA | PaLS |
|---|---|---|---|---|---|
| $\ell_1\ell_2$ | GO | 114 (92) | 7 (7) | 381 (331) |
| KEGG | 43 (43) | 2 (2) | 71 (71) |
| Liblinear | GO | 83 (45) | 0 (0) | 404 (356) |
| KEGG | 56 (55) | 1 (1) | 77 (77) |

procedure, a set of $m$ nets will be generated for each $\mathcal{N}_{h,y}$. Then all mutual ($m^2$) distances are computed, and for each set of $m$ graphs we build the corresponding distance histogram. In particular, for our experiments we set $m = 20$ and $r = \frac{m}{2}$. Mean and variance of the constructed histograms will quantitatively assess the stability of the subnetwork inferred from the whole dataset: the lower the values, the higher the stability in terms of robustness to data perturbation (subsampling).

Data description and preprocessing
The presented approach is applied to PD data originally introduced in [Zhang et al. 2005b] and publicly available at Gene Expression Omnibus (GEO), with accession number GSE20292. The biological samples consist of whole substantia nigra tissue in 11 PD patients and 18 healthy controls. Expressions were measured on Affymetrix HG-U133A platform. We perform the data normalization on the raw data with the rma algorithm of the R Bioconductor affy package with a custom CDF (downloaded from BrainArray: http://brainarray.mbnm.edu) adopting Entrez identifiers.

Software Availability
The Python implementation of $\ell_1\ell_2$ regularization with double optimization is available at http://slipguru.disi.unige.it/Software/LIL2Py Liblinear was originally developed by the Machine Learning Group at the National Taiwan University and it is now available within the Python mlpy library (http://mlpy.fbk.eu). We adopt the l2regLasso.m by solving, with $C = 10^{-4}$, of course. WG is available as a web application at http://bioinfo.vanderbilt.edu/webgestalt/ GSEA is available either as a web application or a Java stand-alone tool at http://www.broadinstitute.org/gsea. PaLS is available online at http://pals.bioinfo.cnio.es as a web application. For three of the network reconstruction algorithms, we adopted the R Bioconductor implementation: the WGCNA package for WGCN, and Mi/NET (Mutual Information NETworks package) for ARACNE and CLR. In particular, we set the WGCNA soft thresholding exponent to 5, while we keep the default value for the ARACNE data processing inequality tolerance parameter (Meyer et al. 2008). Moreover, the ARACNE implementation requires all the features to have non-zero variance on each class and for consistency purposes we applied this in all experiments. RegnANN is instead available from http://sourceforge.net/projects/regnann It is implemented in C and relies on GPGPU programming paradigm for improving efficiency. The $\ell_1\ell_2$ distance $\phi$ is available upon request to the authors either as R script or Python script. The computation of the Ipsen-Mikhailov distance $\ell_1\ell_2$ is included as component of the $glocal$ script.

3 RESULTS AND DISCUSSION
The feature selection results varied accordingly to the chosen method: $\ell_1\ell_2$ identified 458 discriminant genes associated to an average prediction performance of 80.8%, while with Liblinear we selected the top-500 genes associated to an accuracy of 80% (95% bootstrap Confidence Interval: (0.78;0.83)) coupled with a stability of 0.70. The lists have 119 common genes.

The number of enriched pathways greatly varied depending on the selection and enrichment tools. With $\ell_1\ell_2$, we found globally for GO and KEGG, 157, 452 and 9 pathways as significantly enriched, for WG, PaLS and GSEA respectively. Similarly, for Liblinear, the identified pathways were: 139, 481 and 1. Table 1 reports the detailed results for model $\mathcal{M}$, enrichment $\mathcal{E}$ and database $\mathcal{D}$.

If we consider the $\ell_1\ell_2$ selection method and the enrichment performed within the GO, we may note that no common GO terms were selected across enrichment methods. A significant overlap of results was found only between WG and PaLS, with 30 GO common terms. Similar considerations may be drawn with the results from the Liblinear feature selection method. Within the GO enrichment we did not identify any common GO term among the three enrichment tools. Considering only WG and PaLS, we were able to select 12 common GO terms.

If we consider the $\ell_1\ell_2$ selection method and the enrichment performed within KEGG, two common pathways are identified across enrichment methods. A significant overlap of results was found between WG and PaLS, with 43 common terms. For Liblinear, only one common pathway was selected among the three enrichment tools. A significant overlap of results was found between WG and PaLS, with 55 common terms.

Following the pipeline, we also performed a comparison of the three network reconstruction methods. We considered the most disrupted networks, keeping for the analysis those pathways that had a $glocal$ distance greater or equal to the chosen threshold $\tau = 0.05$. The choice of such threshold was made considering the distribution of $glocal$ distances $\phi$ for the methods $\mathcal{M}$. For instance, if we consider the Liblinear selection method and the KEGG database, we have a cumulative distribution as depicted in Figure 2(a). The threshold $\tau$ is set to 0.05 and allows retaining at least 50% of pathways. The plot in Figure 2(b) represents the $glocal$ distances distribution for all enrichment methods $\mathcal{E}$ with respect to the two components of the $glocal$ distance: the Ipsen distance $\ell_1\ell_2$ and the Hamming distance $H$. The red curved line represents the threshold $\tau$ in this space. The plot in Figure 2(c) is detailed for subnetwork inference method $N$. After retaining the most distant pathways, we performed a comparison of common terms for fixed selection method $\mathcal{M}$ and database $\mathcal{D}$. The results are reported in Table 2. In Tables 3 and 4 we report the most disrupted GO terms and KEGG pathways that have a $glocal$ distance $\phi$ greater or equal to the chosen threshold $\tau$.

As an example of a selected pathway within KEGG, the networks (thresholded at edge weight 0.1 for graphic purposes) inferred...
Fig. 3. (a) Networks inferred by WGCNA algorithm for the ALS KEGG pathway for PD patients (above) and controls (below), on the same pathway for different inference algorithm. (b) WGCNA is the method showing the highest stability on the two classes. (c) Same pathway reconstructed with RegnANN.

by WGCNA (together with the corresponding stability) on the Amyotrophic Lateral Sclerosis KEGG pathway (ALS - 05014) are displayed in Figure 3. We also plot the inferred network by the RegnANN algorithm. Similarly, in Figure 4 we plot the Pathogenic E. coli infection KEGG pathway, reconstructed by WGCNA, its stability plot, and the corresponding inferred networks by the RegnANN algorithm.

Discussion

The variability in the results, as expected, strongly depends on the method of choice. For feature selection, the nature of the method is key. In the proposed pipeline we limited the impact of this step by choosing two approaches within the regularization methods family. Both classifiers adopt a \( \ell_2 \)-regularization penalty term, combined with different loss functions and, for \( \ell_1 \), \( \ell_2 \) with another regularization term. We used similar but not equal model selection protocols. Both guarantee that the results are not affected by selection-bias. In this work, the main source of variability was the choice of the gene enrichment module. Therefore, the experimenter must be careful in choosing one method or another and in using it compliantly with the experimental design. For instance, GSEA was designed for estimating the significance levels by considering separately the positively and negatively scoring gene sets within a
Table 4. Summary of most disrupted KEGG pathways common between WG and PaLS, for different models $\mathcal{M}$. Each pathway is associated to a glocal distance $\phi \geq 0.05$ for all subnetwork reconstruction algorithms $\mathcal{N}$. KEGG pathways are sorted according decreasing average $\phi$. Bold fonts represent the KEGG pathways shared by model $\mathcal{M}$.

| ID     | Pathway name                          | ID     | Pathway name                          |
|--------|----------------------------------------|--------|----------------------------------------|
| 01100  | Metabolic pathway                      | 04630  | Jak-STAT signaling pathway             |
| 05130  | Pathogenic Escherichia coli infection  | 01100  | Metabolic pathway                      |
| 04910  | Insulin signaling pathway              | 05130  | Pathogenic Escherichia coli infection  |
| 00310  | Lysine degradation                     | 04623  | Cytosolic DNA-sensing pathway          |
| 04140  | Regulation of autophagy                | 00030  | Arginine and proline metabolism        |
| 03050  | Proteasome                              | 04910  | Insulin signaling pathway              |
| 00230  | Purine metabolism                      | 05212  | Pancreatic cancer                      |
| 05014  | Amyotrophic lateral sclerosis*          | 03030  | DNA replication                        |
| 00980  | Metabolism of xenobiotics by cytochrome P450 | 05213 | Endometrial cancer                     |
| 00620  | Pyruvate metabolism                    | 04660  | T cell receptor signaling pathway      |
| 05213  | Endometrial cancer                     | 04310  | Wnt signaling pathway                  |
| 00270  | Cysteine and methionine metabolism     | 05210  | Colorectal cancer                      |
| 00240  | Pyrimidin metabolism                   | 04912  | GnRH signaling pathway                 |
| 05120  | Epithelial cell signaling in Helicobacter pylori infection | 05332 | Graft-versus-host disease              |
| 05110  | Vibrio cholerae infection              | 04520  | Adherens junction                      |
| 00020  | Citrate cycle (TCA cycle)              | 04621  | NOD-like receptor signaling pathway    |
| 00562  | Inositol phosphate metabolism          | 04370  | VEGF signaling pathway                 |
| 00600  | Sphingolipid metabolism                | 04662  | B cell receptor signaling pathway      |
| 05218  | Melanoma                               | 04722  | Neurotrophin signaling pathway         |
| 00010  | Glycolysis / Gluconeogenesis           | 05214  | Glioma                                 |
| 00051  | Fructose and mannose metabolism        | 04330  | Notch signaling pathway                |
| 04722  | Neurotrophin signaling pathway         |        |                                        |

*Note: This is the only selected pathway shared across all enrichment methods $E$."

list of genes selected with filter methods based on classical statistical tests. It is worth noting that, if one uses the preranked option, as we did, negative regulated groups might not be significant at all (we indeed discarded them). WG uses the Hypergeometrical test to assess the functional groups but, differently from GSEA, does not use any significance assessment based on permutation of phenotype labels. PaLS is the simplest methods, being just a measure of occurrences of a given descriptor in the list of selected genes. However, enrichment methods from different categories are complementary and can identify different but equally meaningful biological aspects of the same process. Thus, the integration of information across different methods is the best strategy.

Moreover, the assessment of the reconstruction distance between case and control version of the same pathways help in providing a quantitative focus on the key pathway involved in the process. The use of a distance mixing the effects of structural changes with those due to the differences in rewiring moreover warrants a more informative view on the difference assessment itself. The limited effect of different feature selection methods is confirmed by the plots in Figure 5.

For $\ell_1\ell_2$, the only most disrupted pathway shared by the three enrichment tools $E$ and the three reconstruction methods $N$ is ALS. This pathway is relevant in this context because, like PD, ALS is another neurodegenerative disease therefore they share significant biological features in particular at the mithocondrial level. Moreover at the phenotypic level the skeletal muscles of the patients are severely affects influencing the movements. In Figure 6 it is evident that a high number of interactions are established among the genes going from the control (below) to the affected (above) pathways. It is also interesting to underline that CYCS (Entrez ID: 54205) one of the hub genes (represented by a red dot in the graph) within the pathway was identified by $\ell_1\ell_2$ as discriminant. This gene is highly involved in several neurodegenerative diseases (e.g., PD, Alzheimer’s, Huntington’s) and in pathways related to cancer. Furthermore its protein is known to functions as a central component of the electron transport chain in mitochondria and to be involved in initiation of apoptosis, known cause of the neurons loss in PD. Across variable selection algorithms $M$, five highly disrupted pathways were found as shared between two of the three enrichment methods (see Table 4 bold items). In particular, we represented in Figure 4 the corresponding inferred networks. To further highlight the different outcomes occurring from the same dataset when diverse inference methods are employed, we reconstructed the ALS and Pathogenic E. coli infection by the RegnANN algorithm, which tends to spot also second order correlation among the network nodes, see Figures 3 and 4.

Two genes in the E. coli infection pathway were selected both by $\ell_1\ell_2$ and Liblinear, namely ABL1 (Entrez ID: 71) and TUBB6
(Entrez ID: 84617). ABL1 seems to play a relevant role as hub both in the WGCNA and in the RegnANN networks. ABL1 is a protooncogene that encodes protein tyrosine kinase that has been implicated in processes of cell differentiation, cell division, cell adhesion, and stress response. It was also found to be responsible for different inference algorithm. (b) WGCNA is the method showing the highest stability on the two classes. (c) Same pathway reconstructed same in term of local distance, instead a wider range of variability in -omics

Moving from gene profiling towards pathway profiling can be an effective solution to overcome the problem of the poor overlapping in -omics signatures. Nonetheless, the path from translating a discriminant gene panel into a coherent set of functionally related gene sets includes a number of steps each contributing to injecting variability in the process. To reduce the overall impact of such variability, it is thus critical that, whenever possible, the correct tool for each single step is adopted, accurately focussing on the desired target to be investigated. This mainly holds for the choice of the most suitable enrichment tool and biological knowledge database, and, to a lower extent, to the inference method for the network reconstruction: all these ingredients are planned for different objectives, and their use on other situations may result misleading. As a final observation and a possible future development to explore, the emerging instability can be tackled by obtaining the functional groups identification as the result of a prior knowledge injection in the learning phase, rather than a procedure a posteriori [Zyczinski et al., 2011, 2012].

4 CONCLUSION

ACKNOWLEDGEMENT

The authors at FBK want to thank Shamar Droggetti for his help with the enrichment web interfaces.

Funding: The authors at DISI acknowledge funding by the Compagnia di San Paolo funded Project Modelli e metodi computazionali nello studio della fisiologia e patologia di reti molecolari di controllo in ambito oncologico. The authors at FBK acknowledge funding by the European Union FP7 Project HiperDART and by the PAT funded Project ENVIROCHANGE.

REFERENCES

Aliébs, A., Cañada, A., and Díaz-Uriarte, R. (2008). PaLS: filtering common literature, biological terms and pathway information. Nucleic Acids Res., 36(Web Server issue), W364–W367.

Ambroise, C. and McLachlan, G. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. PNAS, 99(10), 6562–6566.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., K., D., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology, the gene ontology consortium. Nature Genetics, 25(1), 25–9.

Barabasi, A. L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nature Review Genetics, 12, 56–68.

Baralla, A., Mentzen, W., and de la Fuente, A. (2009). Inferring Gene Networks: Dream or Nightmare? Ann. N.Y. Acad. Sci., 1158, 246–256.

Barla, A., Mosci, S., Rosasco, L., and Verri, A. (2008). A method for robust variable selection with significance assessment. Proceedings of ESANN 2008.

Barla, A., Jurman, G., Visintainer, R., Filosi, M., Riccadonna, S., and Furlanello, C. (2011). A machine learning pipeline for discriminant pathways identification. In CIBB 2011, pages 1–10. ISBN-9788890643705.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: Structure and dynamics. Physics Reports, 424(4–5), 175–308.

Buchanan, M., Caldarelli, G., De Los Rios, P., Rao, F., and Vendruscolo, M., editors (2010). Networks in Cell Biology. Cambridge University Press.

Cover, T. and Thomas, J. (1991). Elements of Information Theory. Wiley.

De Mol, C., Mosci, S., Traskine, M., and Verri, A. (2009). A regularized method for selecting nested groups of relevant genes from microarray data. Journal of Computational Biology, page 8.
Fig. 5. Plots of Hamming vs. Ipsen distances (H vs. $\epsilon$) for all possible combinations of $M$, $D$, $E$ and $N$. In our analysis we considered the glocal distance $\phi$, defined as the normalized product of $H$ and $\epsilon$.

De Smet, R. and Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nature Review Microbiology*, 8, 717–729.

Faith, J., Hayete, B., Thaden, J., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J., and Gardner, T. (2007). Large-Scale Mapping and Validation of *Escherichia coli* Transcriptional Regulation from a Compendium of Expression Profiles. *PLoS Biol.*, 5(1), e8.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9, 1871–1874.

Furlanello, C., Serafini, M., Merler, S., and Jurman, G. (2003). Entropy-based gene ranking without selection bias for the predictive classification of microarray data. *BMC Bioinformatics*.

Grimaldi, M., Visintainer, R., and Jurman, G. (2011). Regnann: Reverse engineering gene networks using artificial neural networks. *PLoS ONE*, 6(12), e28646.

He, F., Balling, R., and Zeng, A.-P. (2009). Reverse engineering and verification of gene networks: Principles, assumptions, and limitations of present methods and future perspectives. *J. Biotechnol.*, 144(3), 190–203.

Horvath, S. (2011). Weighted Network Analysis: Applications in Genomics and Systems Biology. Springer.

Huang, D., Sherman, B., and Lempicki, R. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, 37(1), 1–13.

Ipsen, M. and Mikhailov, A. (2002). Evolutionary reconstruction of networks. *Phys. Rev. E*, 66(4), 046109.

Jurman, G., Visintainer, R., and Furlanello, C. (2011). An introduction to spectral distances in networks. In *Proc. WIRN 2010*, pages 227–234.

Jurman, G., Visintainer, R., Riccadonna, S., Filosi, M., and Furlanello, C. (2012). A glocal distance for network comparison. arXiv:submit/0397475 [math.CO].

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1), 27–30.

Leek, J., Scharpf, R., Corrada Bravo, H., Simcha, D., Langmead, B., Johnson, W., Geman, D., Baggerly, K., and Inzarry, R. (2010). Tackling the widespread and
critical impact of batch effects in high-throughput data. Nat Rev Genet, 11, 733–739. Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla-Favera, R., and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinform., 7(7), S7. Meyer, P., Lafitte, F., and Bontempi, G. (2008). minet: A R/Bioconductor Package for Inferning Large Transcriptional Networks Using Mutual Information. BMC Bioinform., 9(1), 461. Nemenman, I., Escola, G., Hlavacek, W., Unkefer, P., Unkefer, C., and Wall, M. (2007). Reconstruction of Metabolic Networks from High-Throughput Metabolite Profiling Data. Ann. N.Y. Acad. Sci., 1115, 102–115. Newman, M. (2003). The Structure and Function of Complex Networks. SIAM Review, 45, 167–256. Newman, M. (2010). Networks: An Introduction. Oxford University Press. Sharan, R. and Ideker, T. (2006). Modeling cellular machinery through biological network comparison. Nature Biotechnology, 24(4), 427–433. Strogatz, S. (2001). Exploring complex networks. Nature, 410, 268–276. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS, 102(43), 15545–15550. The MAQC Consortium (2010). The MAQC-II Project: A comprehensive study of common practices for the development and validation of microarray-based predictive models. Nature Biotechnology, 28(8), 827–838. Zhang, B., Kirov, S., and Snoddy, J. (2005a). WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res, 33(Web Server issue), W741–8. Zhang, Y., James, M., Middleton, F. A., and Davis, R. L. (2005b). Transcriptional analysis of multiple brain regions in parkinson’s disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms. Am J Med Genet B Neuropsychiatr Genet, 137B(1), 5–16. Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. J.R. Statist. Soc. B. Zycinski, G., Barla, A., and Verri, A. (2011). Svs: Data and knowledge integration in computational biology. Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, pages 6474 – 6478. Zycinski, G., Squillario, M., Barla, A., Sanavia, T., Verri, A., and Di Camillo, B. (2012). Discriminant functional gene groups identification with machine learning and prior knowledge. Submitted.
Fig. 6. (a) Pathway target cumulative histogram. (b) Hamming versus Ipsen (H vs. $\epsilon$) distance, and thresholding of high populated pathways.
