An integrated approach (CLuster Analysis Integration Method) to combine expression data and protein-protein interaction networks in agrigenomics: Application on *Arabidopsis thaliana*.

Daniele Santoni, Aleksandra Swiercz, Agnieszka Żmieńko, Marta Kasprzak, Marek Blazewicz, Paola Bertolazzi, and Giovanni Felici

Keywords: co-expression data, protein-protein interaction networks, microarray analysis, clustering, graph and cliques, KEGG Pathways

Daniele Santoni: Institute for Systems Analysis and Computer Science "Antonio Ruberti", National Research Council of Italy, Viale Manzoni 30, 00185, Rome, Italy, Tel.: +39067716423
Fax: +39067716461, Email: daniele.santoni@iasi.cnr.it.

Aleksandra Swiercz: Institute of Computing Science, Poznan University of Technology, ul. Piotrowo 2, 60965 Poznan, Poland, and Institute of Bioorganic Chemistry, Polish Academy of Sciences, ul. Noskowskiego 12/14, 61704 Poznan, Poland Tel: +48616653030, Fax: +4861 8771525 , Email: aleksandra.swiercz@cs.put.poznan.pl.

Agnieszka Żmieńko: Institute of Bioorganic Chemistry, Polish Academy of Sciences, ul. Noskowskiego 12/14, 61704 Poznan, Poland, Tel: +48616653052, Fax: +480618520532, Email: akisiel@ibch.poznan.pl.

Marta Kasprzak: Institute of Computing Science, Poznan University of Technology, ul. Piotrowo 2, 60965 Poznan, Poland, and Institute of Bioorganic Chemistry, Polish Academy of Sciences, ul. Noskowskiego 12/14, 61704 Poznan, Poland Tel: +48616653030, Fax: +4861 8771525, Email: mkasprzak@cs.put.poznan.pl.

Marek Blazewicz: Institute of Computing Science, Poznan University of Technology, ul. Piotrowo 2, 60965 Poznan, Poland and Poznan Supercomputing and Networking Center, ul. Noskowskiego 10, 61704 Poznan, Poland, Tel: +48 61 8582001, Fax: +48 61 8525954, Email: marek.blazewicz@cs.put.poznan.pl

Paola Bertolazzi: Institute for Systems Analysis and Computer Science "Antonio Ruberti", National Research Council of Italy, Viale Manzoni 30, 00185, Rome, Italy, Tel.: +39067716444
Fax: +39067716461, Email: paola.bertolazzi@iasi.cnr.it

Giovanni Felici: Institute for Systems Analysis and Computer Science "Antonio Ruberti", National Research Council of Italy, Viale Manzoni 30, 00185, Rome, Italy, Tel.: +39067716443
Fax: +39067716461, Email: giovanni.felici@iasi.cnr.it.

Corresponding author: Giovanni Felici, Email: giovanni.felici@iasi.cnr.it

ABSTRACT

Experimental co-expression data and protein-protein interaction networks are frequently used to analyze the interactions among genes or proteins. Recent studies have investigated methods to integrate these two sources of information. We propose a new method to integrate co-expression data obtained through DNA microarray analysis (MA) and protein-protein interaction (PPI) network data, and apply it to *Arabidopsis thaliana***.** The proposed method identifies small subsets of highly interacting proteins**.** Based on the analysis of the basis of co-localization and mRNA developmental expression, we show that these groups provide important biological insights; additionally, these subsets are significantly enriched with respect to KEGG Pathways and can be used to successfully predict whether proteins belong to known pathways. Thus, the method is able to provide relevant biological information and support the functional identification of complex genetic traits of economic value in plant agrigenomics research. The method has been implemented in a prototype software tool named CLAIM (**CL**uster **A**nalysis **I**ntegration **M**ethod) and can be downloaded from http://bio.cs.put.poznan.pl/research_fields. CLAIM is based on the separate clustering of MA and PPI data; the clusters are merged in a special graph; cliques of this graph are subsets of strongly connected proteins. The proposed method was successfully compared with existing methods. CLAIM appears to be a useful semi-automated tool for protein functional analysis and warrants further evaluation in agrigenomics research.

INTRODUCTION

The analysis of gene expression profiles derived from DNA microarray analysis (MA) or RNA-seq experiments is a popular approach to the search for genes involved in specified pathways, developmental stages or resistance to stress and pathogens. Moreover, the correlations between the expression profiles of individual genes are commonly used to infer functional relationships among them, under the assumption that two transcripts sharing a common expression profile are involved in common tasks (among others: Heyer et al., 1999; Stuart et al., 2003; Arisi et al., 2011). Similarity of the expression profiles of two genes in the sample space is usually calculated with correlation measures, such as the Pearson correlation coefficient (McShane et al., 2002; Eisen etal., 1998). The correlation matrix can then be analyzed with clustering methods, or it can be used to generate a co-expression network where nodes represent genes and are connected with arcs associated with the value of their correlation. On such a network additional graph-based analysis can be conducted, such as the search for strongly connected components, cliques, or special nodes (e.g., hubs of the network).

The information derived from a set of gene expression experiments can be extremely valuable, depending on the levels of interest and quality of the experiments. However, the post-genomic era is marked by the rapid accumulation of other genomic resources as well: the whole genome sequences of many organisms, proteomic, metabolomics, epigenomic and interactomic data, to name but a few. The need for data integration is a by-product of this increasing availability of large experimental data that are obtained and globally shared at a decreasing cost. Combining data from different steps of the expression of genomic information may be helpful in uncovering gene functions and regulatory mechanisms in the biological systems. Today's agricultural genomics (agrigenomics) research focuses on the use of genomic technologies for crop and breedstock improvement, increasing resistance towards disease and infection, optimizing plant yield, as well as aiding their use in biofuels or pharmacy. However, priority agronomic traits are often genetically complex and interact with the environment. It is therefore highly desirable to combine various types of phenotypic analyses in order to extract additional information from the data and to increase precision in the marker discovery and association studies. Integration methods also allow for supporting analysis of a species of interest with functional data obtained for model organisms (De Bodt et al., 2009).

The two data sources that so far have been most commonly integrated for meta-analysis are the co-expression (mainly MA) data and protein-protein interaction (PPI) network data. A variety of methods combining these two datasets have been proposed. One of the main references in this field is the work described in Ulitsky and Shamir (2007), proposing an approach based on the search for clusters with high similarity in the MA data which are, at the same time, connected in the PPI network. The method, named MATISSE, has been widely used ever since, and several interesting biological results have been derived from it. The same authors also proposed a variation of their method (CEZANNE (Ulitsky and Shamir, 2009)) where, instead of a condition of simple connectivity in the PPI, a confidence score is used to weight the information contained in the PPI. Differently, in Tornow and Mewes (2003), the proposed method searches for modules in the protein interaction network with the superparamagnetic approach, and next evaluates the correlation strength in the gene expression. Based on the distribution of the correlation strength, the authors compute the probability that the observed strength is derived from random coincidence. In Pavlidis et al., (2002), the authors considered the problem of the classification of genes into functional groups according to gene expression and phylogenetic profiles with the SVM method. They investigated three different fashions of integration of the two data sources: before the classifying method, in between, and after the classification. In Shiga et al., (2007), a method was proposed where the integration between the co-expression and PPI is based on a probabilistic clustering model. In Peña-Castillo et al., (2008), the focus was directed towards the discovery of gene functions; a large body of mouse and human functional genomic data was assembled and analyzed by several teams of scientists using machine learning methods, showing that predictions of good quality can be obtained with a combination of different classifiers for a large portion of Gene Ontology (GO) terms. Among many interesting contributions, the authors confirmed that significant improvements can be obtained when co-expression data and PPI networks are integrated. In De Bodt et al., (2009), the adopted method is based on the identification of clusters and on the analysis of their conservation in the different species. The co-expression information was thus reinforced with co-localization and the similarity of its biological role. Wu et al. (Wu, 2012) also showed that the integration of gene expression and PPI networks provides important information for gene prioritization; in their study they proposed a computational method for the integration and successfully applied it to several breast cancer and lung cancer datasets.

A growing number of reports highlight the importance of biological data integration in a broader context and address this problem with various computational strategies. In Rogers et al., (2008), the authors combined transcriptomic and proteomic data to estimate the relationships between the mRNA and protein accumulation profiles in human breast cell lines. The adopted method is based on probabilistic clustering. An algorithm based on simultaneous clustering of multiple networks was proposed in Narayanan et al., (2010), to identify clusters according to different types of information. Li et al. (Li, 2012) constructed a dynamic PPI network, based on PPI data and a series of time-sequenced gene expression data; Wong et al. (Wong, 2012) proposed a very powerful method to extract information on protein functions from a large cross-organism compendium of functional predictions and networks.

The variety of emerging tools and lack of consensus over different approaches in this developing field confirm the importance of improving the methods that integrate different information sources. Nevertheless, an interesting debate is ongoing, related to the reliability that can be granted to biological conclusions derived from large-scale association studies, such as co-expression networks and PPI. A representative sample of this debate can be found in the work of Gillis and Pavlidis (2011, 2012). Here important considerations on the meaning of connections in gene networks are provided. In these two papers the authors challenged the so-called *guilty-by-association* principle (GBA) from different points of view. First, they claimed that the dominant information in gene interaction networks is essentially contained in direct connections between pairs of genes, providing experimental evidence based on both MA co-expression and PPI networks (Gillis and Pavlidis, 2011); besides, they showed that the aggregation of different co-expression networks obtained in different experiments may improve gene function predictions, and that the integration of the co-expression and PPI network improves predicting power, the latter being in agreement with Peña-Castillo et al., (2008). On the other hand, it has been shown how a very little subset of edges that satisfy certain properties (*critical* and *exceptional* edges) can dramatically affect the whole structure of the network and the conclusions that can be drawn from it based on the GBA principle (Gillis and Pavlidis, 2012). Also this paper seems to suggest the use of the integration of different data sources finalized to the identification of a small and compact set of interactions that may represent the essential information in a very large network.

The approach proposed in this paper adopts a particular technique to integrate the co-expression and PPI data. Its originality is to be found in the fact that the two networks are equally important in the process: both networks are repeatedly clustered according to their own metrics, and then the results of these clusterings are combined into a new graph, whose fully connected components (cliques) are analyzed. We named it CLAIM (Cluster Analysis Integration Method). It provides the cliques together with a measure of their strength and of their robustness with respect to the “noise” that may have been generated in the experiments and in the clustering process.

The main rationale of this method is that really similar proteins will fall into the same cluster regardless of the clustering parameters adopted and of the network used. The robustness measure, later referred to as “level”, represents exactly the convergence of the results w.r.t. to the sources of variation represented by the experiments, the samples, and the algorithm used.

CLAIM is designed to discover subsets of proteins that appear to be co-regulated by important processes associated with the studied organisms and with the experimental conditions; for this reason, we test the results of our method in their relation to known pathways, under the assumption that our cliques are interesting if they are enriched with respect to particular pathways. An additional step consists in using the information contained in the cliques as a method to associate new proteins with a known pathway. Such a potential role of our cliques is tested with a blind leave-one-out test. A final test of validity for our approach is based on the comparison with other integration methods. The tests have been run on *Arabidopsis thaliana* (hereafter, Arabidopsis) datasets, a model organism widely studied, with a genome of 135 Mbp and approximately 25,000 genes. Biological analysis of CLAIM-derived cliques reveals the co-localization and convergence of the mRNA developmental expression profiles of clique elements, thereby providing interesting and sound interpretations of our results.

MATERIALS AND METHODS

Co-expression (MA) data

Affymetrix chip data for time-course transcriptome analysis of Arabidopsis roots, exposed to salinity stress (Dinneny et al.,, 2008), were used in this study. Normalized data were retrieved from the GEO database (GDS3216). Relative gene expression at each treatment time point (0.5h, 1h, 4h, 16h and 32h, versus 0h) was assessed with the R/limma package (Smyth, 2005). A set of 3,943 differentially expressed genes was selected for further analysis using an *F-p value* threshold set to 0.05. The Pearson correlation coefficient ρ_{ij} was computed for each couple of genes *i* and *j*. We assumed that genes showing a large and positive Pearson coefficient (close to 1) are considered strongly related, while 0 indicates the absence of a correlation, and large negative coefficients (close to -1) indicate a strong negative correlation. The Pearson correlation coefficient was converted into a distance that reached its minimum when genes had the same profiles, and increased as the difference in the profiles increased. We used the measure (1-Pearson correlation coefficient) referred to as the *Pearson Correlation Measure* (PCM) (McShane et al., 2002) of two genes:

$$\mathrm{PCM}(i,j)=\mathrm{PCM}(j,i)=1-\rho_{ij}, \text{ where } \mathrm{PCM}(i,j):\mathrm{N} \times \mathrm{N} \rightarrow [0,2].$$

Protein-Protein Interaction (PPI) data

Available information on PPI (Arabidopsis Interactome Mapping Consortium, 2011) containing binary interactions between 4,866 proteins was used. Paths of infinite lengths were removed by considering only the first connected component, whose resulting diameter was equal to 13. On such a sub-network the shortest paths matrix among all pairs of nodes was computed by the implementation of the Floyd-Warshall algorithm (Lawler, 2001). We defined SPM(*i,j*) as the *Shortest Path Measure* of the distance between proteins *i* and *j* in the network, made equal to the length of any of the shortest paths between the two proteins. We assumed the network was undirected and thus we obtained

$$\mathrm{SPM}(i,j) = \mathrm{SPM}(j,i) = \text{shortest path between } i \text{ and } j, \text{ where } \mathrm{SPM}(i,j)\text{: } \mathrm{N} \times \mathrm{N} \rightarrow [0,13].$$

The intersection of the MA and PPI sets resulted in 694 elements. Given that genes and proteins are in a one-to-one relation in the working set, in the following we refer to these elements simply as proteins.

Clusterings

A *k*-means algorithm (Hartigan, 1975; MacKay, 2003) was used to cluster genes with respect to various distances. We use as features for each gene its distance/similarity measure from all the other genes. For CLAIM, PCM and SPM distances were used in separate clusterings and the algorithm was run for different values of *k* ranging from 5 to 30, in increments of 5. In the clusterings based on PCM, we indicated with M_5 the array representing the results for *k* = 5, with M_{10} those for *k* = 10, and so on, up to M_{30}. Analogously in the clusterings based on SPM, the respective arrays were indicated with $S_{5,}$..., S_{30}. For each *k*-means run, 50 different random initial seeds were considered; then, the best clusterization according to a standard clustering quality measure was selected. For weighted clustering five different new measures were obtained as linear combinations of normalized SPM and PCM with different weights, as reported below:

$$W(\delta) = \delta\,(\mathrm{SPM}) + (1 - \delta)\,(\mathrm{PCM})$$

for δ = 0, 0.25, 0.5, 0.75, 1. Trivially for the extreme values of δ (0 and 1) we obtained exactly PCM and SPM respectively, and for δ equal to 0.5 we obtained a fair combination of the two measures (we referred to this measure as the CO-CLUSTER (Hanisch et al., 2002)).

Cliques

Given a graph $G = (V,E)$, a clique C in G is a subset of its nodes such that each pair of nodes in C is connected by an edge in E. When analyzing a graph G, one is typically interested in finding the largest or maximal cliques of G, or to know how many cliques of a given size are present. Maximal cliques were identified using the *igraph* R-Library software, which provides a robust implementation of the maximal clique algorithm (Csardi and Nepusz, 2006).

Synthesis of Clusterings: the H-graphs and the H-cliques

From M_i and S_j we built graph $G_{i,j} = (V, E_{ij})$ where the node set V contains all the 694 proteins, and an edge exists between two nodes if they fall into the same cluster in both M_i and S_j. We restricted the construction of $G_{i,j}$ only to those cases where the number of clusters in M_i and S_j was sufficiently large, by choosing only those pairs (i,j) for which $(i \times j) \geq 51$ (excluding couples: (5, 5), (5, 10) and (10, 5)). The choice of this threshold was motivated by the fact that the higher the number of clusters, the stronger the relationship is between two proteins that belong to the same cluster in both M_i and S_j. When i and j are both small, the co-presence in the same cluster would have been much less significant than in the case when i and j are large. Imposing this restriction, we obtained 33 different graphs $G_{i,j}$. Given $p = 1, 2, \ldots, 33$, we defined a new graph H_p whose nodes were the same as those of any $G_{i,j}$, and an edge between two nodes (i.e., proteins) is present only if that edge is present in at least p of the 33 $G_{i,j}$ graphs.

For each p, we then identified the maximal cliques in H_p and we referred to these cliques as H-Cliques. We defined the *level* of a H-Clique as follows: it has level 1 if it is found in all 33 $G_{i,j}$ graphs; it has level 2 if it is found in 32 of them - and so on - down to level 33, which can be associated with the least relevant cliques. Cliques extracted from the H_p (p = 1, ..., 33) graphs can be ordered according to their level and their size (the number of nodes that compose them). From the definition, it clearly follows that a H-Clique of level 1 is also a H-Clique of level 2, 3, ..., 33; in other words, proteins that belong to a clique of level 1 are proteins that are clustered together according to both PCM and SPM in all the considered clusterings, regardless of the choice of the number of clusters used; accordingly, proteins that belong to a clique of level 2 are proteins that are clustered together according to both PCM and SPM in all the clusterings considered but one, and so on. The computation time of the algorithm is bound by 2 hours on a PC desktop i5 dual core CPU with 8GB RAM running Linux Ubuntu 11.10. A scheme of the method is represented in Figure 1.

MATISSE analysis

MATISSE (Ulitsky and Shamir, 2007) version 1.0 was downloaded and used as an alternative method for identifying groups of genes with similar functionality. From empirical tests we have observed that Matisse favours larger clusters with respect to CLAIM. Each new gene added to the cluster caused a growth of the cluster's score function proportional to the sum of all positive correlations. There was no significant penalty observed for genes which are not strongly related to each other. As a result of the analysis, we have obtained a set of large clusters with loosely coupled genes and the main constraint of their sizes was a predefined value defining the desired maximal size of a cluster. Such clusters resulted in very poor performance in our test of assigning genes to pathways. That is why we decided to force smaller clusters by requesting sets of genes with a limited maximal size. We have performed Matisse analysis 32 times, for a maximal size of clusters varying between 3 and 34 (other changeable parameters remained at default). Moreover, the Matisse score of each cluster was explicitly normalized by the number of maximal possible count of edges in that set of genes. Both interventions significantly improved the classifier's performance. Such clusters were similar to those obtained by CLAIM and gave similarly good results in our blind test.

Performing enrichment analysis

The H-cliques were analyzed in terms of their pathway enrichment. Pathway maps related to Arabidopsis were downloaded from KEGG (Kanehisa et al., 2006). For each clique we computed, using R scripts, the p-value of the hypergeometric test (Lee, 2010) for each pathway associated to at least one protein of that clique. The smallest p-value was taken into account and associated to each clique. A clique was considered to be enriched when it was enriched at least in one pathway, i.e. for at least one pathway the hypergeometric test provided a p-value smaller than 0.05. The fraction of enriched cliques was computed out of the total number of cliques. Such an enrichment test is designed to discount the bias due to randomness and we can assume, provided that a sufficiently small p-value as 0.05 is satisfied, that the presence of one or more interesting proteins in one of the subsets found by CLAIM cannot be ascribed to chance but to the ability of the method.

Pathways Prediction

KEGG pathway maps for Arabidopsis were used (Kanehisa et al., 2006). We analyzed all the selected cliques in terms of their intersection with the pathways, under the hypothesis that a protein occurring in a clique that intersects a certain pathway is likely to be involved in that pathway. Under this assumption, we defined a score for each pair protein-pathway; the larger the score, the stronger the probability that the given protein belongs to that pathway. Given a protein *p* and a pathway *PT*, we indicate N_p cliques containing *p* with C_i, *i*=1,…,N_p, and then define:

$$S(p,PT) = \sum_{i=1}^{N_p} 1/l_i \frac{|C_i \cap PT| - \delta(p,PT)}{|C_i|}$$

where sum *S*(*p,PT*) is over all cliques containing the protein *p*, l_i is the level of clique *i,* and $\delta(p,PT)$ is equal to 1 if protein *p* is in pathway *PT*, and 0 otherwise. The presence of $\delta(p,PT)$ guarantees that the score is computed without the influence of the protein *p*, which conversely would introduce a bias into the prediction when *p* belongs to that pathway. Prediction is accomplished by assigning a protein to the pathway for which the score is above a threshold.

RESULTS

Identification of sets of related genes (H-Cliques) with CLAIM

CLAIM integrates co-expression data and PPI data to identify associations among proteins. As a case study, we used MA experiments focused on evaluating time-course transcriptome changes in Arabidopsis roots under salinity stress (Dinneny et al., 2008). The choice of this dataset is directly motivated by our current research interests. Also, the early response of Arabidopsis to salt stress (up to 32h, see Materials and Methods) manifests in waves of gene expression changes, with only 1.4% of genes being differentially expressed across the whole analyzed time period. It therefore provides a good test set for the method based on data clustering. For PPI information we turned to the Plant Interactome Database (Arabidopsis Interactome Mapping Consortium, 2011). Currently this is the most comprehensive interactome data source for this organism. It contains about 10,900 binary interactions of over 4,800 unique proteins, all of them being experimentally identified. The analysis was restricted to the largest connected component and further to a set of 694 genes shared between PPI and MA (Figure 2). Given that genes and proteins are in a one-to-one relation in the working set, in the following we refer to these elements simply as proteins.

The proteins in the working set have been clustered separately according to two different distance measures. The first measure was based on the Pearson correlation between expression profiles (PCM) and the second was based on the shortest path distance between two proteins in the PPI network (SPM). A weighted graph H has then been constructed, where two proteins were connected if they belonged to the same cluster according to both measures. The clustering algorithms runs have been repeated for different values of the parameter that controls the number of clusters.
A simple example reported in Figure 3 clarifies this procedure. The maximal cliques of this graph (H-Cliques) were used to orient biological analysis and to predict the pathways to which a protein belongs (see below).
We wish to highlight that H-cliques represent groups of proteins whose interactome topology is suggested by both the PCM and SPM. The cliques are ordered according to two parameters: the *level* — that inversely depends on the weight of edges that define the clique — and the *size*, i.e. the number of nodes that compose it. The level of

interest of a clique decreases as its level increases. For the analysis we consider only *maximal* cliques, e.g., cliques that are not contained in any other clique of a larger size.
As mentioned above, the level of a clique is a measure of its validity and robustness w.r.t. not interesting sources of variations, such as those coming from the type of distance used, the experiments, the clustering algorithm, and its parameters. This is indeed confirmed by control experiments that have been run using randomly shuffled distance matrices, where the application of CLAIM resulted in many small to mid-size cliques with a very poor level and no clique with a good level. To reinforce this finding we also verify that the quality of the clustering significantly worsens on the randomly shuffled data, regardless of the number of iterations, the number of clusters, or the starting point.

Enrichment of H-Cliques

The results of CLAIM for the *Arabidopsis* dataset have been evaluated from different points of view. As the first step, in order to test the significance of the identified cliques, we evaluated the coherence of the cliques composition with respect to the KEGG pathways by performing enrichment analysis. We restricted the analysis to the cliques containing at least one annotated protein (n=2,372, out of a total number of 4,563). 1,679 of those cliques (~71%) turned out to be enriched ($p<0.05$) according to our definition (see Materials and Methods). By limiting the analysis to the cliques with a level smaller than 30 (n=564) we found 454 enriched cliques, reaching an enrichment percentage of 80%. Additionally, we performed the same analysis with the cliques obtained using a weighted combination of the two measures (weighted-clustering) as reported in the Materials and Methods section. Weighted-clustering was used as an alternative and simple method to extract clusters from the two measures, in order to test the ability of our algorithm to integrate the two sources of data. In weighted-clustering, new distance functions were obtained as convex combinations of the normalized values of the PCM and SCM. These new distances were then used to extract alternative clusters or sets using the same algorithm (k-means) as for CLAIM. The new clusters were evaluated in their pathways' enrichment and compared with those obtained with CLAIM. The percentage of enriched cliques was 48% and 44% for the separate clustering of MA and PPI respectively. For CO-CLUSTER (0.5 MA + 0.5 PPI) a percentage of 52% was reached and for the two cases 0.25 MA + 0.75 PPI and 0.75 MA + 0.25 PPI a percentage of 42% and 51% was obtained, respectively. The enrichment analysis was finally computed for clusters obtained with MATISSE. This method searches for clusters with a high similarity in the MA and interconnected in the PPI. The choice of MATISSE was motivated by its popularity (162 citations before 2013, according to GoogleScholar) and the fact that it has become a usual reference for the integration of MA and PPI data. For MATISSE-derived clusters, the percentage of pathway enrichment was ~68% (n= 250 enriched clusters out of 366). The results of the enrichment analysis related to all the test sets are reported in Figure 4.

Pathway prediction power of H-Cliques

The high correspondence of cliques with the known pathways revealed by GO enrichment analysis provides evidence for the biological significance of gene sets computed with our method. Besides confirming the validity of our analysis, it provides a rationale to assign proteins whose functional roles are unknown to some of the known pathways. We have blindly tested the assignment correction against the proteins whose functions have been already described. For this we assumed that we do not know the functional class of a protein and we used the CLAIM rule to assign it to a pathway. We then verified whether this assignment was correct or not. For each threshold of the score function (see Pathways Prediction paragraph in the Materials and Methods section), we computed the percentage of correct predictions, and divided into true positive (a prediction is a true positive if the protein is assigned to the pathway it belongs to), and true negative results (i.e., a protein that does not belong to any pathway is not assigned to any pathway). Such experiments are naturally summarized by the Receiver Operating Characteristics Curves (ROC curves; Fawcett, 2006) where the true positive rate is plotted against the false positive rate depending on different values of a parameter of the classifier. In this case, the parameter adopted was the threshold on the score function: when the score of a protein-pathway pair was above the threshold, the protein was assigned to the pathway. A synthetic way to establish the value of a classifier from its ROC curve is the Area Under the ROC Curve (AUC). The AUC would be equal to 1 if there exists a perfect classifier for at least one value of the threshold, while it would equal one half in the case of random classifiers. In the ROC curve plotted in Figure 5 (right Panel), we report the true and false positive rates over the total number of protein-pathway pairs that can be obtained by combining all the proteins present in the H-cliques of CLAIM with all the known pathways of Arabidopsis, obtained from the KEGG repository (Kanehisa et al., 2006).

We point out that the apparent concavity in the curve of between approx. 0.18 and 0.4 FPR is due to only two values of FPR and that the number of elements that are poorly recognized in that portion is very limited. To any extent, this part of the curve is not practically relevant as one would not like at all to operate with an FPR larger than 0.15.

In order to test the pathway predictive power of our method against this acknowledged standard, we used the clusters determined by MATISSE for pathway predictions in a similar manner as we used the cliques derived from CLAIM. For each of the MATISSE clusters we have used the weights provided by the algorithm normalized over the dimension of the cluster (used as the reciprocal of the level in the formula provided in the Pathways Prediction paragraph, in the Materials and Methods section). We have considered only the clusters of a size comparable to those of CLAIM (3 to 30 nodes). The reason for limiting the cluster sizes and normalizing the score function was the fact that MATISSE*'s* score favours larger sub-networks. Despite the fact that such an approach might be helpful in finding large connected sub-graphs, it does not turn out to be effective in finding strong connections between proteins; besides, it would skew the analysis. The ROC curve for the MATISSE-based pathway prediction power is presented in Figure 5 (left Panel). Although it must be stressed that the results of the two methods could not be compared straightforwardly (e.g., the scales of the scores and the coefficients in the score formula differed for the two methods), it is interesting to point out that the ROC plots show a pathways prediction power for CLAIM surely comparable – if not superior – to that of MATISSE; the related AUC, in fact, results in 0.8837 for MATISSE and 0.9200 for CLAIM.

Biological analysis of H-Cliques

The information conveyed by the maximal cliques obtained by CLAIM can be used for different types of analysis. Cliques of good quality (e.g., with a small value of their level) can be analyzed to see if they disclose some biological information of interest within the scope of the analysis. Here we have focused on maximal cliques whose size *s* is greater than 2, that count up to 4,563 (see Supplemental Material, Dataset S1– sheet

M1). As a starting step, we analyzed some of the smallest and most significant cliques: 12 cliques of size 3 with a level equal to 1 and 6, represented in Table 1. Their low level (≤ 6) guarantees that the proteins belonging to each clique are strictly linked to each other and in general fall into the same cluster both in the MA and PPI clusterings. In order to check whether the combination of the proteins that fall into a common clique reflects their biological relationship, we analyzed their Gene Ontology (GO) annotations as well as their expression profiles (on the mRNA level). We compared the proteins' cellular localization, molecular function and the biological processes in which they are involved, as listed in the curated TAIR GO annotations (Berardini et al., 2004) (Supplemental Material, Dataset S1 – sheet M2). As expected, for proteins with a known GO category classification we observed a lot of congruence in their cellular localization (cliques: 02 – endomembrane system, 05 - nucleus, 09 – membrane system, 10 – (plastid-)membrane and 12 - nucleus) and biological processes (cliques: 02 - various defence responses, 03 – oxidation-reduction process, 05 – transcription-related and defence-related, 06 – defence related, 08 – stress response, 11 – stress response and 12 – regulation of transcription). This obvious correspondence implies that it may be possible to infer the biological annotation of the uncharacterized proteins from the GO functions of their clique partners. This assumption was further confirmed by analysis of the mRNA developmental expression patterns for each clique. The expression profiles reflect the average mRNA levels of each protein at various life cycle stages of the Arabidopsis plants. They were calculated with the Genevestigator/Condition Search/Development tool by analysis of over 6,500 Affymetrix gene chip samples from various gene expression experiments (Hruz et al., 2008) (Supplemental Material, Dataset S1, sheet M4). The correlations between some or even all of the clique members are often striking, with the clique 04 being the most prominent example. The observed similarities can be seen even more clearly by focusing on a particular sample set (in this case on the root expression data only – about 600 samples in the Genevestigator database – as the MA data used for cliques generation were derived from roots as well (see Supplemental Material, Dataset S1, sheet M5)).

Biological analysis of predicted protein-pathway associations

Considering the high degree of functional similarities observed among proteins assigned to a common clique, and the ability of CLAIM to accurately predict protein function using GO enrichment analysis information, we turned to using it to predict the functional role of proteins for which no pathway was assigned within the KEGG repository. We analyzed the computed protein-pathway couples with a score greater than 0.2 (for simplicity, in our analysis we did not consider couples associating a protein to the general pathway ath01100 - metabolic pathways). A complete list of 160 computed protein-pathway couples was analyzed, and GO classes were retrieved and made available (Supplemental Material, Dataset S1 – sheet M3). The first, second, and fourth highest scores were obtained for protein AT2G47400, coupled with the pathways ath01064 (*biosynthesis of alkaloids derived from ornithine, lysine and nicotinic acid*), ath00195 (*photosynthesis*) and ath00960 (*tropane, piperidine and pyridine alkaloid biosynthesis*), respectively. According to TAIR, AT2G47400 is a small peptide which belongs to the CP12 gene family, thought to be involved in the formation of a supramolecular complex with glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and phosphoribulokinase (PRK) embedded in the Calvin cycle. Its targeting to chloroplast is well documented and it was reported to be involved in the negative regulation of the reductive pentose-phosphate cycle (Zybailov et al., 2008; Marri et al., 2005; Marri et al., 2010; Ferro et al., 2010), which is in agreement with our predictions. The third highest score was obtained for the couple AT3G21200 – ath00195 – *photosynthesis*. Protein AT3G21200 (PGR7) is coded by a nuclear gene conserved in plants, algae and bacteria, yet it has no molecular function assigned. According to recent experimental data, it is targeted to chloroplasts. Moreover, the pgr7 gene mutation results in lowering of the efficiency of photosynthetic electron transport (Ferro et al., 2010), which confirms the correctness of our pathway prediction. Assigning a protein to more than one pathway was quite common and fully expected, as it is a consequence of the pathway hierarchy. Yet, this multiple assignment can provide more confidence for the findings when the pathways are functionally related. For example, protein AT2G24550, for which any functional category has not been specified until now, has been assigned to three pathways, all related to nucleic acid biosynthesis (highlighted in yellow in the Supplemental Material, Dataset S1 - sheet M3). It can be noticed that the assignments to pathway ath00195 – *photosynthesis*, were most common (involving 40 various proteins). The MA dataset used in this case study consists of genes differentially expressed in the Arabidopsis roots in response to salt stress and is highly enriched in chloroplast-related proteins (Dinneny et al., 2008). Proteins targeted to these organelles are usually assigned to this – very general – pathway, apart from the more specific functional categorization.

DISCUSSION

As anticipated, the integration of different sources of information on genes and their products has recently received increasing attention in literature. The main data integration approaches and studies differ in the methods they adopt and in the type of answers that are sought. In this work we propose CLAIM, a novel approach to the couple co-expression and PPI network data that identifies groups of functionally related genes/proteins. It has been designed to capture common features of the two sources of information by the clique searching approach. The presence of a clique in a special graph *H*, constructed using the results of many runs of a clustering algorithm on each data source, guarantees a very robust relationship among the genes/proteins belonging to that clique. The analysis was performed on the model plant Arabidopsis; out of the 694 genes, we selected a total of 4,563 maximal cliques of different size and strength. Larger and weaker cliques were more abundant. Nonetheless, the parameter *level*, associated with each clique, represents a further quality measure to select robust groups of nodes that exhibit a strong functional relationship. The identification of the cliques is a first step that may significantly help to restrict the work of the biologist in performing a semi-automated functional analysis and the related annotation process. We show that CLAIM exhibits a higher percentage of H-cliques significantly enriched in functionally-related genes (p-value of 0.05) with respect to other methods that derive clusters or subsets based on the same information. Moreover, using a score function derived from the H-cliques, we define a reliable predictor of the association between a protein and a pathway. The basic assumption in using CLAIM for pathway prediction is that if a protein belongs to a clique and other proteins in that clique are assigned to a common pathway, then that protein is likely to belong to that pathway, too. This assumption has been validated by blind analysis of the proteins with a known functional annotation, testing the ability of the method to assign them to the correct pathways.

Combining data from various levels of gene expression has shown to support the study of processes involving a network of interacting genes, like the plant floral transition (He et al., 2010). Also, the integration of data from various levels of genome expression may facilitate functional analysis of genes. The fast expanding agrigenomics field demands extensive phenotypic data to support the identification of complex genetic traits of economic value in plants or animals (Gedil and Rabbi, 2012). In this context, the pathway-prediction approach adopted in CLAIM may be of high importance to agrigenomics studies. Functional annotation of genes in cultivated plants is mostly based on sequence homology-based analyses to model species. At the same time, according to The Arabidopsis Information Resource (TAIR), 42% of Arabidopsis protein-coding genes are not assigned to any GO term within Molecular Function category or are classified as "Unknown molecular function". The same applies to Biological Process category. Interestingly, those still functionally uncharacterized genes display extremely high variation and are likely to contribute to plant adaptive evolution, as recent large-scale population sequencing studies revealed (Xu et al., 2011, Cao et al., 2011). Meta-analysis of data, utilizing integration methods like CLAIM may therefore help to reveal new promising, although less obvious, candidates. As a result, it will reinforce the progress of agrigenomics-based crop improvement.

Although the clique-based pathway prediction approach has been designed for CLAIM, it can actually be used to derive pathways predictions from any collection of gene sets that are supposed to be enriched with respect to pathways, and for which a score measure is available. We therefore adopted a similar scoring method to obtain pathway predictions from the subsets derived from MATISSE, the tool developed for identifying functional modules (Ulitsky and Shamir, 2007). In comparison with MATISSE, our method has appeared to achieve better results according to its pathway prediction power, as determined from the ROC curve analysis. It must be pointed out that the two methods have not been designed for the same purposes; nonetheless, we have tried to extract from them and to score analogous information. We also acknowledge that the same authors proposed an extension of the method where confidence scores are computed on the edges of the PPI network with the use of additional information collected from experiments (CEZANNE (Ulitsky and Shamir, 2009)). We have limited the actual comparisons with the 2007 version of the method because CLAIM does not use the additional information used by CEZANNE and the comparison could suffer from that bias. As for the other methods, we note that some of them (Narayanan at al., 2010; Jung et al., 2010) propose alternative methods for clustering proteins (i.e., finding a partition of them) resulting in much larger subsets of proteins than those found by CLAIM; besides, they do not exhibit a simple way to control the number of subsets that are found and their dimension. Similarly, the interesting work of De Bodt et al., (2009) is substantially devoted to improving the PPI of the studied organism, rather than finding special small structures based on the pairwise relations between proteins. The same considerations apply to the different types of weighted-clustering that we have tested. While the enrichment analysis makes perfect sense for comparing the methods, the prediction power may be strongly distorted by the different number of negative examples that are considered in the different methods. For this reason, it is not a straightforward operation to compare the contribution of CLAIM with other methods that, in different ways, try to exploit and integrate the co-expression and PPI data. Yet it seems that, due to the integration of the MA and PPI data *after* an independent clustering (instead of coupling the data at the beginning as MATISSE and other methods do), CLAIM is more effective in the prediction of protein functions than the other methods tested. Also, our algorithm extracts additional information from the PPI network, which is the distance between proteins; the use of this at the clustering stage had, in our opinion, a visible impact on the outcomes.

Additional validity of the results generated with CLAIM comes from the analysis of the biological meaning of the H-cliques. By simple comparison of the GO annotations of proteins assigned to a common clique (in the Biological Process, Molecular Pathway and Cellular Compartment categories), we show how interesting biological information surfaces from their analysis. Although the PCM clustering process was based on a relatively small and specific dataset (the expression profiles of genes in Arabidopsis roots subjected to salinity stress, measured at six time points), we were also able to demonstrate that genes classified as strongly related according to CLAIM (i.e. falling into the same clique of a low level), indeed often display surprisingly congruent expression patterns across the whole lifespan of a plant. Due to the satisfactory ability to correctly predict the pathway assignment of clique elements in the GO term-enriched cliques, CLAIM appears to be a useful semi-automated tool for protein functional analysis.

Authors' contributions

DS and GF conceived the project and contributed to the computational analysis and draft of the manuscript. AŻ pre-processed the microarray data and contributed to the biological interpretation of the results. AS and PB contributed to the development of the method and to the validation of the data. MB and MK contributed to the analysis of the results and to the comparison with other methods. DS and MB implemented the software. All the authors contributed to the final writing and approved the final manuscript.

Support

This work has been partially supported by the scientific cooperation program between the National Research Council of Italy (CNR) and the Polish Academy of Sciences (PAN). Funding covered travel expenses for meetings among authors and conference participations (www.cnr.it, www.english.pan.pl); NCN grant No. 2012/05/B/ST6/03026 supported the collection of experimental data and their analysis (www.ncn.gov.pl); the FLAGSHIP ”InterOmics” project (PB.P05) funded by the Italian MIUR and CNR institutions supported the work on method design and testing (www.cnr.it).

References

Arabidopsis Interactome Mapping Consortium. (2011). Evidence for network evolution in an Arabidopsis interactome map. Science 333(6042), 601−6077. Available: http://interactome.dfci.harvard.edu/A_thaliana/doc/AI_interactions.xls

Arisi I, D'Onofrio M, Brandi R, Felsani A, Capsoni S, Drovandi G, et al. (2011). Gene expression biomarkers in the brain of a mouse model for Alzheimer's disease: Mining of microarray data by logic classification and feature selection. Journal of Alzheimer's Disease 24(4), 721-738.

Berardini TZ, Mundodi S, Reiser R, Huala E, Garcia-Hernandez M, Zhang P, et al. (2004). Functional annotation of the Arabidopsis genome using controlled vocabularies. Plant Physiol 135(2), 1−11.

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet 43, 956-963.

Csardi G, and Nepusz T. (2006). The igraph software package for complex network research. InterJournal. Complex Systems 1695. Available: www.necsi.edu/events/iccs6/papers/c1602a3c126ba822d0bc4293371c.pdf Accessed: 11 January 2013.

De Bodt S, Proost S, Vandepoele K, Rouzé P, and Van de Peer Y. (2009). Predicting protein-protein interactions in Arabidopsis thaliana through integration of orthology, gene ontology and co-expression. BMC Genomics 10, 288.

Dinneny JR, Long TA, Wang JY, Jung JW, Mace D, Pointer S, et al. (2008). Cell identity mediates the response of Arabidopsis roots to abiotic stress. Science 320(5878), 942−5.

Eisen M, Spellman P, Brown P, and Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U.S.A. 95, 14863–14868.

Fawcett T. (2006). An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874.

Ferro M, Brugière S, Salvi D, Seigneurin-Berny D, Court M, Moyet L, et al. (2010). AT_CHLORO, a comprehensive chloroplast proteome database with subplastidial localization and curated information on envelope proteins. Mol Cell Proteomics 9(6), 1063−84.

Gedil M, and Rabbi I. (2012). Leveraging “agrigenomics” for crop improvement. IITA R4D Review 8, 55-57.

Gillis J, and Pavlidis P. (2011). The Role of indirect connections in gene networks in predicting functions. Bioinformatics 27(13), 1860-1866.

Gillis J, and Pavlidis P. (2012). ‘‘Guilt by association’’ is the exception rather than the rule in gene networks. PLOS Computational Biology 8-3-v1002444.

Hanisch D, Zien A, Zimmer R, and Lengauer T. (2002). Co-clustering of biological networks and gene expression data. Bioinformatics 18, 145-154.

Hartigan JA. (1975). Clustering algorithms. (Wiley. MR0405726).

He F, Zhou Y, Zhang Z (2010). Deciphering the Arabidopsis floral transition process by integrating a protein-protein interaction network and gene expression data. Plant Physiol 153, 1492-1505.

Heyer LJ, Kruglyak S, and Yooseph S. (1999). Exploring expression data: identification and analysis of coexpressed genes. Genome Res 9, 1106–1115.

Hruz T, Laule O, Szabo G, Wessendorp F, Bleuler S, Oertle L, et al. (2008). Genevestigator V3: a reference expression database for the meta-analysis of transcriptomes. Advances in Bioinformatics 420747.

Jung HS, Okegawa Y, Shih PM, Kellogg E, Abdel-Ghany SE, Pilon M, et al. (2010). Arabidopsis thaliana PGR7 encodes a conserved chloroplast protein that is necessary for efficient photosynthetic electron transport. PLoS One 5(7), e11688.

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, et al. (2006). From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34, D354–D357.

Lawler EL. (2001). Floyd-Warshall method, in Combinatorial optimization: networks and matroids. (Courier Dover Publications, pp. 86−92).

Lee JK (editor). (2010). Statistical Bioinformatics for Biomedical and Life Science Researchers, (Wiley-Blackwell, Hoboken, New Jersey), 2010.

Li M, Wu X, Wang J, and Pan Y. (2012). Towards the identification of protein complexes and functional modules by integrating PPI network and gene expression data. BMC Bioinformatics 13, 109.

MacKay D. (2003). Chapter 20. An example inference task: clustering, in Information Theory, Inference and Learning Algorithms. (Cambridge University Press. pp. 284–292).

Marri L, Pesaresi A, Valerio C, Lamba D, Pupillo P, Trost P, et al. (2010). In vitro characterization of Arabidopsis CP12 isoforms reveals common biochemical and molecular properties. J Plant Physiol 167(12), 939−50.

Marri L, Sparla F, Pupillo P, and Trost P. (2005). Co-ordinated gene expression of photosynthetic glyceraldehyde-3-phosphate dehydrogenase, phosphoribulokinase, and CP12 in Arabidopsis thaliana. J Exp Bot 56(409), 73−80.

McShane LM, Radmacher MD, Freidlin B, Yu R, Li MC, and Simon R. (2002). Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. Bioinformatics 18(11), 1462−9.

Narayanan M, Vetta A, Schadt EE, and Zhu J. (2010). Simultaneous clustering of multiple gene expression and physical interaction datasets. PLoS Comput Biol 6(4), e1000742.

Pavlidis P, Weston J, Cai J, and Noble WS. (2002). Learning gene functional classification from multiple data types. J Comp Biol 9(2), 401-411.

Peña-Castillo L, Taşan M, Myers CL, Lee H, Joshi T, Zhang C, et al. (2008). A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. Genome Biology 9,S2

Rogers S, Girolami M, Kolch W, Waters KM, Liu T, Thrall B, et al. (2008). Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models, Bioinformatics 24(24), 2894-2900.

Shiga M, Takigawa I, and Mamitsuka H. (2007). Annotating gene function by combining expression data with a modular gene network. Bioinformatics 23(13), 468−478.

Smyth GK. (2005). Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor. R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds), Springer, New York, pages 397−420.

Stuart JM, Segal E, Koller D, and Kim SK. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science 302(5643), 249−55.

Tornow S, and Mewes HW. (2003). Functional modules by relating protein interaction networks and gene expression. Nucleic Acid Res 31, 6283-6289.

Ulitsky I, and Shamir R. (2007). Identification of functional modules using network topology and high-throughput data. BMC Syst. Biol. 1:8.

Ulitsky I, and Shamir R. (2009). Identifying functional modules using expression profiles and confidence-scored protein interactions. Bioinformatics 25(9), 1158–1164.

Wong AK, Park CY, Greene SC, Bongo LA, Guan Y, and Troyanskaya OG. (2012). IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. Nucleic Acids Res 40(W1), 484-90.

Wu C, Zhu J, and Zhang X. (2012). Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. BMC Bioinformatics 13, 182.

Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, et al. (2011). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol 30, 105-111.

Zybailov B, Rutschow H, Friso G, Rudella A, Emanuelsson O, Sun Q, et al. (2008). Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. PLoS One 3(4), e1994.

Table 1. Composition of the 12 "best" cliques of CLAIM according to the level.

Level	Size	Proteins ID		
1	3	AT2G24550	AT2G38300	AT3G03000
1	3	AT1G76520	AT4G22212	AT5G63770
1	3	AT1G19700	AT1G62180	AT4G02940
1	3	AT1G23390	AT4G34920	AT5G28770
1	3	AT1G32640	AT4G37890	AT5G01840
1	3	AT1G28480	AT4G14060	AT5G11090
1	3	AT2G02810	AT4G11310	AT5G11650
1	3	AT3G02140	AT4G11280	AT5G10380
1	3	AT2G22510	AT2G47770	AT4G28040
1	3	AT1G64150	AT3G48740	AT5G17170
6	3	AT1G31280	AT2G40000	AT5G62520
6	3	AT1G24260	AT3G30530	AT5G39810

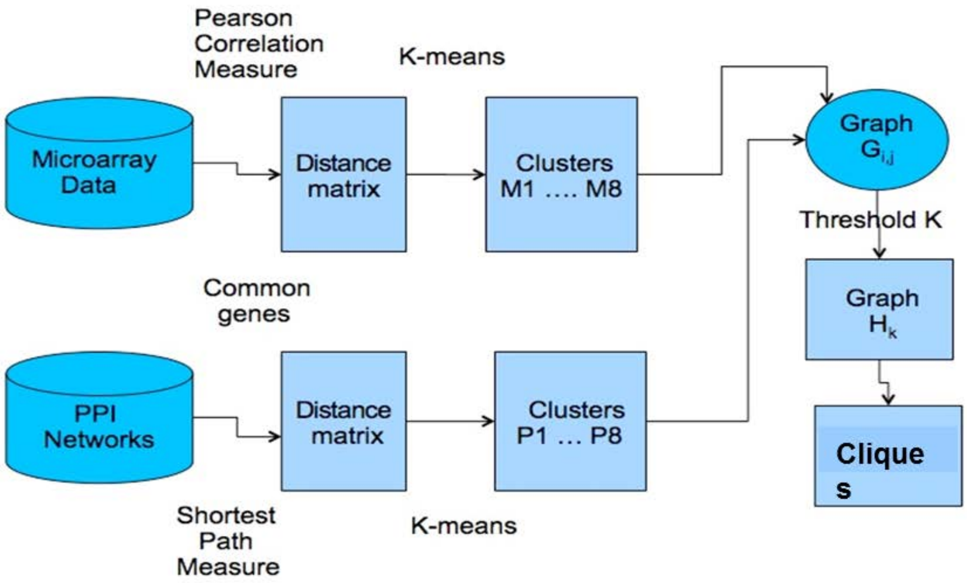


Figure 1. The flowchart of the computational process. The analysis of Microarray data and PPI networks proceeds in parallel, until the clusterings obtained by the two different sources of information are joined by the construction of graphs G_{ij}. Then graphs H_k are used to represent aggregated information on the “weight” of and edge throughout all graphs G_{ij} ; finally, cliques in H_k identify subsets of protein whose co-regulation behaviour is strongly confirmed.

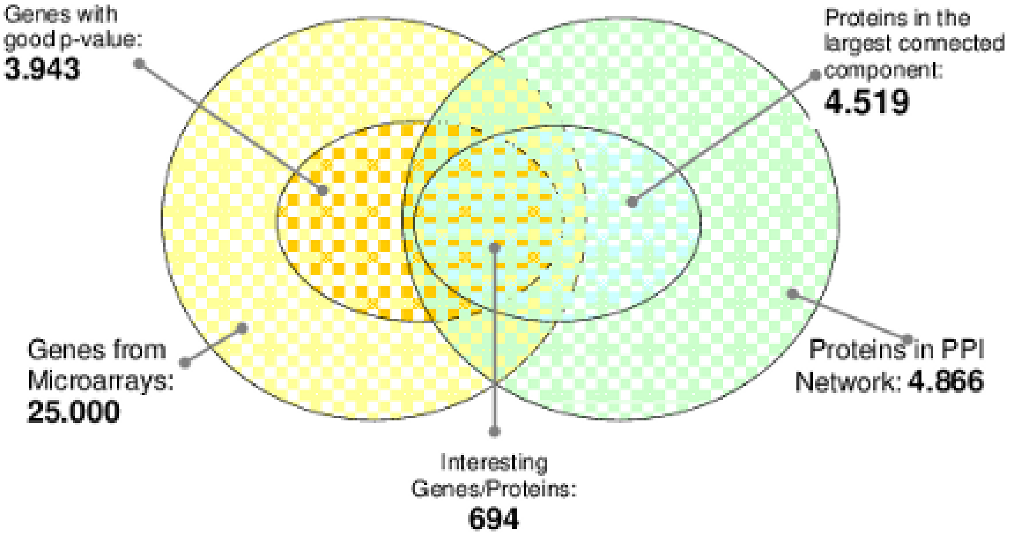
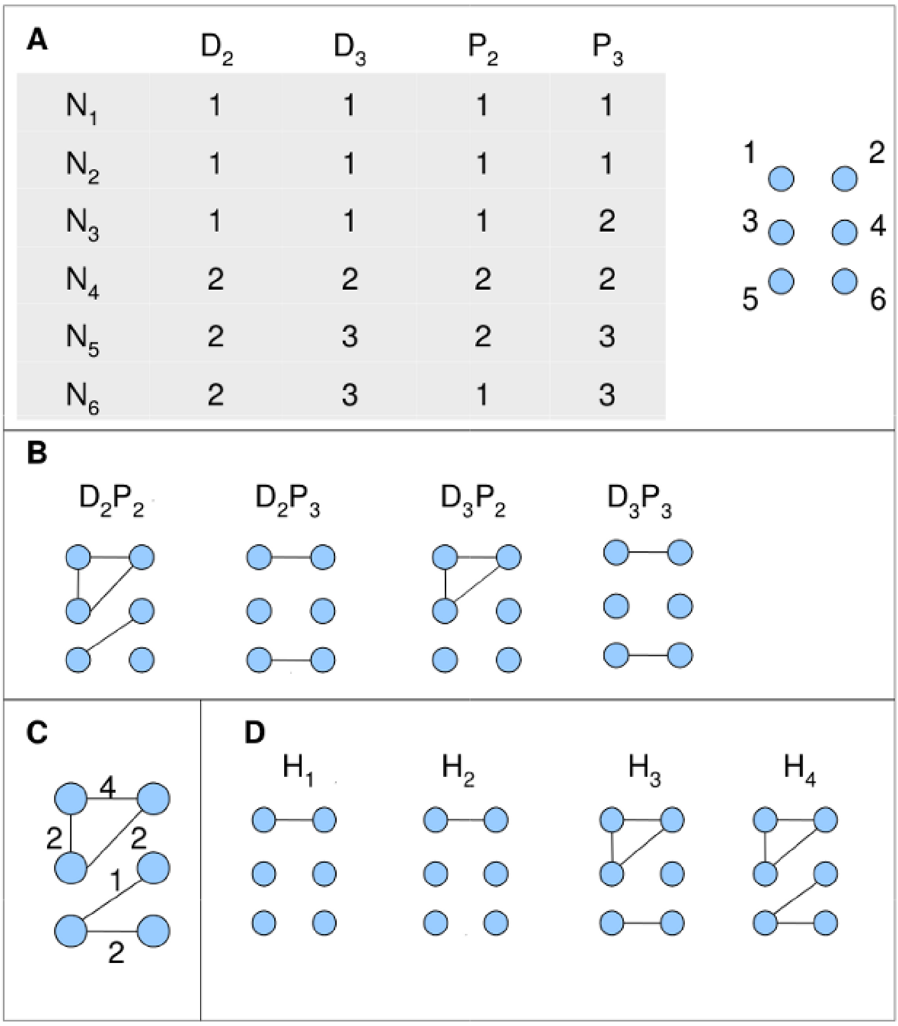
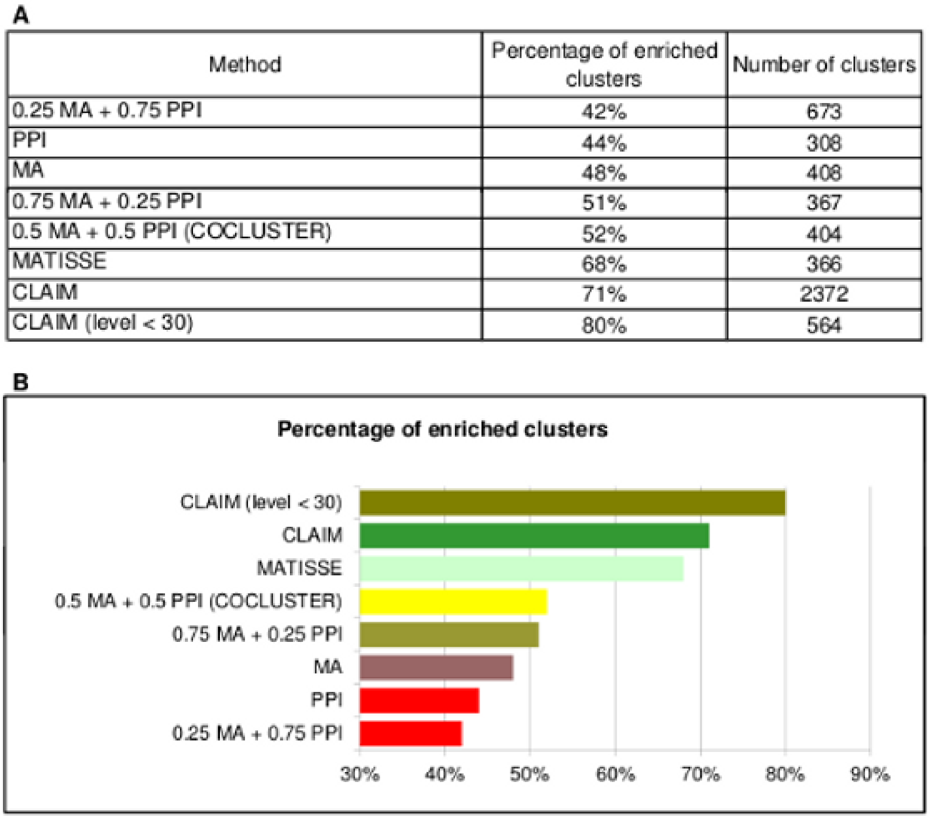


Figure 2. **Identification of the core set of genes/proteins**. A relevant kernel of 694 proteins results from the intersection of interesting sets in MA and PPI dataset.



	D_2	D_3	P_2	P_3
N_1	1	1	1	1
N_2	1	1	1	1
N_3	1	1	1	2
N_4	2	2	2	2
N_5	2	3	2	3
N_6	2	3	1	3

Figure 3. An example of the procedure leading to the identification of cliques. Six nodes N_1, N_2 ... N_6, and two classes of clusterization D and P are considered. D contains two partitions: D_2 (two clusters), and D_3 (three clusters). The same stands for P made of P_2 (two clusters) and P_3 (three clusters). In Panel A a table summarizes partitions D_2, D_3, P_2, and P_3, attributing each node to a cluster. In Panel B four graphs represent each couple of partitions belonging to class D and class P. The nodes are always located in the same position (labels removed) and each edge in the graph represents a pair of nodes falling into the same clusters for both considered partitions. For example, the edge linking node 1 to node 2 in the graph D_2P_3 is due to the presence of node 1 and node 2 in cluster labeled 1 in column D_2 and to cluster labeled 1 in column P_3; in the same way the edge linking node 5 to node 6 in the graph D_3P_3 is due to the presence of node 5 and 6 in the cluster 3 in column D_3 and in the cluster 3 of column P_3. In Panel C there is graph H. The weight of an edge in H is the number of graphs D_iP_j in which the considered edge occurs. In the example, the edge linking node 2 to node 3 has a weight equal to 2 because it occurs in graph D_2P_2 and D_3P_2. In Panel D graph H is represented with different values of the level (from 1 to 4; $H_{1,}$,$H_{2,}$ H_3 and H_4). We thus have, at level 1, the clique {1,2}; at level 2 again {1,2}; at level 3 cliques {1,2,3} and {5,6} and at level 4 cliques {1,2,3}, {4,5} and {5,6}. The best cliques are then {1,2} and {1,2,3**}.**



A

Method	Percentage of enriched clusters	Number of clusters
0.25 MA + 0.75 PPI	42%	673
PPI	44%	308
MA	48%	408
0.75 MA + 0.25 PPI	51%	367
0.5 MA + 0.5 PPI (COCLUSTER)	52%	404
MATISSE	68%	366
CLAIM	71%	2372
CLAIM (level < 30)	80%	564

Figure 4. Enriched clusters w.r.t. to KEGG pathways. Panel A: total number of clusters and percentage of enriched ones (p-value ≤ 5%) for 8 different methods. CLAIM results are reported leaving out H-cliques of large size (level ≥30). Panel B: Graphic representation of the percentage of enriched clusters (p-value ≤ 5%) for the 8 methods.

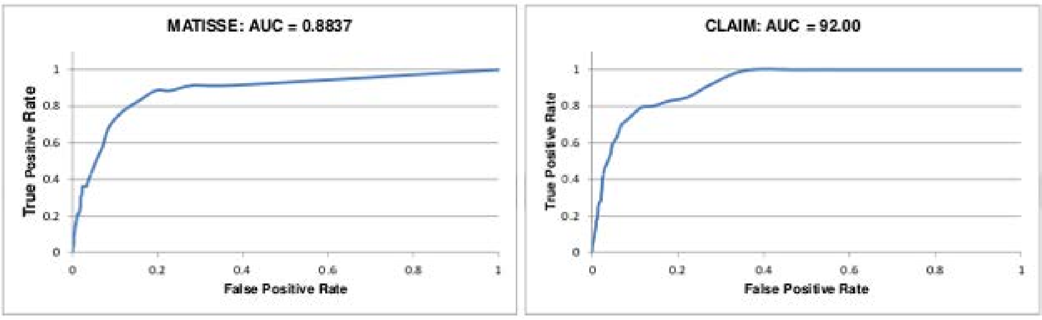


Figure 5. Comparison of predicting power of CLAIM and MATISSE. ROC curves represent true positive and false positive rates pair for different values of the decision threshold on the classification score (200 intervals from min to max values of thresholds). The training set is composed of all protein-pathway pairs where the methods produces a score. AUC values are reported over the plots.