EXPERT REVIEWS

Functional genomics and networks: new approaches in the extraction of complex gene modules

Expert Rev. Proteomics 7(1), 55–63 (2010)

Pablo Minguez and Joaquin Dopazo†

†*Author for correspondence*
Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain
jdopazo@cipf.es

The engine that makes the cell work is made of an intricate network of molecular interactions. Nowadays, the elements and relationships of this complex network can be studied with several types of high-throughput techniques. The dream of having a global picture of the cell from different perspectives that can jointly explain cell behavior is, at least technically, feasible. However, this task can only be accomplished by filling the gap between data and information. The availability of methods capable of accurately managing, integrating and analyzing the results from these experiments is crucial for this purpose. Here, we review the new challenges raised by the availability of different genomic data, as well as the new proposals presented to cope with the increasing data complexity. Special emphasis is given to approaches that explore the transcriptome trying to describe the modules of genes that account for the traits studied.

Keywords: algorithm • functional genomics • gene module • network • pathway • systems biology

The cell as a network of complexity

Since the completion of the human genome sequence in 2004, it has become apparent that the number of genes cannot solely explain the complexity of higher organisms [1]. A comparison between species reveals that, while genome sizes across them do not differ very much, the difference in the number of estimated protein–protein interactions (PPIs) increases drastically with the difference in complexity among organisms. For instance, humans have 22,258 protein-coding genes and 650,000 predicted interactions, while *Caenorhabditis elegans* has 20,158 genes but only a third of the predicted interactions (the number of genes was taken from Ensembl genome browser release 55 and interactome size predictions taken from Stumpf *et al.* [2]). This observation suggests that the complexity in the organisms must have a strong post-transcriptional component. Actually, it has been proposed that this complexity resides in the interactions between elements of the cell [2,3]. The cell can, thus, be seen as an intricate network of relationships that determines its activity. This network is composed of two types of elements: nodes and edges. Nodes represent molecules (e.g., proteins, DNA, miRNAs and other noncoding RNAs) or metabolites (e.g., lipids, carbohydrates or small molecules). The edges represent any type of relationship among the elements in the nodes. Beyond the elements that compose the network, the topology of the relationships linking them constitutes the real identity of the network. This property is of crucial importance when trying to understand the role of the network in a cellular process [4] and cannot be studied under a reductionist, gene-based perspective but under a holistic view of the cell.

Graph theory has helped biology tudy these networks and has established the basis for their description. The first discovery was that biological networks are scale-free networks [5] instead of random networks. Scale-free networks are defined by a distribution of the connections degree (defined as the number of connections of a node) that approximates to a power law $P(k) \sim k^{-\gamma}$, γ being less than 3. This indicates that the network has a small number of highly connected nodes, called hubs, while most of the nodes have a few connections in their neighboring regions. Indeed, identifying hubs is a hot topic in functional analysis [6–9].

 10.1586/EPR.09.103 ISSN 1478-9450

Another important aspect in the description of the biological networks is its dynamics [10]. Understanding the dynamics of the network implies knowing the kinetics of the relationships and its nature (activation or repression) [11], although a more detailed description would be possible and necessary in some cases (e.g., protein post-translational modifications). Although beyond the scope of this review, systems biology assigns a major role to the dynamics in the characterization of the state of a cell [12,13].

Types of biological networks

From a very broad perspective, many systems, ranging from ecosystems to cells, including functional processes within the cell, can be described in terms of networks. This fact has important implications beyond a mere visual representation of them that affect the way networks can be studied. A new biology that borrows methodologies from other fields, such as mathematics or physics, is emerging in order to study these complex entities.

Within the domain of molecular networks, the term 'network' refers to different types of relationships between molecules (typically proteins or genes and, in some cases, chemical compounds). Probably, the most relevant types are genetic networks, which account for genetic linkages among genes [14], and physical networks, which are complex systems in which the molecules involved are related by different types of interactions that can be physical, functional or regulatory. There are many possible examples of networks of this type, such as regulatory networks [15], co-expression networks (which account for the dynamics of coordinated gene expression) [16], metabolic networks [17], protein signaling networks [18] or, more generally, PPI networks, which would also include protein complexes [2].

This review focuses on physical networks; more precisely, we analyze the novel methodologies that exploit prior knowledge from PPI networks and curated pathways. Special stress is made on methods that apply such knowledge into the context of gene dynamics in transcriptomic studies to obtain modules of genes with significant impact over specific phenotypes.

Evolution of functional genomic methodologies

The availability and popularization of high-throughput technologies has paved the way for the study of different aspects of the cell behavior, such as gene-expression dynamics, PPIs, regulation, epigenetics and genetic mutations, at a genome-wide scale in an amazingly short time. Functional genomics is a field of molecular biology that makes use of the wealth of 'omic' data produced by these technologies to understand cell functionality by studying the relationships among the cell's molecular components.

Microarrays can be considered the most paradigmatic among the high-throughput technologies used in functional genomics. Many of the caveats for this methodology are common to almost any other high-throughput technology. Since the first proposals for the analysis of whole-genome gene-expression data obtained from microarrays in the late 1990s [19], this technology has matured through a series of periods in which different interests were dominating. Although microarray experiments can be used to address a large variety of biological problems, scientific literature on this subject concentrates on three main types of objectives: 'class comparison', 'class prediction' and 'class discovery' [20]. The functional interpretation of the experiments is typically made on the basis of those genes selected as relevant by tests or methods that address the above objectives. For this purpose, predefined modules of genes (gene sets) related among them by any biological property (i.e., common function, regulation or chromosomal location) are used. Functional enrichment methods are used to find if one or more of these gene modules is significantly over-represented among the relevant genes selected in the experiment [21,22]. Over-representation of a given gene module indicate that genes with a particular property have been activated or deactivated in the experiment. There are a number of available tools [23–25] that use different functional definitions to build up gene modules, such as gene ontology (GO) terms [26] or the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [27]. A large number of functional enrichment methods with different acceptation by the scientific community [101] have been proposed [21].

However, functional enrichment methods present some limitations associated with the imposition of a previous threshold [28] to the genes analyzed. Thus, genes that are relevant in a genomic experiment, in which functional enrichment is further studied, are selected exclusively on the basis of a statistic test based on experimental measurements (e.g., gene-expression intensities). The gene selection process applied in this first step does not take into account that these genes are acting cooperatively in the cell and, consequently, their behavior must be coupled to some extent. In this selection process, under the unrealistic simplification of independence among gene behaviors, much information is lost. The biological properties (e.g., function and regulation) that define the gene modules used in the functional enrichment test entails dependences [16,29,30], which are ignored and mostly lost by the application of such a threshold. Therefore, there is a paradoxical incongruence in the way functional hypotheses are tested by the functional enrichment approach, whose practical consequence is a considerable loss of statistical power [31].

Actually, it is a long-recognized fact that genes with similar overall expression often share similar functions [19,29,32]. This observation is consistent with the hypothesis of modular-behaving gene programs, where modules of genes are activated in a coordinated way to carry out functions. Under this scenario, a different class of hypothesis based on modules of functionally related genes rather than on individual genes can be tested. The activity of such modules can be studied and tested by analyzing the joint behaviors of their components. In the simplest formulation, lists of genes ranked by a measurement derived from a given experiment (e.g., differential expression when comparing cases and healthy controls) can be built. Then, the distribution of gene modules across such lists can be studied without the necessity of imposing any arbitrary threshold on the measurement. Each functional module related to the experiment and accounting for the trait studied will, consequently, be found associated to the extremes of the ranked list with highest probability.

Different methods have been proposed for this purpose such as the gene set enrichment analysis (GSEA) [33,34] or the significance analysis of function and expression (SAFE) [35], that use a nonparametrical version of a Kolmogorov–Smirnov test. Other strategies are

also possible, such as the direct analysis of functional terms weighted with experimental data [36], or model-based methods [37]. With similar accuracy, although conceptually simpler, other methods have also been proposed, such as the parametrical counterpart of the GSEA, the parametric analysis of gene set enrichment (PAGE) [38] or the segmentation test, Fatiscan [39,40]. Revisions on gene set methods can be found in different reviews [31,41,42].

The methods mentioned previously consider gene modules as categorical, unstructured entities. This abstraction, although operatively useful, is far from the reality. Proteins, and by extension their templates (genes), form an intricate network of relationships where genes, proteins and other metabolites are involved [43,44]. In this network of life, every element has a specific role determined by their position. A clear example of structured functional classes are those defined by databases, such as KEGG [45], BioCarta, Reactome [46], WikiPathways or the Pathway Interaction Database [47]; initiatives that are making an invaluable effort to compile of knowledge on metabolic and signaling pathways. Other example are the PPIs datasets (ppis) produced by high-throughput techniques, such as yeast two hybrid (Y2H), tandem affinity purification (TAP) and high-throughput mass spectrometry (MS). Reviews on these and related methodologies can be found in [48,49]. They represent binary physical interactions between proteins that taken in the right perspective are seen as a global network, known as the interactome.

Novel methodologies in functional genomics

Nowadays there is no better approximation to represent complex systems, such as multigenic diseases, than the network perspective [50]. Thus, new methods are necessary to deal with this new challenge. This review summarizes the new achievements in this field with a focus in methodologies that study the transcriptome. Strategies able to extract and interpret the relevant networks and pathways with an accurate treatment of their features from the data produced by genome-scale experiments will also be discussed.

The main difference between curated pathways, such as KEGG or BioCarta, with respect to PPI networks, is that, while the description of the former represent directed interactions (e.g., A activates B) the relationships are normally undirected for the latter (e.g., A and B interact). Also, for the connected character of the PPI network, many, if not all, of the nodes are linked directly or indirectly through other nodes. In this respect, physical PPI networks constitute a more accurate description of the functional interoperability of the molecules in the cell than the conventional functional modules, which describe functions as watertight compartments. In fact, this vision is changing, and there are recent descriptions of complex functionalities, such as metabolism, in which classical pathways are substituted by a complex metabolic network [17,51]. Certainly, subnetworks within the global network can play specific roles and different methods with the aim to extract such submodules, which will be commented on later, have been proposed.

Another difference between PPI networks and other conventional functional modules is the way that they are defined. While conventional modules are typically the result of a manual curation process, the PPI networks are obtained in different experiments (e.g., Y2H and MS), which define physical relationships among pairs of proteins. A consequence of this fact is that the relationships that define the topology of the network itself do not allow an easy delimitation of submodules within.

The classification of the methods that exploit the information on the functional or physical relationships between genes or proteins to analyze omics experiments (e.g., gene-expression or genome-wide genotyping) is not an easy task. We have used a naive separation between methods oriented to pathways or to PPI networks, which constitutes the most comprehensible classification for the end user from an operative standpoint.

Methods applied to pathways from curated repositories

Contrary to PPI data, metabolic and signaling pathways from curated repositories have been used extensively in the functional profiling of genome-scale experiments. Together with GO terms, they are by far the most used functional definitions in the context of the conventional functional enrichment approaches, as well as in gene-set analyses (both discussed previously). Under such approaches, all the genes annotated within the pathways are considered with the same weight, and the internal structure is dismissed. However, under the systems biology perspective, the topology of the pathway is a much more important aspect than the simple list of individual components. A pathway can be activated or deactivated depending on the relative relationships among the genes that are deregulated. For instance, in a signaling pathway, the signal transmitted from input to output genes can be drastically affected if the genes located between them are deregulated. The relative position of up- or down-regulated genes is a factor that cannot be underestimated when identifying pathways related to traits, such as disease phenotypes.

In any case, it is obvious that considering functional classes as discrete, unstructured entities composed by members with an equal weight, was quite an unrealistic assumption, which reduced the statistical power in any testing framework. Thus, a new family of methods was proposed that could be called core-based algorithms, as the main concept introduced is that not all genes contribute in a similar manner to the pathway activity in a determined cell condition [52]. Through different flavors, these algorithms search for the set (core) of the genes within the pathway that more significantly account for its functionality. Tomfohr *et al.* proposed an algorithm that translates gene-expression levels into pathway-activity levels derived from singular-value decomposition (SVD) [53]. For each of the pathways analyzed, they generated a gene-expression matrix and took the first eigenvector that is claimed to be representative of the activity of the pathway. In a similar approach, condition-responsive genes (CORGs) could be identified and used in a second step to discriminate among samples in a microarray experiment [54]. Eigenvectors are also used in other approaches to remove high-frequency components, mostly composed by noise, from the networks analyzed [55]. This idea of existing cores of genes within functional modules, which contribute to their detection by functional profiling methods, owing to their internal coexpression, was applied by Montaner *et al.* to generate a coherence index to be used for measuring functional module coregulation in GO terms

and KEGG pathways [52]. Interestingly, they found that only 30% of the modules defined by GO terms and 57% of the modules defined by KEGG pathways display an internal correlation higher than that expected by chance.

The leitmotiv in these approaches is the removal of different types of noise within the pathways, especially redundancy and genes not relevant for the activity of the pathway under the studied condition (e.g., genes not expressed in the tissue under study). Although these approaches produce comparatively better results than their gene set of functional enrichment counterparts, they still do not use the information of the topological features of the pathways.

Several solutions attempted to go beyond the mere cleansing of the pathways by taking into account, from different perspectives, the internal structure of the pathways. They belong to 'topological-based algorithms'. In early attempts, prior biological information was introduced by means of distances between the genes within the pathways in order to correct their correlation within microarray experiments. Thus, a combined measurement that includes co-expression and distance according to the topology of the pathways has been used to improve the performance of coclustering methods [56] or in supervised studies, using metabolic pathways [57]. Other more sophisticated algorithms use the network information to correct gene expression measurements by more complex algorithms, such as Markov random field [58,59].

More recently, at least three novel approaches are the future directions for this type of study. Thomas *et al.* present a set of algorithms with a wide scope, both in the hypothesis tested and in the type of data they can deal with, but similar in their approach to the problem [60]. Of particular interest is SEPEA_NT1, which compares the genes within every pathway to the rest of the genes in the system. For the genes in a pathway, the algorithm calculates two scores: the heavy-ends rule (HER), which takes higher values when the genes associated to a class (e.g., disease) are in the terminal part of a pathway, and the distance rule (DR), which increases the weight of the genes associated to the class if they are close to each other in the pathway. These two scores are combined and the significant pathways associated to the case of study are extracted using a permutation test.

Efroni *et al.* identified signaling pathways associated to cancer stages by determining two scores for every pathway: a consistency score and an activity score [61]. The consistency score identifies those pathways whose genes in input and output positions are consistent in terms of expression. Thus, the state of every gene (up or down) is estimated by fitting their expression distribution to a mixture of two γ distributions that represent the active and inactive states over the whole experiment. Then, for every interaction the probability of occurrence is calculated by a joint probability of the input and output genes. The global consistency of a pathway is given by the average of the consistency of every interaction by comparing the probability of the output genes and its real state. The pathway activity score is the average of all the probabilities for the interactions in a pathway. The result is the transformation of gene-expression values into a matrix where every pathway shows its two scores over the samples. Finally, the algorithm detects pathways that are able to discriminate between phenotypic classes.

Another related algorithm is the so-called impact analysis [62,63]. This approach combines elements from classical enrichment analysis with topological measurements, calculating two independent probabilities: the probability of having differentially expressed genes in a pathway (P_{NDE}), and the probability of the pathway of being deregulated, estimated as the propagation of the expression changes of each gene along the pathway topology (P_{PERT}). P_{PERT} is calculated over the perturbation factor (PF) of every gene in the pathway. This PF takes into account the expression change of a gene and the PFs of the genes that are up- and down-stream in the cascade. It also introduces the concept of activation and inhibition in the calculation of the PF. A bootstrap procedure is used to assess the significance of the observed total pathway perturbation.

Obviously, the more detailed the description of the biology behind the model tested, the better. Thus, the last two proposals seem to represent the most sophisticated way to exploit the information available on pathway topologies to understand the functional consequences of the changes in gene-expression levels [61,62]. It is obvious that the application of such methods requires knowledge of the internal structure of the entities tested (e.g., KEGG pathways, BioCarta and Reactome).

Methods applied to PPI networks

Protein–protein interactions play a central role at almost every level of cell activity – they are involved in the structure of organelles (structural proteins), transport machinery (nuclear pore importins), response to a stimulus (signaling cascades), regulation of gene expression (transcription factors), protein modification (kinases) and many other processes. The inference and proper use of this type of information are of crucial importance to understand cell behavior. We review the more relevant methodologies that have been proposed for the identification of phenotype-responsive PPI subnetworks.

A subnetwork is a subset of the whole network (interactome). Owing to their putative functional role, subnetworks with internal node connectivity higher than its connectivity to other nodes are of special interest. Many attempts have been made to explore the whole interactome in order to detect such subnetworks. Most of the approaches used were based on the application of clustering methods to weighted matrices that describe the network of PPIs. For example, the number of experiments that support each interaction has been used as the value for the corresponding entry of the PPI matrix [64]. Other authors use the shortest paths among pairs of nodes to measure the relationship among them [65] or topological features of the network, such as the 'betweeness' [66,67]. Central nodes, with a high betweenness, can help to define the boundaries of the subnetworks because many of the shortest paths pass through them, and the action of removing them from the network would lead to the disconnection of subnetworks. Other approaches make use of other properties, such as the conservation of subnetworks across species [68].

The subnetworks obtained by these methodologies constitute gene modules that may be enriched in proteins with related biological functionalities, as indicated by its significant enrichment in GO terms [69] or by its co-occurrence within the literature [67]. Indeed,

it has also been shown that there are subnetworks associated to diseases [50,70–72]. It has also been reported that genes deregulated in cancer confer fragility to the interactome network [73,74]. The analysis of the human interactome reported that proteins encoded by genes mutated in inherited genetic disorders are likely to interact with proteins known to cause similar disorders [71].

Nevertheless the interactome obtained from high-throughput techniques conforms with an abstract scaffold that describes all the possible PPIs, but it does not provide information about particular conditions, cell developmental stage or cell type in which a particular PPI occurs (if any). To infer a case-specific interactome, it is necessary to integrate other types of data that provide information that allows inferring the active PPIs at a particular condition.

As in conventional functional profiling analyses based on categorical, unstructured functional classes (e.g., GO terms) PPI data can be used for the functional interpretation of genome-wide experiments. The aim is to extract the gene modules, defined as subnetworks, which are associated to a particular cell phenotype beyond the random expectation. A few approaches have been proposed to statistically test such associations, which are now described.

Extracting subnetworks from genomic experiments

Actually, the majority of the algorithms that have been proposed in this field use different scoring systems to measure the association of subnetworks to the differences between two experimental conditions in a transcriptomic experiment. In an early paper in 2002, Ideker *et al.* laid the foundations for this type of approach [75]. The authors introduced a scoring-based measure of groups of genes, generated using the interactome as scaffold. All possible subnetworks are scored based on their differential expression over two or more classes in a microarray experiment before searching for the highest-scoring subnetwork by a procedure based on simulated annealing. Other different approaches using distinct scoring procedures have been proposed [76,77].

An interesting approach to evaluate the association of subnetworks to a determined disease through their enrichment in pre-established gene signatures already associated to the disease has recently been proposed [78]. The gene signatures are defined by differential expression tests. The relative expression of the genes in the signature is mapped to the global network of PPIs. From the combination of the interactomic and transcriptomic information, a high-scoring matrix (HSM) is extracted as a subnetwork that is highly transcriptionally affected in the disease. Finally, the hypothesis that a particular gene signature is enriched into the subnetwork is tested. In brief, PPI subnetworks are used as gene modules in a conventional functional enrichment analysis, although the topological relationships among the proteins in the network are not exploited.

Other proposals followed the same rationale, although applying an edge-based algorithm [79]. In this case, a greedy search is performed through the interactome, taking proteins that are known to be associated to the disease under study as seeds. The subnetworks found are then evaluated according to an aggregative and normalized score of their edges. The score for each edge is calculated taking into account the correlation in the expression measurements of the connected proteins and their individual differential expression over the classes compared in the microarray experiment. Then, the highest-scoring subnetwork is found by a similar procedure as that used by Ideker *et al.* [75]. Finally, the authors evaluated if the subnetworks were enriched in any functional class comparing its proteins annotation to the background annotations. This final evaluation of the subnetworks has become popular [67,69]. There are even specific tools for doing this task, such as Bingo [80], a Java applet that can be integrated into the Cytoscape visualization tool [81], which performs GO enrichment analysis to the nodes that configure subnetworks. Although highly informative, in the case of negative results, this test does not guarantee that the studied subnetwork is not a functional module, owing to the lack of annotation. In fact, functional analysis using PPIs does not always overlap with functional profiling counterparts using biologically relevant terms, such as GO or KEGG [82,83].

Similarly, other authors proposed detection of the active subnetworks, taking as seeds proteins with more than five interactions [84]. They score the subnetworks by using a multivariate analysis of variance over the gene-expression data of the phenotype under study. The sequence of aggregation of nodes is formed according to a threshold based on the score performance. Once the candidate subnetworks are extracted, a permutation test evaluates their significance.

Alternatively, Minguez *et al.* proposed a method that uses the structural features of the subnetworks as a statistical criterion for its evaluation [85]. The authors generate subnetworks from lists of differentially expressed genes among the experimental conditions compared. This list of genes can, in fact, be preselected by any other criteria, given that the method is independent from this selection process. From this list, they derive the associated minimal connected network (MCN), which is the minimal network that connects all the nodes in the list. The shortest paths among all the pairs of nodes in the list are calculated using the Dijkstra algorithm [86]. The paths that connect two proteins either directly or through an additional protein (initially not contained in the list) are introduced into the MCN. Indeed, the inclusion of proteins not contained in the list constitutes an interesting feature of this proposal. Actually, it has been reported that proteins not preselected by expression profiling experiments, and detected because they connected two or more preselected proteins, were related to diseases [78,87,88].

Expert commentary

Functional genomics is a fast-growing discipline. As systems biology viewpoints are gaining influence, the new methodologies are closer to a more realistic management of the biological modules, and this is reflected in the evolution of the methodologies reviewed in this manuscript.

For curated signaling pathways, studies that take into account the topology of the pathways point to a promising direction [60,61,63]. The well-defined structure and annotation of these type of data represent an excellent starting point as proof-of-concept for new algorithms. The methods to be developed in this area must take into account some of the elements these three methods have separately, such as:

- Inclusion of topological information with special attention in input and output genes;
- Incorporation of the direction and nature of the interactions (activations and inhibitions);
- Evaluation of the consistency of the expression data with the network topology;
- Proper evaluation of the pathways to check their statistical significance.

An obvious step forward could be to increase the scope of studies that exploit the topological relationships between proteins in pathways. Thus, initiatives that attempt to generate global curated pathways provide a standpoint for these future generalization efforts [17,51]. In a recent imaginative attempt, the authors generated a novel entity that represents a network of pathways where the nodes are the pathways themselves and the edges are defined by their shared genes [89].

Working with PPI networks is far more difficult than dealing with curated pathways – the data are intrinsically noisy and the network has not a predefined structure of subnetworks. On the other hand, it constitutes an excellent substrate for a more exploratory biology, where new modules of genes can be discovered beyond the already known curated pathways and associated to phenotypes, such as diseases. As a corollary, there are at least two aspects to consider when approaching the development of methodologies to deal with omics data in the context of networks. The first is that the topology of the network has to be taken into account. This feature can provide tools to study the networks (graph theory), as well as ways to classify them and their constituting elements (e.g., hubs and their implications in diseases). The second important aspect to be considered is the proper statistical evaluation of the relationship between the subnetwork extracted and the phenotype under study.

Five-year view

One of the practical consequences of the introduction of systems biology concepts in the area of functional genomics is the use of networks to describe different aspects of the cell biology. This has constituted a revolution with long-term implications that are difficult to predict. The biological systems (diseases or particular cellular processes) will no longer be seen as static, isolated entities. Rather, a vision of them as part of a global network of interactions with no limits in its scope, from molecules to organisms, will gradually gain acceptance.

Biology has been climbing from a reductionist approach, based on the study of isolated genes, to a more holistic approach based on pathways, which constitute an important advancement in the knowledge of the biological phenomena. Pathways, however, are only one step toward the whole description of the cell biology. The perspectives for the future of functional genomics are clearly associated with the development of new methodologies able to deal with these new challenges. New methods should surpass the present scenario, in which the network is divided into pathways, and provide more general models to study realistic situations in which global metabolic networks or systems, including proteins, lipids, carbohydrates and drugs, could be studied easily. An example of a visionary proposal with clear implications in biomedicine is a recent study of networks of drug–drug relationships to predict new targets for a drug through side-effect similarities [90]. Another interesting viewpoint is the network of disease genes linked by known disorder–gene associations, which constitutes a useful tool for extracting links between diseases and by extension between their treatments [91].

In parallel with the development of new strategies for exploring this new information, the quality of the data we use to generate such information also need to be improved. The implications for this include the development of high-throughput techniques able to generate more accurate PPIs, and protein and other metabolites interactions, but also a better annotation of these interactions. There is still a big gap between curated pathways and network data coming from high-throughput experiments, which needs to be filled. Some recent works exemplify the generation of pathways from PPI data and describe the strategies followed to obtain a complete map of the interactome of model species [92,93].

Financial & competing interest disclosure

This work is supported by grants from project BIO BIO2008-04212 from the Spanish Ministry of Science and Innovation. The National Institute of Bioinformatics (www.inab.org), is a platform of Genoma España. The CIBER de Enfermedades Raras is an initiative of the ISCIII. This work is also partly supported by a grant (RD06/0020/1019) from Red Temática de Investigación Cooperativa en Cáncer (RTICC), Instituto de Salud Carlos III (ISCIII), Spanish Ministry of Science and Innovation. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Key issues

- As high-throughput techniques are producing massive amounts of data, the challenge now resides in filling the gap between data and information.
- Almost all biological systems can be described in terms of networks setting up the new basics of a new biology that need to be able to accurately describe these complex systems.
- Data integration is of crucial importance when approaching the description of cell phenotypes.
- Both, curated pathways (metabolic and signaling) and protein–protein interaction networks represent an excellent scaffold for the study of complex traits, including diseases.
- New methodologies need to be developed for the correct treatment of networks within this context.

References

Papers of special note have been highlighted as:
- • of interest
- •• of considerable interest

1 International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 431(7011), 931–945 (2004).

2 Stumpf MP, Thorne T, de Silva E *et al.* Estimating the size of the human interactome. *Proc. Natl Acad. Sci. USA* 105(19), 6959–6964 (2008).

3 Copley RR. The animal in the genome: comparative genomics and evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363(1496), 1453–1461 (2008).

4 Yeger-Lotem E, Sattath S, Kashtan N *et al.* Network motifs in integrated cellular networks of transcription-regulation and protein–protein interaction. *Proc. Natl Acad. Sci. USA* 101(16), 5934–5939 (2004).

5 Barabasi AL, Albert R. Emergence of scaling in random networks. *Science* 286(5439), 509–512 (1999).

6 He X, Zhang J. Why do hubs tend to be essential in protein networks? *PLoS Genet.* 2(6), E88 (2006).

7 Sporns O, Honey CJ, Kotter R. Identification and classification of hubs in brain networks. *PLoS ONE* 2(10), E1049 (2007).

8 Said MR *et al.* Global network analysis of phenotypic effects: protein networks and toxicity modulation in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* 101(52), 18006–18011 (2004).

9 Lim J, Hao T, Shaw C *et al.* A protein–protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* 125(4), 801–814 (2006).

10 Tyson JJ, Chen K, Novak B. Network dynamics and cell physiology. *Nat. Rev. Mol. Cell. Biol.* 2(12), 908–916 (2001).

11 Tegner J, Bjorkegren J. Perturbations to uncover gene networks. *Trends Genet.* 23(1), 34–41 (2007).

12 Kitano H. Systems biology: a brief overview. *Science* 295(5560), 1662–1664 (2002).

• **Basic review in which it is suggested that in order to properly understand biology at the system level, the structure and dynamics of cellular and organismal function must be studied, rather than the characteristics of isolated parts of a cell or organism.**

13 Bruggeman FJ, Westerhoff HV. The nature of systems biology. *Trends Microbiol.* 15(1), 45–50 (2007).

14 Kelley R, Ideker T. Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.* 23(5), 561–566 (2005).

15 Lee TI, Rinaldi NJ, Robert F *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298(5594), 799–804 (2002).

16 Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302(5643), 249–255 (2003).

17 Ma H, Sorokin A, Mazein A *et al.* The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol. Syst. Biol.* 3, 135 (2007).

18 Papin JA, Hunter T, Palsson BO, Subramaniam S. Reconstruction of cellular signalling networks and analysis of their properties. *Nat. Rev. Mol. Cell Biol.* 6(2), 99–111 (2005).

19 Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* 95(25), 14863–14868 (1998).

20 Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.* 7(1), 55–65 (2006).

21 Dopazo J. Functional interpretation of microarray experiments. *Omics* 10(3), 398–410 (2006).

22 Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 21(18), 3587–3595 (2005).

23 Al-Shahrour F, Diaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics* 20(4), 578–580 (2004).

24 Draghici S, Khatri P, Bhavsar P, Shah A, Krawetz SA, Tainsky MA. Onto-tools, the toolkit of the modern biologist: onto-express, onto-compare, onto-design and onto-translate. *Nucleic Acids Res.* 31(13), 3775–3781 (2003).

25 Zeeberg BR, Feng W, Wang G. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* 4(4), R28 (2003).

26 Ashburner M, Ball CA, Blake JA *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25(1), 25–29 (2000).

27 Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32(Database issue), D277–D280 (2004).

28 Pan KH, Lih CJ, Cohen SN. Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proc. Natl Acad. Sci. USA* 102(25), 8961–8965 (2005).

29 Mateos A, Dopazo J, Jansen R, Tu Y, Gerstein M, Stolovitzky G. Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. *Genome Res.* 12(11), 1703–1715 (2002).

30 Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray data sets. *Genome Res.* 14(6), 1085–1094 (2004).

31 Dopazo J. Formulating and testing hypotheses in functional genomics. *Artif. Intell. Med.* 45(2–3), 97–107 (2009).

32 Lee JM, Sonnhammer EL. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.* 13(5), 875–882 (2003).

33 Mootha VK, Lindgren CM, Eriksson KF *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 34(3), 267–273 (2003).

34 Subramanian A, Tamayo P, Mootha VK *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* 102(43), 15545–15550 (2005).

35 Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* 21(9), 1943–1949 (2005).

36 Smid M, Dorssers LC. GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate gene ontology terms. *Bioinformatics* 20(16), 2618–2625 (2004).

37 Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20(1), 93–99 (2004).

38 Kim SY, Volsky DJ. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* 6, 144 (2005).

39 Al-Shahrour F, Arbiza L, Dopazo H *et al.* From genes to functional classes in the study of biological systems. *BMC Bioinformatics* 8, 114 (2007).

40 Al-Shahrour F, Diaz-Uriarte R, Dopazo J. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics* 21(13), 2988–2993 (2005).

41 Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37(1), 1–13 (2008).

42 Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 23(8), 980–987 (2007).

43 Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5(2), 101–113 (2004).

44 Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. *Nature* 402(6761 Suppl.), C47–C52 (1999).

•• **Seminal paper in which the concept of the functional module is introduced.**

45 Kanehisa M, Araki M, Goto S *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36(Database issue), D480–D484 (2008).

46 Vastrik I, D'Eustachio P, Schmidt E *et al.* Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.* 8(3), R39 (2007).

47 Schaefer CF, Anthony K, Krupa S *et al.* PID: the Pathway Interaction Database. *Nucleic Acids Res.* 37(Database issue), D674–D679 (2009).

48 Falk R, Ramström M, Ståhl S, Hober S. Approaches for systematic proteome exploration. *Biomol. Eng.* 24(2), 155–168 (2007).

49 Berggard T, Linse S, James P. Methods for the detection and analysis of protein–protein interactions. *Proteomics* 7(16), 2833–2842 (2007).

50 Badano JL, Katsanis N. Beyond Mendel: an evolving view of human genetic disease transmission. *Nat. Rev. Genet.* 3(10), 779–789 (2002).

51 Duarte NC, Becker SA, Jamshidi N *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl Acad. Sci. USA* 104(6), 1777–1782 (2007).

52 Montaner D, Minguez P, Al-Shahrour F, Dopazo J. Gene set internal coherence in the context of functional profiling. *BMC Genomics* 10(1), 197 (2009).

53 Tomfohr J, Lu J, Kepler TB. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics* 6, 225 (2005).

54 Lee E, Chuang HY, Kim JW, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.* 4(11), E1000217 (2008).

55 Rapaport F, Zinovyev A, Dutreix M, Barillot E, Vert JP. Classification of microarray data using gene networks. *BMC Bioinformatics* 8, 35 (2007).

56 Hanisch D, Zien A, Zimmer R, Lengauer T. Co-clustering of biological networks and gene expression data. *Bioinformatics* 18(Suppl. 1), S145–S154 (2002).

57 Rahnenführer J, Domingues FS, Maydt J, Lengauer T. Calculating the statistical significance of changes in pathway activity from gene expression data. *Stat. Appl. Genet. Mol. Biol.* 3, Article 16 (2004).

58 Wei Z, Li H. Markov random field model for network-based analysis of genomic data. *Bioinformatics* 23(12), 1537–1544 (2007).

59 Wei P, Pan W. Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. *Bioinformatics* 24(3), 404–411 (2008).

•• **Rigorous and statistically sound proposal for introducing network information in the analysis of genomic data.**

60 Thomas R, Gohlke JM, Stopper GF, Parham FM, Portier CJ. Choosing the right path: enhancement of biologically relevant sets of genes or proteins using pathway structure. *Genome Biol.* 10(4), R44 (2009).

61 Efroni S, Schaefer CF, Buetow KH. Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PLoS ONE* 2(5), E425 (2007).

62 Draghici S, Khatri P, Tarca AL *et al.* A systems biology approach for pathway level analysis. *Genome Res.* 17(10), 1537–1545 (2007).

63 Tarca AL, Draghici S, Khatri P *et al.* A novel signaling pathway impact analysis. *Bioinformatics* 25(1), 75–82 (2009).

64 Pereira-Leal JB, Enright AJ, Ouzounis CA. Detection of functional modules from protein interaction networks. *Proteins* 54(1), 49–57 (2004).

65 Rives AW, Galitski T. Modular organization of cellular networks. *Proc. Natl Acad. Sci. USA* 100(3), 1128–1133 (2003).

66 Girvan M, Newman ME. Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA* 99(12), 7821–7826 (2002).

67 Wilkinson DM, Huberman BA. A method for finding communities of related genes. *Proc. Natl Acad. Sci. USA* 101(Suppl. 1), 5241–5248 (2004).

68 Sharan R, Suthram S, Kelley RM *et al.* Conserved patterns of protein interaction in multiple species. *Proc. Natl Acad. Sci. USA* 102(6), 1974–1979 (2005).

69 Luo F, Yang Y, Chen CF, Chang R, Zhou J, Scheuermann RH. Modular organization of protein interaction networks. *Bioinformatics* 23(2), 207–214 (2007).

70 Brunner HG, van Driel MA. From syndrome families to functional genomics. *Nat. Rev. Genet.* 5(7), 545–551 (2004).

71 Gandhi TK, Zhong J, Mathivanan S *et al.* Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet.* 38(3), 285–293 (2006).

72 Dezso Z, Nikolsky Y, Nikolskaya T *et al.* Identifying disease-specific genes based on their topological significance in protein networks. *BMC Syst. Biol.* 3, 36 (2009).

73 Hernández P, Huerta-Cepas J, Montaner D *et al.* Evidence for systems-level molecular mechanisms of tumorigenesis. *BMC Genomics* 8, 185 (2007).

74 Rambaldi D, Giorgi FM, Capuani F, Ciliberto A, Ciccarelli FD. Low duplicability and network fragility of cancer genes. *Trends Genet.* 24(9), 427–430 (2008).

75 Ideker T, Ozier O, Schwikowski B *et al.* Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18(Suppl. 1), S233–S240 (2002).

•• **Constitutes the earliest description of the definition of subnetworks for a predefined set of genes (commonly obtained upon the application of a differential expression test in a microarray experiment).**

76 Calvano SE, Xiao W, Richards DR *et al.* A network-based analysis of systemic inflammation in humans. *Nature* 437(7061), 1032–1037 (2005).

77 Huang SS, Fraenkel E. Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Sci. Signal.* 2(81), 40 (2009).

78 Liu M, Liberzon A, Kong SW *et al.* Network-based analysis of affected biological processes in Type 2 diabetes models. *PLoS Genet.* 3(6), E96 (2007).

79 Guo Z, Wang L, Li Y *et al.* Edge-based scoring and searching method for identifying condition-responsive protein protein interaction sub-network. *Bioinformatics* 23(16), 2121–2128 (2007).

80 Maere S, Heymans K, Kuiper M. BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21(16), 3448–3449 (2005).

81 Shannon P, Markiel A, Ozier O *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13(11), 2498–2504 (2003).

82 Liu J, Hughes-Oliver JM, Menius JA Jr. Domain-enhanced analysis of microarray data using GO annotations. *Bioinformatics* 23(10), 1225–1234 (2007).

83 Minguez P, Dopazo J. *Protein Interactions for Functional Genomics, in Biological Data Mining in Protein Interaction Networks.* Li X-L, Ng S-K (Eds). IGI Global, PA, USA, 223–238 (2009).

84 Hwang T, Park T. Identification of differentially expressed subnetworks based on multivariate ANOVA. *BMC Bioinformatics* 10, 128 (2009).

85 Minguez P, Götz S, Montaner D, Al-Shahrour F, Dopazo J. SNOW, a web-based tool for the statistical analysis of protein-protein interaction networks. *Nucleic Acids Res.* 37(Web Server issue), W109–W114 (2009).

86 Dijkstra E. A note on two problems in connexion with graphs. *Numerische Mathematik* 1, 269–271 (1959).

87 Xu J, Li Y. Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics* 22(22), 2800–2805 (2006).

88 Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* 3, 140 (2007).

89 Francesconi M, Remondini D, Neretti N *et al.* Reconstructing networks of pathways via significance analysis of their intersections. *BMC Bioinformatics* 9(Suppl. 4), S9 (2008).

90 Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. *Science* 321(5886), 263–266 (2008).

91 Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. *Proc. Natl Acad. Sci. USA.* 104(21), 8685–8690 (2007).

• **Challenging paper that presents a network of disorders and disease genes linked by known disorder–gene associations. Such networks can be considered platforms to explore, in a single graph-theoretic framework, all known phenotype and disease–gene associations.**

92 Mak HC, Daly M, Gruebel B, Ideker T. CellCircuits: a database of protein network models. *Nucleic Acids Res.* 35(Database issue), D538–D545 (2007).

93 Schwartz AS, Yu J, Gardenour KR, Finley RL Jr, Ideker T. Cost-effective strategies for completing the interactome. *Nat. Methods* 6(1), 55–61 (2009).

Website

101 Profiling tool: functional profiling sites http://bioinfo.cipf.es/docus/tools-citations/functional_profiling

Affiliations

- Pablo Minguez
Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain
and
European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany
- Joaquin Dopazo
Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain
and
CIBER de Enfermedades Raras (CIBERER), Valencia, Spain
and
Functional Genomics Node, (INB) at CIPF, Valencia, Spain
jdopazo@cipf.es