Genetic networks for the functional study of genomes

Antonio G. Pisabarro, Gumer Pérez, José Luis Lavín and Lucía Ramírez

Advance Access publication date 25 June 2008

Abstract

The high-throughput analytical techniques used in genome, proteome and metabolome studies produce large sets of data that must be studied using appropriate tools. The construction of networks linking different genetic elements and/or functions makes it possible to obtain an integrated view of the cell molecular biology and will eventually help us to predict complex phenotypes from molecular data. Genetic networks can be constructed using different types of data such as genes involved in the control of complex phenotypic traits, genes controlling global gene expression, genetic elements involved in the same metabolic process, gene products interacting physically between them. The connections linking these genetic elements in the network reflect the genetic, physical and/or functional interaction among them. All these networks share common properties and reflect the different layers of the cell's complexity. In this review, we will study how different types of networks can be constructed, how the different networks complement each other and how this information can be used to obtain an integrated picture of the cell.

Keywords: genetic networks; QTL analysis; expression profiling; synthetic lethality; interactome networks

INTRODUCTION

The purpose of this work is to review how genetic and functional networks can be constructed using different types of data produced by high-throughput analytical technologies. In a specific context, these genetic networks are useful tools for the functional analysis of genomes because they reveal different types of interactions among genetic elements, and because they can be used as a roadmap for the study of genomes in organisms with reduced pre-genomic information. The ease of genome sequencing makes the whole-genome sequence of an organism the first genetic or molecular piece of information about it, in an increasing number of cases. This is specially the case in some metagenomic studies. In other cases, such as the identification of pathogenicity islands in bacterial isolates, the whole-genome sequence is an approach that produces results faster than the classical cloning and analysis techniques. When this gene finding strategy is used, however, the impact of the genetic environment of the whole cell elements is not frequently taken into account due to the lack of instruments to study this type of system interactions.

The molecules and functions that form a living cell are densely interconnected. These connections can be physical (protein–protein, protein–DNA, protein–RNA, etc.), genetic (epistatic interactions), metabolic (enzymes, substrates and products, metabolic fluxes), functional (redundant metabolic paths, redundant genetic elements and gene products), etc. Moreover, in many cases, a given element participates in different types of interactions. For instance: the cloning of a structural gene in a high copy number plasmid and its expression under the control of an inducible promoter modifies the pools of available transcription factors, nucleotides, aminoacids and the structures in which the expressed protein participates. As a consequence of all these interactions, apparently
distant cell processes and functions are affected by the expression of a transgene, or by the change in the culture conditions.

Networks are graphical tools to facilitate the understanding of the complete set of connections than can be established among different molecular and/or functional elements in a cell. Networks facilitate the formulation of hypothesis explaining the cell’s functioning as a whole system. Moreover, according to some authors, the intrinsic properties of the networks can explain some of the phenotypes expressed by complex organisms [1].

Among the main goals of genetic studies, two of the more relevant are to explain how the cell’s or the organism’s genetic information is shaped into its phenotype, and how does evolution proceed. The knowledge of these processes may let us predict the response of the organism to environmental aggressions and how to control this response for biotechnological applications. The ever increasing volume of data produced by high-throughput technologies enables us to shift the focus of study from single genes to complete genomes, transcriptomes, proteomes and metabolomes, to produce more integrated views of the cell structure and function. This approach, however, requires the use of new concepts and tools to deal with highly complex systems. Among the new tools there are new experimental approaches such as the use of bioinformatics for studying and understanding the vast amount of information present in a genome.

The functional study of a new genome starts when a satisfactory assembly of the sequence has been produced. Then, it is necessary to annotate it as accurately and completely as possible to refine gene models, and to determine the structural and regulatory regions in the genome. Genome annotation is done at different levels: (i) nucleotide level to determine the start and stop codons and the donor and acceptor intron splice sites for each gene, (ii) protein level to determine the structure and function of the coded protein and (iii) process level to determine the interactions established between genes and gene products, and to identify the processes where these genes or gene products are involved. Whereas gene annotation at the nucleotide level is highly efficient and several protocols based on sequence similarity and ab initio gene prediction have been developed, protein and process annotations are less developed because of their intrinsic complexity [2, 3]. The prediction of a protein’s structure from its sequence is still not efficient as to deal with completely new protein sequences, the prediction of stable physical interactions among proteins and gene products are also tentative, and the functional interactions between genetic elements and pathways is even more complicated [4]. As a complement of the standard methods, functional networks have been used in some cases as a tool for the annotation of whole genomes [5].

The study of functional genomics depends on the sensitivity of the analysis for detecting subtle phenotype variations (including interactions between gene products) and to correlate them to genomic variations [6]. Phenotypic traits have been classically classified as Mendelian and complex. The last ones are affected by genes (quantitative trait genes, QTGs) at many loci (quantitative trait loci, QTLs), the environment, and the interaction between the QTGs and the environment. Genetic linkage mapping is the method for determining the position of genes on the genome. The mapping of the genes controlling complex traits (QTL mapping) permits to determine the position of the main loci responsible for the trait (major QTLs), and their contribution to the variation of the trait in the population (effect size). The mode by which different genes contribute to the quantitative trait may be different: structural genes, transcriptional regulators, functional regulators, etc. [7]. Consequently, all the genes contributing to the final phenotype are functionally interconnected, and these connections can be depicted as functional networks that reflect the genetic architecture of the complex trait [8].

The high-throughput transcriptome studies have permitted to monitor the quantitative variation of the expression of individual genes in different members of a population. This type of analysis has revealed that gene expression is a complex trait controlled by multiple quantitative loci and by environmental conditions [9]. The mapping of the QTLs responsible of gene expression variability (eQTLs, [10, 11]) permits to identify genome regions acting as master regulators of the expression of many genes, and to establish networks of co-regulated genes [9, 12, 13].

Another approach for studying the functional interactions between genes has been assayed in the model yeast *Saccharomyces cerevisiae*. It consists in the construction of complete series of double mutants and the ulterior identification of pairs of genes whose simultaneous deletion causes lethality (synthetic lethals) or other measurable changes in
parameters such as growth rate [14, 15]. The genes involved in a synthetic lethal pair are functionally, although not necessarily physically, connected. The elements involved in many redundant metabolic and/or genetic pathways can be identified by this type of analysis since their simultaneous deletion can produce a measurable phenotype. The functional networks built up using this type of interactions are complementary to other networks based on quantitative traits and on the molecular interaction between gene products.

The last level of interactions to be considered in this review is that established among gene products within a cell. In a first step, we can consider the physical interactions occurring between pairs of proteins. In a more general analysis, all the physical interactions between the different molecular species (proteins, DNA, RNA, metabolites, etc.) present simultaneously in the cell can be included in the analysis. The comprehensive description of all these interactions forms the cell’s protein interactome [16]. The interactome study permits to construct new networks relating gene products that interact physically in the cell. These are currently the densest networks under study, and provide a more integrated view of the cell function enabling the construction of models to explain complex processes and pathologies [17, 18].

Most of the genome functional studies have been developed in the baker’s yeast *S. cerevisiae* as a model organism. Consequently, a preliminary question arises about the validation of these models in other organisms. In the case of interactome and synthetic lethal studies, the basics of the functional networks have also been observed in other organisms [15, 19]. In the case of the analysis of complex loci, the theoretical models about their architecture are also compatible with results from other organisms [8, 20]. Consequently, the network-based picture of the functional organization of the cell is a robust theoretical framework.

**NETWORKS**

Biological networks are graphical visualizations of the interactions established between different genetic or functional elements within a cell (Box 1). These elements can be genes, transcripts, DNA binding sites, proteins, metabolites or functions, depicted in the graph as nodes or vertices, connected to each other by links or edges representing their functional interactions (e.g. protein–protein interactions, enzymatic reactions, DNA–protein interactions, etc.) [16, 21].
Two nodes connected by a link are considered neighbours (Figure 1). In the case of quantitative data, the relationship between the nodes can be measured via correlation coefficients [22].

A network can be described by some numerical parameters. The most basic characteristic of a node is its $k$-degree, which indicates the number of links it has to other nodes. Different network nodes may have different $k$-values (Figure 1), and the distribution of the $P(k)$-values gives an elementary measure of the network’s topology. Most biological networks’ $P(k)$-values follow a power law, $P(k) \approx k^{-\gamma}$, where $\gamma$-values range between 2 and 3. In this type of networks, a small but significant number of nodes (hubs) are connected to many others, while the large majority of nodes have only one or very few links. Power law topology produces robust networks stable against random perturbations [23]. Hub knockouts have larger phenotypic effects than the knocking of non-hub nodes [24]. The relevant presence of hubs in biological networks points to a modular organization of the cell functions.

A second parameter used to describe and classify the network’s architecture and the relative position of particular nodes is the path length. Path length measures the number of steps that have to be taken to travel from one node to another. The mean path length is a measure of the network’s diameter (Figure 1). Networks with short mean path lengths are classified as small-world, a characteristic of both biological and random networks.

The third parameter for describing a network is the clustering coefficient ($C_i$) that indicates the number of links existing between the neighbours of a given node divided by the maximum number of links possible between them. A high clustering coefficient means that, for example, if node A is connected to nodes B and C, there is a high probability that node B has a direct link to node C or, in other words, that nodes A, B and C form a triangle. Biological networks usually have a clustering coefficient higher than that found in random networks. This is an attribute of regular networks, and suggests a community structure or a modular organization, which is a general property of complex networks. Highly complex networks with high $C_i$ values are difficult to study, and they can be simplified by removing the least connected nodes to reveal the modular backbone of highly connected hubs ($k$-decomposition of the network in which only nodes with a $k$-degree equal or higher to $k$ are drawn) [25].

In addition to the complexity caused by high clustering coefficients, biological networks are contextual since most of the interactions can vary in different environmental or growth conditions. Biological networks that combine the separate layers of information from gene regulatory networks, signal transduction networks and metabolic networks are expected to enhance the understanding of the cell function and dynamics. These types of networks are known as biological context networks [26, 27].

**COMPLEX TRAIT NETWORKS**

The genetic architecture of a trait is defined by the knowledge of the genes controlling it [8].
Complex phenotypic traits are controlled by many genes, and are highly affected by the environmental conditions and the genetic background. The tendency of the genetic background to modify the phenotypic expression of most genes indicates that few (if any) Mendelian traits are truly monogenic and that most of them are genetically complex [20, 28].

Complex quantitative traits have been extensively considered in the design of genetic breeding processes [29, 30] although a definitive model to fully explain their structure is not yet available. As summarized by Farrall [31], the classical structural model for complex traits was set up by Fisher and it described them as controlled by a vast number of genes with minute effects. According to this model, evolution must proceed as a gradual and imperceptible process. However, this infinitesimal model can be discarded as it is possible to map genetically QTLs. This mapping would have been impossible in an infinitesimal scenario, since the statistical power required for the detection of individual genes would have never been reached. The finding that genetically mapped QTLs were responsible for as much as 20% of the trait genetic variation prompted the refinement of Fisher’s model. Two theoretical modifications were made by Kimura and Orr following Robertson’s suggestion that the distribution of allelic effects must be exponential: (i) non-deleterious large mutations can be fixed and (ii) the adaptive space at the beginning of an evolutive process must be larger and allow effects more pronounced than in later evolutive steps. According to this modified exponential model for the control of quantitative traits, there must be a small number of loci with a relative large effect plus large number of genes with progressively smaller effects. An additional prediction of this model is that the oldest allelic variants must show the larger effects. This model, however, is not free of problems since it predicts a continuous exponential distribution of effects whereas in QTL mapping experiments the distribution of gene effects is much coarser.

QTL mapping of a complex trait lets us to determine the number of genome regions controlling it, the distribution of their gene effects and their additive or non-additive action [8, 30, 32]. The final objective of QTL mapping is to identify the quantitative trait nucleotides (QTNs) [33] whose mutation causes quantitative variation in the trait. For mapping a complex trait, its quantitative variation is scored in the members of a segregating population. The segregating populations can derive from crosses of inbred lines, backcrosses or be formed by haploids in those species in which this type of material can be propagated [34]. The positions of the QTLs are determined by correlating the variation of the quantitative trait in the members of the segregating population with the presence in them of specific genomic regions identified by a molecular marker or limited by two adjacent markers. The output of QTL mapping is a graph that represents the likelihood (LOD value) of a positive correlation between a given genomic segment and a significant variation in the quantitative trait (Y-axis), against the chromosome linkage length (X-axis).

QTL analyses using large mapping populations have permitted the identification of genes responsible for as much as half of the population trait’s variability in different biological systems (fungi [35, 36], plants [37, 38], mice [39, 40] and Drosophila [41, 42]). QTLs explaining as little as 5% of the phenotype variance can be detected with samples of a few hundred individuals [43, 44]. The rest of the variation that cannot be attributed to these major QTLs can be due to others with even smaller effect or to epistatic effects that cannot be directly determined. Moreover, QTL mapping based on bi-parental populations reveals only the portion of the genetic trait’s architecture in which both parents differ (i.e. they are heterozygous). For a more comprehensive study of the genetic architecture of complex loci the study of multiple interconnected populations or multiple populations studied in the same environment is necessary [8].

QTN identification requires a further high resolution of the genetic linkage map, since linkage distances of 10–30 centimorgan (cM; acceptable in the initial study) can correspond to large genome regions with multiple genes (10–30 Mb and 100–300 genes in humans). Map resolution can be increased by high-resolution crosses, the use of congenic strains, near isogenic lines and linkage disequilibrium studies [8, 45]. The goal of this work is to reduce the interval of the map to ~1 cM. QTN detection has the additional difficulty that even in a small genome region, there would be neutral sequence variants linked to the functional variant. Assuming the sequence variants observed in humans, a region of 1 Mb corresponding to 1 cM is expected to contain nearly 1000 SNPs that must be individually evaluated for their quantitative effect. The final proof for a QTN is the complementation test [46].
The effect of a QTL can vary in different genetic environments because of gene interactions known as epistasis [45] (Figure 2). The importance and actual effect of epistasis is not fully clear. Some studies suggest that it is more relevant than the additive effect [6, 47, 48], whereas others indicate the opposite [49]. The inclusion of epistasis in QTL mapping complicates it very much since the number of interactions can be huge. The method for studying epistasis is to limit this number either by studying only those involving specific QTLs or by selecting only those with a larger effect. Another alternative is to concentrate the study in very limited genomic regions (around 1 cM).

As an example of QTL analysis of a complex trait, we will summarize the study of the QTLs controlling S. cerevisiae yeast morphology carried out by Nogami et al. [6] using the natural variation produced by crossing two natural populations. The authors mapped QTLs for 67 traits and discovered 364 correlations between their segregation and the inheritance of the expression levels. Additionally, they found that 25% of the morphological characters showed transgressive segregation (see subsequently) and that 40% of them had epistatic interactions.

The results described in this article did not overlap with others previously obtained from yeast systematic deletion strains (see subsequently) indicating that multiple complementary approaches could be necessary for uncovering all the functional relationships occurring within genomes. As one of these complementary approaches, we have studied the genomic regions controlling the variation of enzymatic activities in different individuals of a segregating population [7]. The rationale is that the identification of QTLs controlling variations in enzymatic activity (aQTLs) will reveal not only the structural genes, but also the functional regulators of the activities.

GENE EXPRESSION NETWORKS

The high-throughput techniques used for transcriptome studies assist to profile the quantitative variations in gene expression in the members of a segregating population. This type of analysis has shown that gene expression is a complex trait amenable to QTL analysis. In this case, expression QTLs (eQTLs [10, 11]) reveal the genome regions controlling the expression of a given gene [50, 51]. The study of global gene expression profile in a segregating population is known as Genetical Genomics [52], and it is a powerful tool for finding candidate genes [13, 46, 53, 54]. An advantage of the genetical genomics approach over the classical gene-by-gene studies is that, whereas traits in classic expression studies are selected on the basis of their phenotypic divergence or biological interest, genome-wide expression studies provide data on a large and unbiased set of traits revealing new levels of complexity [13]. There are, however, some limitations to this experimental approach that have not fully resolved yet: (i) eQTLs might produce changes smaller than the current discriminating levels in microarray studies, (ii) many eQTLs show allelic series, (iii) the identification of the correct tissue in which the expression difference occurs is problematic because the environmental conditions vary among tissues and cell types in multicellular organisms and (iv) pleiotropy is a common feature of the genetics of gene expression [13].

Despite these difficulties, evaluations of the quantitative variation of gene expression in S. cerevisiae has revealed that 38% of the genes that show differences of expression in two yeast strains showed quantitative variation in hybrids [31]. In yeast, the effect of the eQTLs explained up to the 27% of the genetic
(heritable) expression variance, and only 23% of the traits had an eQTL that explained more than 50% of the genetic variance. In mice, the mapped eQTLs explain the 25% of expression variance, and in human the effect sizes varies between 27% and 29%. Therefore, even the strongest QTL that underlie variation in gene expression typically explains quarter or less of the variation [13].

The comparison of the distribution of gene expression levels of a segregating population and the corresponding parental strains reveals various types of genetic complexity [13]. Most heritable yeast transcripts show a transgressive segregation (TS) pattern [6, 55] in which values in the segregating population extend significantly beyond the range defined by the progenitor strains. A smaller proportion of traits show directional genetics (DG) [55] with a distribution of trait values in the segregating population that is significantly concentrated within the range defined by the progenitor strains [56]. Finally, it has been shown that non-additive action (NA) is common (roughly 50% of the transcripts) in Arabidopsis, Drosophila and maize, and that the extreme forms of NA interaction (overdominance and underdominance) are not rare [13].

There are two types of eQTLs: cis-eQTLs located in the vicinity of the structural gene, and trans-eQTL corresponding to classical trans-acting factors. This group has a direct role on the widespread phenomenon of pleiotropy [31, 51]. cis-eQTLs usually have individual larger effects on the expression than trans-eQTLs, and most of them had been identified before the use of large-scale genome analyses [51]. A recent study suggested that most of the total quantitative variation in expression corresponded to trans-acting loci [12]. Surprisingly, only a minor portion of them appeared to co-localize with transcription factors. In fact, trans-acting variation appeared to be dispersed across a range of molecular functions, [9, 12]. The genetic elements forming a trans-eQTL are, themselves, controlled by other cis- and trans-eQTLs forming, by this way, complex regulatory networks (Figure 3).

In a study carried out using Arabidopsis thaliana, West et al. [57] found that the majority of genes exhibited heritable transcriptional variation that is controlled by eQTL. This analysis revealed more than 36,000 eQTLs, and up to 11 eQTLs detected per transcript. One-third of the transcripts had a cis-eQTL, whereas the remaining were trans-eQTLs that, in this case, were no uniformly distributed across the chromosomes but concentrated in hotspots that regulated hundreds to thousands of e-trait. In a similar study, Keurentjes et al. [58] studied the transcription level of 24,065 genes and found 4523 eQTLs detected for 4066 genes. The differences between these two
studies may be due to the smaller number of replicates used in the last one that might have led to an underestimation of the number of eQTLs. Notwithstanding, this second study also found clusters of co-regulated genes that might correspond to master regulators genes controlling the expression of many other genes. The occurrence of regulatory hotspots has also been confirmed in other studies [59, 60], and it makes it possible to construct genetic networks of co-regulated genes whose expression is correlated [22] (Figure 3). A question arising from these studies is whether co-regulated genes involved in the same network belong to the same biological process or metabolic pathway [22, 51].

SYNTHETIC LETHAL NETWORKS

A different approach for studying the functional interactions among genetic elements is based on the detection of pairs of genes whose individual deletion can be tolerated but that cause lethality when simultaneously deleted [15, 61]. A remarkable observation is that the systematic deletion of the genes of *S. cerevisiae* (yeast gene-deletion set [62]) has revealed that out of ~6000 heterozygous diploid strains, each of which is deleted for a single copy of a specific gene, only a few of them (~3%) were haplo-insufficient whereas the others grew normally as diploids on a rich medium [63]. Only ~1000 genes were found to be essential for viability in haploids growing in standard lab conditions [64, 65]. This defines a set of 1000 essential genes (complementarily, there are ~5000 mutants viable in haploid status) [15]. The mating of two viable deleted haploids and the subsequent tetrad dissection permits the production of new deletion combinations some of which are tolerated, whereas others are lethal (synthetic lethal interaction, SLI) revealing a functional relationship between the two deleted genes. The use of yeast-gene deletion sets permits the systematic study of SLIs using high-throughput technologies such as synthetic genetic arrays (SGAs) [66, 67] in which a single deletion strain is used as query against the whole deletion mutant collection, and the analysis of synthetic dosage lethality (SDL) [68] in which the search is for genes whose overexpression is lethal in a genetic background defined by a single deletion.

The systematic study of SLIs reveals synthetic genetic interaction networks (SGINs) that interconnect genes whose products operate in functionally related pathways (Figure 4). The SGINs have the characteristics of ‘small world’ networks in which highly connected hub genes that play a central role in the fitness of the organism, are densely connected in a local neighbourhood. An additional characteristic of these networks is the deep intrinsic buffering of cellular functions through redundant or overlapping pathways [15]. The synthetic genetic interactions were significantly more abundant

![Figure 4: Synthetic lethal networks. Thirteen different deletion haploids are mated among them in all combinations and examples of the different behaviours of the offspring are shown: grey dots indicate the lack of a differential phenotype, white dots indicate a reduction in the growth rate (aggravating interaction), black dots indicate an increase in the growth rate (alleviating interaction) and dashes indicate a lethal interaction. These interactions are used to construct the synthetic lethal network in which their different types are indicated.](http://bfg.oxfordjournals.org/)

Pisabarro et al.
between genes with the same mutant phenotype, between genes encoding proteins with the same subcellular localization, and between genes encoding proteins within the same protein complex [69].

Synthetic methodologies allow a quantitative assessment of the relative fitness of double-mutant meiotic progeny. This type of studies permits a quantitative mapping of epistatic relationships. The genetic epistatic interactions can be aggravating, reflecting the activity of genes in two separate but compensatory pathways, and alleviating, when genes work in the same pathway and share alleviating interactions with each other [15]. The incorporation of this information into the SGIN adds an additional layer of complexity to the network (Figure 4).

An important question arising at this point is about the validation of the SLIs and SGINs in other biological systems for which there are no gene deletion sets. The use of RNAi libraries to target all predicted genes permits to explore this question in many biological processes in Caenorhabditis elegans, fly and mammalian cell lines [70, 71]. Despite its currently modest size, the C. elegans genetic network recapitulates the topology of genetic networks in yeast [72], suggesting that a general network structure is conserved in eukaryotes. For instance, a study of genetic interactions among signalling and transcriptional regulators in Caenorhabditis elegans has uncovered ∼350 interactions [73], and a study of the control of spindle assembly in this nematode also revealed similar sites of interactions to those described for S. cerevisiae [74].

Finally, it is important to stress that despite only 1000 genes have been found essential for viability in yeast haploid cells, Boone et al. [15] estimate that there are ∼200 000 digenic lethal combinations. Thus, using combinations of drugs, we can find a 200-fold wider repertory of drugs that work in a way that exploits a fundamental weakness of cellular networks.

**INTERACTOME NETWORKS**

The finding that organisms with very different complexity levels (i.e. humans, Caenorhabditis, Drosophila) have a similar number of genes, has suggested that it is not the number itself but the complexity of the interactions between gene products one of the basis of such complexity differences. Besides that, the one-gene/one-protein at-a-time approach of the last 30 years has provided some indication of function for only 5–10% of all predicted proteins so far [75]. The construction of interactome networks is an attempt to explain the organism’s complexity on the basis of the interactions among its molecular constituents.

The aim of interactome studies is to identify all the physical interactions established among proteins, nucleic acids and other molecules within a cell [76]. Interactome networks have been studied in plants [25], fungi [77], animals [78], bacteria [79], virus [80–83], cell organelles such as the mitochondria [84] and pathological conditions [85]. The goal is to generate comprehensive network maps in which the nodes (proteins, RNAs, DNA binding sites or metabolites) are connected by edges corresponding to molecular interactions (protein–protein interactions, enzymatic reactions, DNA–protein interactions, etc.). As biological systems are very fluid, the interactions between the edges can be formed and disappear as consequence of changes in the cell’s environmental conditions (see earlier). The systematic study of these changes permits a deeper understanding of the molecular biology of the cellular or organismal processes.

The construction of interactome networks requires two technical activities: the isolation of all the open read frames (ORFome) translated in the cell, and the detection of the physical interactions between proteins [86–88]. The construction of the ORFome is aimed at avoiding the sampling bias caused by differences in the transcription levels of different genes, and to consider all the splicing variants present in the cell. The systematic study of protein interactions has been predominantly based on the Yeast-two-Hybrid (Y2H) approach [89] that consists in the construction of two recombinant proteins, one of them fused to the DNA binding domain (BD) of a transcriptional activator protein (yeast Gal4 or bacterial LexA) called ‘bait’, and the second fused to the transcriptional activation domain (AD) called ‘prey’ (Figure 5). When bait and prey proteins interact, the transcription activator is reconstructed and the expression of a reported gene denotes the physical interaction between bait and prey.

This system has two major advantages: DNA is manipulated to study both bait and prey [90] so ORFome resources are readily employed, and the system is readily adapted to high-throughput. Once protein interaction complexes are initially detected, they can be confirmed by other techniques such as the physical purification of the complexes...
by co-affinity (co-AP) and subsequent identification of its members by mass spectrometry [91]. In the case of humans and Caenorhabditis, up to 80% of the interactions detected by the Y2H system could be validated by co-AP [86, 92].

The interactome maps produced by the current technologies include only a sample of all the interactions actually occurring in a cell because only those that can be established by bait and prey in the nuclear environment are susceptible for this type of analysis. This limitation impedes the study of interactions involving proteins targeted to specific cellular localizations such as membranes, and limits the study of interactions established when at least one of the members has been post-translationally modified. In consequence, it is convenient to complement the interactome studies with others such as the construction of genetic interaction networks discussed earlier. The first interactome networks constructed in S. cerevisiae used a collection of ~6000 bait proteins and identified around 3000 interactions. This number is considered by the authors as corresponding to only the 10–15% of the total number of interactions within the cell [93, 94]. Moreover, the interactomes detected in both studies showed <15% overlap, suggesting that they represented only a limited portion of the total interaction landscape and a high signal-to-noise ratio.

Some general rules can be deduced from the interactome studies carried so far. The interactome

---

**Figure 5:** Interactome networks. These networks are built up using the information provided by Y2H analyses in which different proteins (baits) are fused to the DNA BD and a second protein collection (preys) are fused to the activating domain of a transcription factor. Whenever there is a physical interaction between baits and preys, the transcription factor promotes the expression of a reporter gene. The results of this type of analysis are shown in the panel where the black dots indicate the expression of the reporter gene. The data of this panel are used to construct the interactome network depicted on the right.
networks show a power law degree distribution of \( k \)-values. This means that most proteins interact with few partners, whereas a few hub proteins are central in protein interaction clusters [16, 95, 96]. As it has been discussed above, the power law topology of interactome maps suggest that knocking out hub proteins would be more likely to produce lethality than knocking less connected nodes. In the interactome networks, hubs very likely represent essential evolutionary conserved proteins central in the dynamic organization of cellular properties [25, 97, 98].

The conservation of interactome networks in different organisms (i.e. the power of an interactome constructed for a model organism to predict the interactions established by the corresponding proteins in another system) has been investigated by testing the degree in which the yeast interactome can be used to predict the interactome network in *Caenorhabditis* [99–101], and it has been found that \( \sim 20\% \) of the interactions predicted to occur in *C. elegans* on the basis of proven interactions between their yeast orthologues, were actually detected [19, 102]. Moreover, protein–protein interactions can be computationally predicted from co-evolution events [103, 104]. These results indicate that protein interaction networks represent a basic functional organization of the cell that appears to be conserved across evolution.

**PHENOTYPE PREDICTION USING GENETIC NETWORKS**

The integrated study of the different types of genetic networks discussed above, and their contextual modulation provide a new integrated picture of the cell’s biology. However, how can this information be used to predict the phenotype of a cell or an organism? The identification of the causal links between genotype and phenotype is a central challenge that is approached from different perspectives. In a first step, genes involved in different biological responses were identified using combination of gene perturbation and high-throughput phenotype assays [105–108]. The linkage between individual genes and phenotypes had also been established by QTL analysis [109]. However, the complexity of phenotypic traits makes it necessary to take into account simultaneously many genes and their network interactions. In this case, the use of networks make it possible to functionally classify genes and to identify probable candidate genes [1] and gene pairs involved in genetic interactions [110] responsible for some phenotypes.

A more predictive approach has been recently published by Carter *et al.* [108] who developed a mathematical model of genetic interaction based on the classical approach of observing how genetic perturbations interact to affect phenotypes revealing functional relationships and pathway order. The authors analysed the genetic interactions as a quantitative influence and used these results to direct the integration of molecular (physical) interaction data. The influences could be positive or negative and modify a fraction of the measurable phenotype (e.g. the expression of a gene) inferred to be caused by a genetic element in the system (e.g. a regulatory protein). The measured phenotype is modelled by multiple influences acting throughout the inferred network. By this method, they succeeded in predicting quantitative gene expression profiles and precise filamentation phenotypic effects in *S. cerevisiae*. A different strategy has been used by Kaleta *et al.* [111] who used a model based on the chemical organization theory [112] to correctly predict the known growth phenotypes of *Escherichia coli* on 16 different substrates and the lethality of knockout experiments in 101 out of 116 cases. Finally, Chen *et al.* [1] used the analysis of eQTL profiles to determine gene networks perturbed by susceptibility loci that lead to disease in mice. By this way, they identified liver macrophage populations involved in the development of a metabolic syndrome.

**CONCLUSION**

The study of the complex biologic systems using the information provided by high-throughput analytical technologies requires new conceptual tools for producing integrated pictures of the whole functional system. Various approaches have revealed a cluster organization of cellular and functional elements. Moreover, these clusters are organized in networks that share common general characteristics and topology irrespective of the nature of their nodes and edges, and the essential of these networks appears to be conserved across evolutionary branches. Consequently, this area of study will very likely produce new exciting results in the near future that will open the door to a more integrated understanding of the cell’s biology [113–115].
Key Points

- The integrated study of cell functions and processes is an unavoidable challenge in molecular biology and biotechnology. This study is fuelled by the flow of information produced by different high-throughput analytics technologies.
- The construction of genetic networks is an efficient approach for the integrated study of the cell’s molecular biology. These networks are graphic representations of genetic elements connected by the physical and/or functional interactions established among them.
- The genetic elements used in the networks can be, among others, genes controlling complex traits (QTLs controlling phenotypes, gene expression, enzymatic activities), gene pairs of synthetic lethals or proteins and other gene products within a cell.
- The genetic networks have a common topology where a small number of genetic elements are highly connected to many others (hubs), whereas most of them have a small number of connections. This type of network is very stable against random perturbations reflecting the inherent stability of biological systems.
- The biological networks observed share common characteristics irrespective of the nature of their nodes, and are essentially conserved across evolutionary branches.

Acknowledgements

This work has been supported by funds of the AGL2005-08005-C02-01 and GEN2006-27843-E, and by institutional support from the Public University of Navarre.

References


