The origins and evolution of functional modules: lessons from protein complexes

Jose B. Pereira-Leal*, Emmanuel D. Levy and Sarah A. Teichmann

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Modularity is an attribute of a system that can be decomposed into a set of cohesive entities that are loosely coupled. Many cellular networks can be decomposed into functional modules—each functionally separable from the other modules. The protein complexes in physical protein interaction networks are a good example of this, and here we focus on their origins and evolution. We investigate the emergence of protein complexes and physical interactions between proteins by duplication, and review other mechanisms. We dissect the dataset of protein complexes of known three-dimensional structure, and show that roughly 90% of these complexes contain contacts between identical proteins within the same complex. Proteins that are shared across different complexes occur frequently, and they tend to be essential genes more often than members of a single protein complex. We also provide a perspective on the evolutionary mechanisms driving the growth of other modular cellular networks such as transcriptional regulatory and metabolic networks.

Keywords: functional module; modularity; protein complex; evolution; duplication; protein interaction network

1. INTRODUCTION

Modularity is a widespread concept in computer science, cognitive science (Cohen & Tong 2001), organization theory (Langlois 2002) and other scientific and technological fields. It allows a complex system or task to be broken down into smaller, simpler functions. At the same time, the individual modules can be modified or operated on independently. Thus, changes in one part of a system should not affect other parts.

In biology, the concept of modularity has a long history (see Winther (2001) for a recent review). In comparative anatomy, structural modules representing the parts of an organism, usually at the adult stage, have been discussed since Cuvier & Saint-Hilaire in the late eighteenth century. Concomitantly embryologists were recognizing developmental modules as parts that change over embryonic time. In the 1930s, Needham postulated that development consists of distinct processes that while operating in coordination, can be dissociated into separate elements (Needham 1933). He proposed that these can evolve separately from each other, thereby laying the foundation for the present-day study of the evolution of development.

During the development of multi-cellular organisms, cells differentiate into separate cell lineages, which make up different developmental modules, such as the endo-, ecto- and mesoderme. Each cell itself can be viewed as a module at a lower level of organization of the organism. These cell ‘modules’ come in different flavours, as illustrated by the lymphocytes (round green) and dendritic cells (grey) in panel (c) of figure 1.

A few years ago, it was proposed that biological processes within individual cells are modular (Hartwell *et al.* 1999). These modules were called ‘functional modules’. They are discrete entities whose function are separable from those of other modules, and are composed of many types of molecules whose interactions underlie the function of the module (Hartwell *et al.* 1999). The availability of complete genome sequences for more than 200 organisms, as well as a variety of genome-scale datasets of functional genomics and proteomics information has revealed the prevalent modular nature of cellular systems. For example, protein–protein interaction (Rives & Galitski 2003; Spirin & Mirny 2003; Pereira-Leal *et al.* 2004), gene-regulatory (Ihmels *et al.* 2002; Segal *et al.* 2003) and metabolic networks (Ravasz *et al.* 2002) have all been shown to display a modular organization, in which the modules correspond to discrete functional units. This level of modularity within cells is illustrated by the protein complex, signal transduction pathway and metabolic pathway in panel (b) of figure 1. Here, we will be primarily concerned with the origins and evolution of protein complexes as cellular modules.

Modularity is an established notion even at the level of the individual subunits of a protein complex. Proteins are composed of structural domains, which are in most cases autonomous folding units. These modules are the building blocks of proteins, which can be reused and combined in different ways in evolution (Levitt & Chothia 1976; Chothia 1992; Vogel *et al.* 2005). The modular organization of protein structure facilitates the combinatorial generation of complexity as well as structural and functional diversity. It has further been claimed that these domains are themselves modular (reviewed in Soding & Lupas (2003)).

This brief review of the concept of modularity at different scales shows that it is an important principle in

* Author for correspondence (jleal@mrc-lmb.cam.ac.uk).

One contribution of 15 to a Discussion Meeting Issue ‘Bioinformatics: from molecules to systems’.

Figure 1. Hierarchical modularity in biology. (a) Modularity at the protein level—proteins consist of modules formed by domains. The *S. cerevisiae* uridylylate kinase [lukz] contains a single *P*-loop containing nucleotide triphosphate hydrolase domain as defined in the SCOP database (Murzin *et al.* 1995; Andreeva *et al.* 2004). A domain from the same superfamily is also found in the multi-domain protein EF-TU from *T. thermophilus* [1exm], also shown as a red cartoon. The two other domains belong to the SCOP superfamilies *Translation proteins* (pink) and *EF-TU/EF-1alpha/EF2-gamma C-terminal domain* (violet). (b) Modularity at the cellular level—most proteins work in a cooperative manner with other proteins and form functional modules. Here we show three distinct types of functional modules: (i) a protein complex—the *B. taurus* ATP synthase [1e79]. Here six protein chains (three red and three orange) that all contain the *P*-loop containing nucleotide triphosphate hydrolase domain (cartoon representation) as well as several other domains, assemble and form a ring; (ii) a signalling pathway—mating response MAPK pathway in yeast (Schwartz & Madhani 2004); (iii) a metabolic pathway—mevalonate pathway. (c) The diversity of cellular types is a consequence of the distinct arrangement of modules at lower levels, and cells are themselves modules from which tissues are built, such as the dendritic cells (grey) and lymphocytes (green) shown here (microscopy image courtesy of Milka Sarri).

biology. Here, we will focus on the intermediate biological scale of cellular networks, with an emphasis on the protein interaction network. We first discuss the definitions of modules in different cellular networks, and the extent to which it is possible to propose objective criteria defining a functional module. We then

focus on protein complexes, modules in the protein–protein interaction networks, to discuss the origins and evolution of modularity in cellular networks. Finally, we turn our attention to how evolution at the module level contributes to the evolution of the network as a whole.

2. DEFINING MODULES IN CELLULAR NETWORKS

Cellular networks can be partitioned into functional modules, which accomplish discrete biological functions in isolation from other modules in the networks (Hartwell *et al.* 1999). Isolation in this context can signify spatial, chemical and temporal separation. In this section, we will discuss how different ways of partitioning cellular processes relate to the general definition of modularity in terms of isolated functional entities.

In the physical protein interaction network, protein complexes are an obvious example of entities that are spatially and chemically isolated. Stable protein complexes are assemblies of proteins, which form many interactions with each other and therefore are cohesive and strongly connected to each other in the context of the larger protein interaction network. Furthermore, stable protein complexes can frequently be reconstituted in a functional form independently of the rest of the protein interaction network. Thus, they are clearly functional modules.

In stable complexes, protein interfaces are typically large, with buried surface areas frequently larger than 2500 Å², involving long-lived hydrophobic contacts (Janin *et al.* 1988; Janin & Chothia 1990). One example is the F1Fo ATP synthase shown in figure 1. However, not all protein complexes are stable—many assemble and disassemble on far shorter time-scales (less than 1 s). These interactions between proteins involve smaller interfaces, burying surface areas typically less than 2000 Å² and hydrophilic contacts (Lo Conte *et al.* 1999; Chakrabarti & Janin 2002). A recent study on the dynamics of protein interactions during the budding yeast cell cycle revealed many transient complexes, which assemble at specific stages during the cycle (de Lichtenberg *et al.* 2005). Well-known examples are dimers between cyclin-dependent kinases and cyclins, the mitotic checkpoint complex (Lew & Burke 2003) and the chromosomal passenger complex (Vagnarelli & Earnshaw 2004). In these examples, the individual proteins coalesce into transient complexes. Thus, functional modules may be established by strong and weak physical interactions between their components.

Protein complexes, stable and transient, display strong interactions within the complex and weak interactions to components outside the complex. This is analogous to the definition of modularity in the context of evolutionary developmental biology, in which modules should display strong connections within and weak connections outside the module (Winther 2001). This is perhaps the most well-known characteristic of functional modules, underlying a variety of algorithms for the detection of modules in cellular networks, such as of transcriptional clusters (Brazma & Vilo 2000).

A related notion is that there should be more connections among the module's components than outside the module. This is the same as stating that a functional module corresponds to a clique in the network (Watts & Strogatz 1998). This principle has also been successfully used to identify functional modules in cellular networks from a variety of data types, for example from protein–protein interaction

data (Bader & Hogue 2003; Rives & Galitski 2003; Spirin & Mirny 2003; Gagneur *et al.* 2004; King *et al.* 2004; Pereira-Leal *et al.* 2004). In many cases there is an overlap between the two concepts. For example, the F1Fo ATP synthase complex was recovered from a protein–protein interaction network on the basis of there being more connections within the complex than to the outside (Pereira-Leal *et al.* 2004), but this is also a stable complex with high affinity interactions between its subunits (Stock *et al.* 2000).

Not all functional modules require physical interactions of all the components at the same point in time. A signal transduction cascade can be seen as an information-processing module involving a succession of interactions, where the separation is accomplished by the specificity of the interactions, rather than that all components co-localize in time and space. One example, shown in panel (b) of figure 1, is the yeast MAPK pathway controlling mating response (Schwartz & Madhani 2004).

Furthermore, it should be noted that functional modules do not require a physical interaction between components. For example in a metabolic pathway, connections are established by the product of one enzyme being the substrate of the next enzyme in the cascade, independently of any physical interactions between enzymes. An example of such a pathway is shown in panel (b) of figure 1. Hence, in order to define functional modules in cellular networks, we need to consider not only direct physical interactions but also indirect connections such as metabolic interactions, gene regulatory interactions and so forth. Thus, a functional module does not have to be a structure defined in time and in space, like a protein complex, but may also be conceptualized as a process (Schlosser 2004).

There are many anecdotal examples of protein complexes, metabolic and signalling pathways and sets of co-expressed genes that are conserved in evolution. These functional modules can also be evolutionary modules. The success of methods like phylogenetic profiling at identifying functionally associated genes (Pellegrini *et al.* 1999) and clusters of genes that appear to be functional modules (Snel *et al.* 2002; von Mering *et al.* 2003) suggests that this is a good criterion for modularity at the cellular level. However, it was recently shown that not all types of putative modules display the same degree of evolutionary conservation (Snel & Huynen 2004). Therefore, functional modules are not necessarily evolutionary modules in all cases. We will return to this point later when discussing the evolution of functional modules.

In summary, there are distinct types of functional modules in cellular networks. They are defined by the nature of their components and by the types of interaction established within the module (Hartwell *et al.* 1999). From these we can derive a variety of criteria for defining functional modules in cellular networks. Many of these criteria overlap. For example a conserved, stable protein complex displaying many interactions between subunits conforms to three distinct criteria. As a rule of thumb, 'the more criteria a ... unit fulfils, the more justified we are in deeming it a module' (Winther 2001). Any one criterion in

Figure 2. DNA-directed RNA polymerase in the three major branches of the tree of life. The bacterial RNA polymerase structure is from *T. aquaticus* [1iw7]. The eukaryotic RNA polymerase is from *S. cerevisiae* [1i50]. No structure is available for the archaeal RNA polymerase, but its subunit composition is known to be similar to the eukaryotic enzyme (Bell & Jackson 2001), so it is displayed as a transparent version of the eukaryotic protein complex. The three RNA polymerases exhibit striking structural and functional similarities (Woychik & Reinberg 2001). Structurally and functionally equivalent protein chains are shown in the same colour.

isolation will fail to encompass all types of functional modules, as each criterion only captures one aspect of cellular organization.

In this text, we focus on protein complexes as modules in the protein interaction network. Their subunits have more connections and/or are more strongly connected to other subunits than to proteins outside the module, and they form individual structures at a given point in space and time. They are frequently evolutionary modules (Snel & Huynen 2004) existing in other cells and organisms, and many can be reconstituted *in vitro* and can be shown to perform the same function outside the native cell. Thus, of all the types of modules discussed above, protein complexes are the most clearly defined functional modules in cellular networks.

3. THE ORIGINS OF MODULARITY: PHYSICAL PROTEIN INTERACTIONS AND PROTEIN COMPLEXES

Most proteins establish physical interactions with other proteins and many form oligomeric structures (Marianayagam *et al.* 2004). These complexes achieve functions that transcend the sum of the isolated subunits, thus illustrating one of the many positive consequences of modularity. Furthermore, complex formation can provide robustness against mutation and chemical attack (Hartwell *et al.* 1999), as well as propitiate evolvability, i.e. the organism's ability to generate heritable, selectable phenotypic variation (Kirschner & Gerhart 1998; Hansen 2003). But how did modularity emerge in evolution? The evolutionary mechanisms driving the emergence of functional modules, and of modularity in cellular networks are poorly understood. In this section, we will discuss the evolutionary mechanisms driving the formation of protein complexes, and attempt to contextualize such mechanisms into the more general questions of the origins of cellular modularity. We will divide the problem into three questions as postulated by Winther (2001).

First, is modularity in cellular networks an ancestral trait, or was it acquired later, i.e. is a derived character? The genome of Last Universal Common Ancestor (LUCA) is believed to have already coded for the

Table 1. Yeast proteins in complexes. ($p(C)$ is the probability of a protein to be part of a complex. $p(C|E)$ and $p(C|NE)$ represent the conditional probabilities of a protein being part of a complex given that the protein is essential (E) or non-essential (NE). The datasets of protein complexes in yeast are the manually curated complexes from MIPS database (Mewes *et al.* 2002), and the results of two large-scale protein complex purification experiments followed by mass spectrometric identification of proteins: TAP (Gavin *et al.* 2002), HMS-PCI (Ho *et al.* 2002).)

	$p(C)$	$p(C E)$	$p(C NE)$
MIPS	0.23	0.44	0.18
TAP	0.29	0.58	0.21
HMS-PCI	0.35	0.51	0.31

protein complexes that represent the core transcriptional and translational apparatus (Kyrpides *et al.* 1999; Makarova *et al.* 1999). One example is the RNA polymerase, a protein complex that exists in several forms in all known organisms (figure 2). This and other anecdotal examples suggest that protein complexes were established very early in evolution, which is consistent with modularity as an ancestral trait. Can we generalize this? Analysis of the phylogenetic extent of all yeast proteins showed that essential proteins are the most widely conserved in different branches of the tree of life and therefore the most ancient (Pereira-Leal *et al.* 2005). At the same time, essential proteins tend to be subunits of protein complexes (table 1), suggesting that protein complexes were formed early in evolution. Thus, although we are limited by the resolution of the methods available to us, protein complexes are frequently conserved in all domains of life. This suggests that modularity is an ancestral feature of biological systems. However, it was shown that evolutionary conservation increases from monomeric proteins to members of transient complexes and finally to components of stable protein complexes (Teichmann 2002). This general trend suggests that these conservation patterns are linked to the constraints imposed by the protein-protein interfaces rather than being purely a signal of phylogenetic distribution.

The second question is whether modularity in cellular networks emerged as the result of integrating

disconnected parts, or whether it is the result of parcellation of an integrated whole. Division of a long gene or a multi-functional protein complex into many separate genes might be favourable in order to limit the function of each gene to one or a few cellular processes, which would limit the effect of damaging mutations. This is also termed differential suppression of pleiotropic effects of genes involved in different cellular processes (Wagner 1995). At the single protein level we see that both routes are used in evolution—both gene fusion and fission are used to generate new proteins (Snel *et al.* 2000). However, fusion is about four times more prevalent than fission (Snel *et al.* 2000; Kummerfeld & Teichmann 2005). In either case, the individual component proteins often physically interact (Huynen *et al.* 2000), which is the basis of the Rosetta stone method for predicting protein–protein interactions (Enright *et al.* 1999; Marcotte *et al.* 1999).

Besides the heteromeric interaction of separate proteins that have evolved by fusion and fission, homomeric complexes are highly abundant. We can only accurately and simultaneously quantify the stoichiometry and type of interaction of proteins in the datasets of protein complexes of known three-dimensional structure, which may represent a biased view of the protein universe. As shown in figure 3, we have dissected the Protein Quaternary Structure (PQS) database (Henrick & Thornton 1998) into purely homomeric, mixed and purely heteromeric complexes. To do so we prepared a non-redundant set of protein complexes by considering complexes as graphs where the nodes are the protein subunits and lines are the contacts between these subunits, as illustrated by the cartoons in figure 3. Two complexes are considered identical if they have the same number of subunits from the same protein families as well as the same pattern of contacts between subunits. In our non-redundant dataset, we found that about 90% of the complexes contain interactions between identical proteins.

The prevalence of homo-oligomers is undoubtedly due to the advantages conferred by this type of complex: improved stability, allosteric regulation, multivalent binding, control over accessibility and specificity of binding sites as well as increased complexity (Goodsell & Olson 2000; Marianayagam *et al.* 2004). This is accomplished at the same time as reducing genome size and introducing increased error control during synthesis (Goodsell & Olson 2000; Marianayagam *et al.* 2004). Protein interfaces of homomeric complexes are mainly hydrophobic (Bahadur *et al.* 2003, 2004), making the creation of such structures simple and optimizable. The latter property stems from the fact that in a hydrophobic, homomeric interface, one single mutation towards hydrophobicity is advantageous on both sides of the interface. Furthermore, domain swapping may provide an alternative and/or complementary simple and frequent route for the establishment of homomeric interactions (Liu *et al.* 1998; Liu & Eisenberg 2002). In summary, although parcellation of large peptides into interacting subunits of complexes is observed in the generation of hetero-oligomers, it appears that self-interactions lie at the heart of many stable protein complexes. Their establishment is

Figure 3. Widespread presence of homomeric interactions in protein complexes. Homomers, protein complexes formed from multiple copies of the same protein, represent the vast majority (69.8%) of the protein complexes of known three-dimensional structure, as defined in the PQS database (Henrick & Thornton 1998). Of the 30.2% complexes that are not pure homomers, 20% contain at least one interaction between identical chains. Only about 10% of the complexes do not have any homomeric interactions. Three structures exemplify each category. From left to right: the carbonic anhydrase from *M. thermophila* [1qrf] is a homotrimer, the *P. putida* 2-oxoisovalerate dehydrogenase [1ps0] is a heterotetramer of two homodimers and the cap-binding protein from *H. sapiens* [1n52] is a heterodimer. The interactions in each complex are shown as red (homomeric) or green (heteromeric) lines between the nodes in a two-dimensional graph representation of each complex, where nodes of the same shape are identical proteins. The set of protein complexes was filtered according to this graph representation of complexes, so that identical or similar complex structures were not counted multiple times.

sufficiently easy and advantageous to represent a preferred evolutionary route.

The third and final question is whether modularity is a result of the basic mechanisms of network growth, i.e. the result of self-organization by neutral evolution, or instead is a result of selection shaping the structure of the network (Wagner 1995). The evolution of cellular networks has been the subject of much recent interest, in particular regarding the origins of characteristics of the architecture of networks such as scale-free topologies and small-world character. A scale-free network displays a power law distribution of node degree, which translates as most nodes establish few connections, whereas a small number of nodes, the hubs, display a very large number of connections. The small-world character, sometimes described as ‘six degrees of separation’, describes networks with high clustering and short average path-lengths, which is to say that starting from any one node in the network, we can reach any other node in a small number of steps by following the connections in that network (Strogatz 2001; Barabasi & Oltvai 2004). The scale-free structure of the network conveys robustness (Jeong *et al.* 2000),

whereas the speed of perturbation spread and reduced response time are associated with the small-world character.

These advantageous traits may be under selection (Fell & Wagner 2000; Guelzim *et al.* 2002), but theoretical modelling approaches have revealed that networks with topologies comparable to real world networks can be evolved by simulations based on cycles of network growth using very simple principles. For example, if new nodes added to a network are preferentially attached to the most connected nodes, scale-free networks emerge (Barabasi & Albert 1999). However, the biological relevance of this model is unclear (Qin *et al.* 2003; Kunin *et al.* 2004; Pereira-Leal *et al.* 2005). A more biologically reasonable approach considers that networks grow by duplication of nodes, followed by divergence of a node's interactions relative to the interactions of its ancestor (Pastor-Satorras *et al.* 2003; Vazquez *et al.* 2003; van Noort *et al.* 2004). This represents modelling the growth of a network by gene duplication, followed by conservation and then divergence of the interactions of a protein. This does appear to be the principle by which actual protein interaction networks evolve, since at least 40% of the protein interactions in yeast result from duplication with conservation of protein interactions (Pereira-Leal & Teichmann 2005).

Using theoretical approaches, simulated networks with parameters comparable to those of real-world cellular networks have been obtained. In particular, these networks are characterized by a clustering coefficient that is higher than expected in randomized networks. The clustering coefficient is a parameter that captures the cliquishness, or local clustering of a network (Watts & Strogatz 1998). As we saw above, functional modularity can be defined by densely connected sub-graphs, i.e. cliques in the network. A highly clustered network is a network that is rich in these types of modules. Some transcription factor families exemplify this clique formation by evolution through duplication and divergence. A recent case study of several transcription factor families suggests that the duplication of an ancestral homomeric protein underlies the creation of cliques that could be equated with modules in protein interaction networks (Amoutzias *et al.* 2004, 2005). Thus, at least one type of modularity can be a side effect, a simple consequence of gene duplication. For this type of modularity we do not have to invoke selection as a driving force.

How generic are these conclusions about protein complexes and the protein interaction network to other cellular networks? The genome of the LUCA is believed to include most metabolic pathways of central metabolism, suggesting that metabolic networks, as far as the resolution of our current methods takes us, were modular (Kyrpides *et al.* 1999; Makarova *et al.* 1999). However, in these networks, duplication is rarely accompanied by conservation of interactions (Teichmann *et al.* 2001), suggesting that although modularity is an ancestral character of metabolic networks, it is not achieved by duplication with conservation of interactions. In gene regulatory networks, we find that many of the protein complexes are conserved, namely the polymerases and general

transcriptional regulatory complexes (Coulson *et al.* 2001; Coulson & Ouzounis 2003; Babu *et al.* 2004). In many cases, duplication followed by divergence of DNA specificity was used in evolution to expand the repertoires of transcription-associated proteins (Babu *et al.* 2004; Teichmann & Babu 2004). In §4, we will discuss functional divergence of protein complexes in more detail.

4. MECHANISMS OF SPECIALIZATION OF MODULES

Above we discussed protein complexes as an ancestral modular characteristic of protein interaction networks, and the different ways of evolving new protein interactions, including self-interactions in homomeric complexes and gene duplication. Now we ask, once the complexes are established, how do they evolve? In the field of artificial intelligence, theoretical simulations in neural networks have shown that duplication and specialization of entire modules is an effective mode of network growth (Calabretta *et al.* 1998, 2000). In biology duplication, like modularity, is observed at different levels of complexity. For example in genomes, duplication of individual genes is a major mechanism of evolution (Teichmann *et al.* 1998), and is most frequently accompanied by functional specialization (Lynch & Conery 2000; Prince & Pickett 2002; Lynch & Katju 2004).

Above, we mentioned that gene duplication followed by conservation of protein interactions has contributed about 40% of the interactions of the yeast protein interaction network (Pereira-Leal & Teichmann 2005). In the same work, we investigated whether protein complexes as a whole also duplicate, and the functional consequences of such duplications. We developed an algorithm to detect duplication of modules, which is based on pairwise matching of components of different modules. This algorithm uses structural domain assignments and protein sequence similarity to determine whether two subunits of complexes are homologous. We then analysed three datasets of protein complexes in *Saccharomyces cerevisiae*, and showed that even with very conservative criteria, we detect between 7 and 20% of duplicated complexes in all three datasets, as shown in figure 4. In most cases, duplicate complexes share some protein components in addition to duplicated proteins. We term this scenario 'partial duplication' of complexes, while 'complete duplication', in which all subunits are duplicates, are rarer.

We then investigated the functional consequences of these duplications and observed functional specialization in all studied cases. This means that the core catalytic activity or ligand-binding activity of homologous complexes is retained, while the substrate specificity changes. The conservation of the core function appears to be more conserved among duplicate complexes than among families of individual proteins, though this is hard to quantify. The adaptin (AP) complexes shown in figure 4 are a good example of this functional specialization of duplicate complexes: all are adaptor complexes which mediate binding to coat proteins in the context of vesicular trafficking pathways, but each complex acts on distinct sets of

Figure 4. Duplication of protein complexes. The *S. cerevisiae* adaptin complexes AP1 and AP2 [1w63, 1gw5] have homologous protein chains and are known to have arisen by gene duplication events. Two other homologous complexes, AP3 and AP4 (not shown), have emerged by duplication in a similar manner. Using a conservative definition of homology between complexes, the extent of duplication in three different sets of yeast protein complexes is between 7 and 20%, as shown in the bar chart. MIPS: the manually curated catalogue of protein complexes in the MIPS Yeast Genome Database (Mewes *et al.* 2002) TAP: a set of protein complexes identified by high-throughput purification of tagged proteins and mass spectrometry (Gavin *et al.* 2002). HMS-PCI: another set of protein complexes identified by large-scale experiments in yeast (Ho *et al.* 2002). Circles with error bars represent the random expectation (Pereira-Leal & Teichmann 2005).

vesicles and in separate pathways (Boehm & Bonifacio 2001). Though our study was limited to yeast, there are anecdotal examples suggesting that duplication of protein complexes is widespread in evolution. Mammals, for example, display an additional duplicate of the AP complexes, AP-4 (Boehm & Bonifacio 2001).

Duplicated modules across 10 genomes from the three kingdoms of life were also identified by Eisenberg and co-workers (Li *et al.* 2005) using an approach based on cluster analysis of proteins with similar phylogenetic profiles. This approach does not use prior knowledge of protein complexes, but appears to predict physical interactions rather than any other type of functional association between proteins. All 10 genomes investigated in this study displayed duplicated protein complexes.

In summary, there is clear evidence that protein complexes have evolved by extensive duplication within and across complexes. Incremental evolution by partial duplication is the most common scenario, though some complete duplications of all the components of a complex do exist. After duplication, the functions of complexes diverge in such a way that the core function is conserved while functional specialization is achieved.

How general are these results for other biological interactions? In transcriptional regulatory networks, duplicate transcription factors and target genes

frequently inherit regulatory interactions (Teichmann & Babu 2004). Although duplication of groups of co-regulated genes has not been investigated to our knowledge, it is possible that similar mechanisms may operate in these networks. In metabolic pathways, serial duplication of multiple enzymes is observed rarely, if at all. Duplicate enzymes are distributed across the metabolic network without any coherence, because substrate specificity can change rapidly in evolution (Teichmann *et al.* 2001). Therefore, duplication is unlikely to be a mechanism used in evolution to generate novel, specialized metabolic pathways.

5. NETWORKS OF MODULES IN EVOLUTION

Although modularity confers a degree of isolation to the function accomplished by each module, modules do not exist in isolation. Functional modules exchange matter and information with each other, forming complex networks. Now we consider the evolution of functional modules within these higher-order networks. Do modules delimit mutational change in order to isolate change from the rest of network (Wagner 1996; Wagner 2005)? Or in contrast do they represent evolutionarily stable units that, once established, maintain their composition and do not evolve. These questions are particularly important in the understanding of the role of modularity in evolvability: does evolution tinker with the connections between modules, or with the modules themselves (Hartwell *et al.* 1999)?

In a study of protein interactions in the yeast *S. cerevisiae* (Han *et al.* 2004), highly connected proteins were found to fall into two categories: 'party' and 'date' hubs. The first type of hub binds all its partners simultaneously and these interactions are proposed to be established *within* a functional module (Han *et al.* 2004). In contrast, the date hubs interact with different partners at different times, and are proposed to *bridge* distinct functional modules (Han *et al.* 2004). Thus, these data provide a tool to test evolutionary constraints of proteins within and outside modules. When the nucleotide substitution rates of these types of hubs were studied, it was revealed that even though both date and party hubs accept fewer substitutions over time than other nodes (Fraser 2005), party hubs are the most evolutionarily constrained of the three categories, with date hubs only slightly more constrained than non-hubs (Fraser 2005). The party hubs are frequently subunits of protein complexes (Han *et al.* 2004; Fraser 2005). Thus, these results are in agreement with Teichmann (2002), who showed that components of stable protein complexes are more conserved in evolution than those of transient protein complexes. The larger interface sizes in stable compared to transient protein complexes imposes a greater constraint on their sequences, thus resulting in tighter sequence conservation.

The success of phylogenetic profiling (Pellegrini *et al.* 1999) to identify functional modules (for example, Date & Marcotte 2003; von Mering *et al.* 2003) gives further support to the notion of modules as evolutionarily conserved entities. However, it has recently been questioned to what extent a functional

module is also an evolutionary module (Snel & Huynen 2004). The answer is that it depends on the nature of the module. For example protein complexes are more likely to be evolutionarily conserved than co-expression clusters (Snel & Huynen 2004).

In summary, at least for a subset of functional modules, modularity does seem to play a role in evolvability in the sense that modules such as protein complexes are evolutionarily conserved. This suggests that evolution tinkers with the connections between modules, rather than with the modules themselves (Hartwell *et al.* 1999). In fact, signalling pathways can be co-opted in different developmental contexts, suggesting that a functional module can in fact be a reusable unit that can be wired into different contexts. One example is the signalling module containing *sonic hedgehog*, which is used in both scale and feather development (reviewed in True & Carroll 2002).

Does this mean that evolution does not tinker with the modules themselves? It must, otherwise it would not be possible to develop modules that are functional. In this respect, the robustness conveyed by the modularity of cellular networks may be an important factor in developing new functionalities, and hence in its evolutionary adaptability.

In fact, robustness was proposed recently to play an important role in evolvability (Wagner 2005). The argument is as follows. Robustness implies that most mutations will have little or no phenotypic effect—they are neutral. Thus, these mutations are invisible to selection and cannot be a source of innovation, and hence cannot promote evolvability. However, if these mutations in the robust system do not change its primary function, i.e. the system can perform its function irrespective of the mutations, then these neutral mutations may be the seed for later functional innovation and allow ‘exadaptations’—organismal features that become adaptations long after they arise (Wagner 2005). At the molecular level this is possible, and some argue that it is indeed common (James & Tawfik 2003; James *et al.* 2003; Aharoni *et al.* 2005). For example, Tawfik and co-workers recently showed that using laboratory-directed evolution they could optimize the promiscuous activities of enzymes without losing the primary function of the enzyme (Aharoni *et al.* 2005).

Not just promiscuous enzymes, but also proteins that are part of two or more complexes are multi-functional. As described above, the same protein can become a member of two complexes if its binding partner(s) duplicate, giving rise to two homologous complexes by partial duplication. If the duplicate complexes are formed in the cell at the same time, different molecules of the shared protein are part of the different complexes. If the duplicate complexes assemble and disassemble in a dynamic manner, the shared protein forms a link between the two modules, connecting them within the protein interaction network.

The shared components of multiple complexes are pleiotropic, which means that mutations in these genes will have manifold effects. We would expect such proteins to be subject to a more intense purifying selection than proteins that are part of only a single

Figure 5. Proteins in multiple complexes are often essential. Using the three protein complex datasets in *S. cerevisiae* we determined the probability that a protein is essential ($p(E)$) if it is part of a single complex (white bars) or multiple complexes (black bars).

complex. In addition, essential proteins tend to be part of protein complexes more frequently than non-essential proteins (table 1), so we would expect proteins that are members of multiple complexes to be even more likely essential. When we calculate the fraction of proteins that are essential for proteins in one or multiple complexes in three different datasets, there is a consistent trend for a larger fraction of the multi-complex proteins to be essential (figure 5). This result confirms the important role of duplication in conferring robustness against mutations on an organism. Thus, duplication of complexes has several outcomes: proteins shared between complexes link the modules in the protein interaction network, duplicate complexes can specialize functionally or evolve new functions by exadaptation, and duplication reduces pleiotropy of proteins that are part of a single complex, while proteins shared across multiple complexes are more likely to be essential.

6. OUTLOOK

Here, we have addressed the evolution of protein complexes as functional modules from different perspectives. Analysis of complexes of known three-dimensional structure suggests that formation of protein complexes by homomeric interactions of the same protein is a major mechanism for complex evolution (figure 3). We showed, in three different datasets of experimentally defined complexes in yeast, that duplication of complexes has occurred for a considerable fraction of these. The duplication events result in functional specialization of the new product, as exemplified by the adaptin complexes in protein trafficking (figure 4). Most of these duplications are partial duplications of complexes, leaving proteins that are shared between multiple complexes. This contributes to the modular network of protein interactions, and reduces pleiotropy of the duplicated proteins. The proteins shared across several complexes thus have an increased tendency to be essential (figure 5).

These insights into the evolution of protein complexes still leave us with challenges in terms of understanding the modular nature of cellular networks. For instance, the duplication of complexes, with shared components across several complexes, raises the question of how modules are connected in time and space, and how to take into account the dynamics of module assembly and disassembly. If the shared proteins are permanently associated with each of the duplicate complexes, then they are represented by different molecules, even if the duplicate complexes exist at the same time in the cell. If the shared proteins are transiently associated with the duplicate complexes, then they can form links between the modules and, in certain cases, could be viewed as forming higher order modules consisting of multiple protein complexes.

This raises the issue of the overlap between the different definitions of functional modules. For instance, a protein complex such as the F₁F₀ ATP synthase shown in panel (b) of figure 1 is a functional module itself, but is also a member of a metabolic pathway, in this case the mitochondrial electron transfer chain, which is itself a module as illustrated by the first example in panel (b) of figure 1. This shows that cellular organization can be described effectively by a hierarchy of modules.

This work was supported by the Medical Research Council and the EMBO Young Investigators Programme. We thank Sierhei Maslau for critical reading of the manuscript.

REFERENCES

- Aharoni, A., Gaidukov, L., Khersonsky, O., Mc, Q. G. S., Roodveldt, C. & Tawfik, D. S. 2005 The 'evolvability' of promiscuous protein functions. *Nat. Genet.* **37**, 73–76.
- Amoutzias, G. D., Robertson, D. L., Oliver, S. G. & Bornberg-Bauer, E. 2004 Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. *EMBO Rep.* **5**, 274–279. (doi:10.1038/sj.embor.7400096)
- Amoutzias, G. D., Weiner, J. & Bornberg-Bauer, E. 2005 Phylogenetic profiling of protein interaction networks in eukaryotic transcription factors reveals focal proteins being ancestral to hubs. *Gene* **347**, 247–253. (doi:10.1016/j.gene.2004.12.031)
- Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. 2004 SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.* **32**, D226–D229. (doi:10.1093/nar/gkh039)
- Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M. & Teichmann, S. A. 2004 Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* **14**, 283–291. (doi:10.1016/j.sbi.2004.05.004)
- Bader, G. D. & Hogue, C. W. 2003 An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2. (doi:10.1186/1471-2105-4-2)
- Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. 2003 Dissecting subunit interfaces in homodimeric proteins. *Proteins* **53**, 708–719. (doi:10.1002/prot.10461)
- Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. 2004 A dissection of specific and non-specific protein–protein interfaces. *J. Mol. Biol.* **336**, 943–955. (doi:10.1016/j.jmb.2003.12.073)
- Barabasi, A. L. & Albert, R. 1999 Emergence of scaling in random networks. *Science* **286**, 509–512. (doi:10.1126/science.286.5439.509)
- Barabasi, A. L. & Oltvai, Z. N. 2004 Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113. (doi:10.1038/nrg1272)
- Bell, S. D. & Jackson, S. P. 2001 Mechanism and regulation of transcription in archaea. *Curr. Opin. Microbiol.* **4**, 208–213. (doi:10.1016/S1369-5274(00)00190-9)
- Boehm, M. & Bonifacio, J. S. 2001 Adaptins: the final recount. *Mol. Biol. Cell* **12**, 2907–2920.
- Brazma, A. & Vilo, J. 2000 Gene expression data analysis. *FEBS Lett.* **480**, 17–24. (doi:10.1016/S0014-5793(00)01772-5)
- Calabretta, R., Nolfi, S., Parisi, D. & Wagner, G. P. 1998 Emergence of Functional Modularity in robots. In *From animals to animals*, vol. 5 (ed. R. Pfeifer, B. Blumberg, J.-A. Meyer & S. W. Wilson), pp. 497–504. Cambridge, MA: MIT Press.
- Calabretta, R., Nolfi, S., Parisi, D. & Wagner, G. P. 2000 Duplication of modules facilitates the evolution of functional specialization. *Artif. Life* **6**, 69–84. (doi:10.1162/106454600568320)
- Chakrabarti, P. & Janin, J. 2002 Dissecting protein–protein recognition sites. *Proteins* **47**, 334–343. (doi:10.1002/prot.10085)
- Chothia, C. 1992 One thousand families for the molecular biologist. *Nature* **357**, 543–544. (doi:10.1038/357543a0)
- Cohen, J. D. & Tong, F. 2001 Neuroscience. The face of controversy. *Science* **293**, 2405–2407. (doi:10.1126/science.1066018)
- Coulson, R. M. & Ouzounis, C. A. 2003 The phylogenetic diversity of eukaryotic transcription. *Nucleic Acids Res.* **31**, 653–660. (doi:10.1093/nar/gkg156)
- Coulson, R. M., Enright, A. J. & Ouzounis, C. A. 2001 Transcription-associated protein families are primarily taxon-specific. *Bioinformatics* **17**, 95–97. (doi:10.1093/bioinformatics/17.1.95)
- Date, S. V. & Marcotte, E. M. 2003 Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol.* **21**, 1055–1062. (doi:10.1038/nbt861)
- de Lichtenberg, U., Jensen, L. J., Brunak, S. & Bork, P. 2005 Dynamic complex formation during the yeast cell cycle. *Science* **307**, 724–727. (doi:10.1126/science.1105103)
- Enright, A. J., Iliopoulos, I., Kyripides, N. C. & Ouzounis, C. A. 1999 Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90. (doi:10.1038/47056)
- Fell, D. A. & Wagner, A. 2000 The small world of metabolism. *Nat. Biotechnol.* **18**, 1121–1122. (doi:10.1038/81025)
- Fraser, H. B. 2005 Modularity and evolutionary constraint on proteins. *Nat. Genet.* **37**, 351–352. (doi:10.1038/ng1530)
- Gagneur, J., Krause, R., Bouwmeester, T. & Casari, G. 2004 Modular decomposition of protein–protein interaction networks. *Genome Biol.* **5**, R57. (doi:10.1186/gb-2004-5-8-r57)
- Gavin, A. C. *et al.* 2002 Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147. (doi:10.1038/415141a)
- Goodsell, D. S. & Olson, A. J. 2000 Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 105–153. (doi:10.1146/annurev.biophys.29.1.105)
- Guelzim, N., Bottani, S., Bourguin, P. & Kepes, F. 2002 Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.* **31**, 60–63. (doi:10.1038/ng873)
- Han, J. D. *et al.* 2004 Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **430**, 88–93. (doi:10.1038/nature02555)
- Hansen, T. F. 2003 Is modularity necessary for evolvability? Remarks on the relationship between pleiotropy and

- evolvability. *Biosystems* **69**, 83–94. (doi:10.1016/S0303-2647(02)00132-6)
- Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. 1999 From molecular to modular cell biology. *Nature* **402**, C47–C52. (doi:10.1038/35011540)
- Henrick, K. & Thornton, J. M. 1998 PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **23**, 358–361. (doi:10.1016/S0968-0004(98)01253-5)
- Ho, Y. *et al.* 2002 Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183. (doi:10.1038/415180a)
- Huynen, M., Snel, B., Lathe 3rd, W. & Bork, P. 2000 Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* **10**, 1204–1210. (doi:10.1101/gr.10.8.1204)
- Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. & Barkai, N. 2002 Revealing modular organization in the yeast transcriptional network. *Nat. Genet.* **31**, 370–377.
- James, L. C. & Tawfik, D. S. 2003 Conformational diversity and protein evolution—a 60-year-old hypothesis revisited. *Trends Biochem. Sci.* **28**, 361–368. (doi:10.1016/S0968-0004(03)00135-X)
- James, L. C., Roversi, P. & Tawfik, D. S. 2003 Antibody multispecificity mediated by conformational diversity. *Science* **299**, 1362–1367. (doi:10.1126/science.1079731)
- Janin, J. & Chothia, C. 1990 The structure of protein–protein recognition sites. *J. Biol. Chem.* **265**, 16 027–16 030.
- Janin, J., Miller, S. & Chothia, C. 1988 Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.* **204**, 155–164. (doi:10.1016/0022-2836(88)90606-7)
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A. L. 2000 The large-scale organization of metabolic networks. *Nature* **407**, 651–654. (doi:10.1038/35036627)
- King, A. D., Przulj, N. & Jurisica, I. 2004 Protein complex prediction via cost-based clustering. *Bioinformatics* **20**, 3013–3020. (doi:10.1093/bioinformatics/bth351)
- Kirschner, M. & Gerhart, J. 1998 Evolvability. *Proc. Natl Acad. Sci. USA* **95**, 8420–8427. (doi:10.1073/pnas.95.15.8420)
- Kummerfeld, S. K. & Teichmann, S. A. 2005 Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet.* **21**, 25–30. (doi:10.1016/j.tig.2004.11.007)
- Kunin, V., Pereira-Leal, J. B. & Ouzounis, C. A. 2004 Functional evolution of the yeast protein interaction network. *Mol. Biol. Evol.* **21**, 1171–1176. (doi:10.1093/molbev/msh085)
- Kyrpides, N., Overbeek, R. & Ouzounis, C. 1999 Universal protein families and the functional content of the last universal common ancestor. *J. Mol. Evol.* **49**, 413–423.
- Langlois, R. 2002 Modularity in technology and organization. *J. Econ. Behav. Organ.* **49**, 19–37. (doi:10.1016/S0167-2681(02)00056-2)
- Levitt, M. & Chothia, C. 1976 Structural patterns in globular proteins. *Nature* **261**, 552–558. (doi:10.1038/261552a0)
- Lew, D. J. & Burke, D. J. 2003 The spindle assembly and spindle position checkpoints. *Annu. Rev. Genet.* **37**, 251–282. (doi:10.1146/annurev.genet.37.042203.120656)
- Li, H., Pellegrini, M. & Eisenberg, D. 2005 Detection of parallel functional modules by comparative analysis of genome sequences. *Nat. Biotechnol.* **23**, 253–260. (doi:10.1038/nbt1065)
- Liu, Y. & Eisenberg, D. 2002 3D domain swapping: as domains continue to swap. *Protein Sci.* **11**, 1285–1299. (doi:10.1110/ps.0201402)
- Liu, Y., Hart, P. J., Schlunegger, M. P. & Eisenberg, D. 1998 The crystal structure of a 3D domain-swapped dimer of RNase A at a 2.1-Å resolution. *Proc. Natl Acad. Sci. USA* **95**, 3437–3442. (doi:10.1073/pnas.95.7.3437)
- Lo Conte, L., Chothia, C. & Janin, J. 1999 The atomic structure of protein–protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198. (doi:10.1006/jmbi.1998.2439)
- Lynch, M. & Conery, J. S. 2000 The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155. (doi:10.1126/science.290.5494.1151)
- Lynch, M. & Katju, V. 2004 The altered evolutionary trajectories of gene duplicates. *Trends Genet.* **20**, 544–549. (doi:10.1016/j.tig.2004.09.001)
- Makarova, K. S., Aravind, L., Galperin, M. Y., Grishin, N. V., Tatusov, R. L., Wolf, Y. I. & Koonin, E. V. 1999 Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Res.* **9**, 608–628.
- Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D. 1999 Detecting protein function and protein–protein interactions from genome sequences. *Science* **285**, 751–753. (doi:10.1126/science.285.5428.751)
- Marianayagam, N. J., Sunde, M. & Matthews, J. M. 2004 The power of two: protein dimerization in biology. *Trends Biochem. Sci.* **29**, 618–625. (doi:10.1016/j.tibs.2004.09.006)
- Mewes, H. W. *et al.* 2002 MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30**, 31–34. (doi:10.1093/nar/30.1.31)
- Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. 1995 SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540. (doi:10.1006/jmbi.1995.0159)
- Needham, J. 1933 On the dissociability of the fundamental processes in ontogenesis. *Biol. Rev.* **8**, 180–233.
- Pastor-Satorras, R., Smith, E. & Sole, R. V. 2003 Evolving protein interaction networks through gene duplication. *J. Theor. Biol.* **222**, 199–210. (doi:10.1016/S0022-5193(03)00028-6)
- Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. 1999 Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA* **96**, 4285–4288. (doi:10.1073/pnas.96.8.4285)
- Pereira-Leal, J. B. & Teichmann, S. A. 2005 Novel specificities emerge by stepwise duplication of functional modules. *Genome Res.* **15**, 552–559. (doi:10.1101/gr.3102105)
- Pereira-Leal, J. B., Enright, A. J. & Ouzounis, C. A. 2004 Detection of functional modules from protein interaction networks. *Proteins* **54**, 49–57. (doi:10.1002/prot.10505)
- Pereira-Leal, J. B., Audit, B., Peregrin-Alvarez, J. M. & Ouzounis, C. A. 2005 An exponential core in the heart of the yeast protein interaction network. *Mol. Biol. Evol.* **22**, 421–425. (doi:10.1093/molbev/msi024)
- Prince, V. E. & Pickett, F. B. 2002 Splitting pairs: the diverging fates of duplicated genes. *Nat. Rev. Genet.* **3**, 827–837. (doi:10.1038/nrg928)
- Qin, H., Lu, H. H., Wu, W. B. & Li, W. H. 2003 Evolution of the yeast protein interaction network. *Proc. Natl Acad. Sci. USA* **100**, 12 820–12 824. (doi:10.1073/pnas.2235584100)
- Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabasi, A. L. 2002 Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555. (doi:10.1126/science.1073374)
- Rives, A. W. & Galitski, T. 2003 Modular organization of cellular networks. *Proc. Natl Acad. Sci. USA* **100**, 1128–1133. (doi:10.1073/pnas.0237338100)
- Schlosser, G. 2004 The role of modules in development and evolution. In *Modularity in development and evolution* (ed. G. Schlosser & A. Wagner), pp. 519–582. Chicago: The Chicago University Press.

- Schwartz, M. A. & Madhani, H. D. 2004 Principles of MAP kinase signaling specificity in *Saccharomyces cerevisiae*. *Annu. Rev. Genet.* **38**, 725–748. (doi:10.1146/annurev.genet.39.073003.112634)
- Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. & Friedman, N. 2003 Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176.
- Snel, B. & Huynen, M. A. 2004 Quantifying modularity in the evolution of biomolecular systems. *Genome Res.* **14**, 391–397. (doi:10.1101/gr.1969504)
- Snel, B., Bork, P. & Huynen, M. 2000 Genome evolution. Gene fusion versus gene fission. *Trends Genet.* **16**, 9–11. (doi:10.1016/S0168-9525(99)01924-1)
- Snel, B., Bork, P. & Huynen, M. A. 2002 The identification of functional modules from the genomic association of genes. *Proc. Natl Acad. Sci. USA* **99**, 5890–5895. (doi:10.1073/pnas.092632599)
- Soding, J. & Lupas, A. N. 2003 More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays* **25**, 837–846. (doi:10.1002/bies.10321)
- Spirin, V. & Mirny, L. A. 2003 Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA* **100**, 12 123–12 128. (doi:10.1073/pnas.2032324100)
- Stock, D., Gibbons, C., Arechaga, I., Leslie, A. G. & Walker, J. E. 2000 The rotary mechanism of ATP synthase. *Curr. Opin. Struct. Biol.* **10**, 672–679. (doi:10.1016/S0959-440X(00)00147-0)
- Strogatz, S. H. 2001 Exploring complex networks. *Nature* **410**, 268–276. (doi:10.1038/35065725)
- Teichmann, S. A. 2002 The constraints protein–protein interactions place on sequence divergence. *J. Mol. Biol.* **324**, 399–407. (doi:10.1016/S0022-2836(02)01144-0)
- Teichmann, S. A. & Babu, M. M. 2004 Gene regulatory network growth by duplication. *Nat. Genet.* **36**, 492–496. (doi:10.1038/ng1340)
- Teichmann, S. A., Park, J. & Chothia, C. 1998 Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. *Proc. Natl Acad. Sci. USA* **95**, 14 658–14 663. (doi:10.1073/pnas.95.25.14658)
- Teichmann, S. A., Rison, S. C., Thornton, J. M., Riley, M., Gough, J. & Chothia, C. 2001 The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*. *J. Mol. Biol.* **311**, 693–708. (doi:10.1006/jmbi.2001.4912)
- True, J. R. & Carroll, S. B. 2002 Gene co-option in physiological and morphological evolution. *Annu. Rev. Cell Dev. Biol.* **18**, 53–80. (doi:10.1146/annurev.cellbio.18.020402.140619)
- Vagnarelli, P. & Earnshaw, W. C. 2004 Chromosomal passengers: the four-dimensional regulation of mitotic events. *Chromosoma* **113**, 211–222. (doi:10.1007/s00412-004-0307-3)
- van Noort, V., Snel, B. & Huynen, M. A. 2004 The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep.* **5**, 280–284. (doi:10.1038/sj.embor.7400090)
- Vazquez, A., Flammini, A., Maritan, A. & Vespignani, A. 2003 Modeling of protein interaction networks. *Complexity* **1**, 38–44. (doi:10.1159/000067642)
- Vogel, C., Teichmann, S. A. & Pereira-Leal, J. 2005 The relationship between domain duplication and recombination. *J. Mol. Biol.* **346**, 355–365. (doi:10.1016/j.jmb.2004.11.050)
- von Mering, C., Zdobnov, E. M., Tsoka, S., Ciccarelli, F. D., Pereira-Leal, J. B., Ouzounis, C. A. & Bork, P. 2003 Genome evolution reveals biochemical networks and functional modules. *Proc. Natl Acad. Sci. USA* **100**, 15 428–15 433. (doi:10.1073/pnas.2136809100)
- Wagner, G. (ed.) 1995 *Adaptation and the modular design of organisms*. London: Springer-Verlag.
- Wagner, G. 1996 Homologues, natural kinds and the evolution of modularity. *Am. Zool.* **36**, 36–43.
- Wagner, A. 2005 Robustness, evolvability, and neutrality. *FEBS Lett.* **579**, 1772–1778. (doi:10.1016/j.febslet.2005.01.063)
- Watts, D. J. & Strogatz, S. H. 1998 Collective dynamics of ‘small-world’ networks. *Nature* **393**, 440–442. (doi:10.1038/30918)
- Winther, R. G. 2001 Varieties of modules: kinds, levels, origins, and behaviors. *J. Exp. Zool.* **291**, 116–129. (doi:10.1002/jez.1064)
- Woychik, N. & Reinberg, D. 2001 RNA polymerases: subunits and functional domains. In *Nature encyclopedia of life sciences*. Chichester: Wiley.