Review

Choose your partners: dimerization in eukaryotic transcription factors

Grigoris D. Amoutzias[1,2], David L. Robertson[3], Yves Van de Peer[1,2] and Stephen G. Oliver[4]

[1] Department of Plant Systems Biology, VIB, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium
[2] Department of Molecular Genetics, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium
[3] Faculty of Life Sciences, University of Manchester, Manchester, M13 9PT, UK
[4] Department of Biochemistry, University of Cambridge, Sanger Building, 80 Tennis Court Road, Cambridge CB2 1GA, UK

In many eukaryotic transcription factor gene families, proteins require a physical interaction with an identical molecule or with another molecule within the same family to form a functional dimer and bind DNA. Depending on the choice of partner and the cellular context, each dimer triggers a sequence of regulatory events that lead to a particular cellular fate, for example, proliferation or differentiation. Recent syntheses of genomic and functional data reveal that partner choice is not random; instead, dimerization specificities, which are strongly linked to the evolution of the protein family, apply. Our focus is on understanding these interaction specificities, their functional consequences and how they evolved. This knowledge is essential for understanding gene regulation and designing a new generation of drugs.

Dimerization in signal transduction

Dimerization, in a biological system, is defined as the formation of a functional protein complex composed of two subunits [1]. In general dimeric interactions in signal transduction pathways are not very stable. Rather, they are dynamic and act as reversible switches in the process of information flow [2]. Although dimerization is observed in many signal transduction and regulatory gene families [1,3], this review focuses on eukaryotic transcription factors (TFs). There are several families of TFs (e.g. the ETS-domain family, named for the avian E26 transformation-specific retrovirus) that form heterotypic dimers (see Glossary), but here we will discuss homotypic dimerization (i.e. dimerization among members of the same TF family). These family members can homodimerize or heterodimerize: two monomers might first dimerize in solution and then bind DNA (dimer pathway) or,

Glossary

2R event: The two rounds of whole-genome duplication that probably happened in the common ancestor of vertebrates. A gene subfamily consisting of one gene would duplicate to two and then to four genes.
bHLH: Basic-region helix–loop–helix. This is the largest family of dimerizing TFs in humans, with 118 members. Many are key regulators of developmental processes, differentiation or the cell cycle.
Bilateria: Animals with a bilateral symmetry; that is, they have a front and a back end, as well as an upside and downside. They have bodies that develop from three different germ layers: the endoderm, mesoderm and ectoderm. Mollusks, worms, insects and vertebrates belong to this group.
bZIP: Basic region leucine zipper. This is the second-largest family of dimerizing TFs in humans. They are encoded by 51 genes, many of which are well-studied oncogenes.
Choanoflagellates: Flagellate unicellular eukaryotes. They are considered to be the closest living relatives of the metazoa. The last unicellular ancestors of animals are thought to have resembled modern choanoflagellates.
Cnidaria: Animals with radial symmetry and two germ layers, endoderm and ectoderm. Corals, sea anemones and jellyfish are some of the animals that belong to this group.
Enhanceosome: A complex of TFs and other proteins that assemble and bind cooperatively to the enhancer region of a gene.
Eumetazoa: Animals with true tissues organized into germ layers, and an embryo that goes through a gastrula stage. Cnidaria and bilateria belong to this group, although sponges do not.
HD-ZIP: Homeodomain leucine zipper. The homeodomain is a helix–turn–helix DNA-binding domain, consisting of three alpha helices. The third helix confers DNA-binding specificity.
Heterodimer: A dimer formed between two non-identical proteins.
Heterotypic dimer: A dimer formed between two proteins derived from different gene families.
Homodimer: A dimer formed between two identical protein molecules.
Homotypic dimer: A dimer formed between two proteins derived from the same gene family.
Hub: A highly connected (promiscuous) protein in the protein–protein interaction network.
MADS box: Named after the four originally identified members: MCM1, AGAMOUS, DEFICIENS and SRF1. This family expanded significantly in plants, with 107 genes in *Arabidopsis thaliana*. Many of these genes encode regulators of organ development in plants.
Metazoa: Animals. Eumetazoa and sponges belong to this group.
NFAT: Nuclear factor of activated T cells. It is closely related to the NF-κB family of TFs.
NF-κB: Nuclear factor-kappa B. This is a very well-studied gene family that is involved in the regulation of the immune system, among other processes.
NR: Nuclear receptor. This is the third-largest family of dimerizing TFs in humans, with 48 members. They constitute major pharmaceutical drug targets for a wide range of diseases.
Opisthokonta: Monophyletic group of eukaryotes, including both the metazoan and fungal lineages, together with the phylum choanozoa. Their flagellate cells propel themselves with a single posterior flagellum.
Paralogs: Genes or proteins of the same phylogenetic group in a given organism that are very similar in sequence. They originate from a common gene by one or more gene-duplication events.
PAS domain: Named after three proteins in which it occurs: Per, period circadian protein; Arnt, Ah receptor nuclear translocator protein; and Sim, single-minded protein.
Peripheral member: A poorly connected protein in the protein–protein interaction network.
Protein array: A high-throughput technology that spots whole proteins or protein domains onto glass slides at high density. It is analogous to a DNA microarray but is used for assaying protein–protein interactions.
Scale-free network: A network whose distribution of connectivity (number of proteins with K interactions versus K interactions) decays in a power-law fashion. It has a few hubs and many peripheral members. Its special topology is linked to robustness and fast information flow.
STAT: Signal transducers and activators of transcription. This family has attracted a lot of attention because of its involvement in cancer and its potential as a target for therapeutics. Usually, it is connected to one of the hubs of the network.
Vertebrates: Animals with backbones or spinal columns, for example, fishes, frogs, reptiles and mammals.

Corresponding author: Oliver, S.G. (steve.oliver@mole.bio.cam.ac.uk).

0968-0004/$ – see front matter doi:10.1016/j.tibs.2008.02.002 Available online 9 April 2008

alternatively, they might first bind DNA as monomers and then dimerize (monomer pathway) [4].

Depending on the choice of partner and the cellular context, each unique dimer triggers a sequence of regulatory events that lead to a particular cellular fate. Elucidation of the rules of dimerization and how it has evolved is crucial not only for understanding biological complexity but also for developing novel therapeutic strategies [5]. In the past 10 years, the advancement of high-throughput technologies and the integration of data (by systems biologists) from genomics, transcriptomics, protein–protein interaction experiments and the scientific literature has allowed us, for the first time, to obtain a global snapshot of this complex mechanism. The aim of this review is to discuss: (i) the functional importance of dimerization at the molecular and cellular level, (ii) the specificity of dimerization interactions, (iii) the higher order network architectures that emerge owing to this specificity, (iv) the new regulatory properties that emerge as a consequence of the network architecture, (v) how these complex systems evolved and (vi) how this knowledge might be exploited therapeutically.

Dimerizing transcription factor families

The best studied TF families that form homotypic dimers are the bHLH, bZIP, NR, MADS-box, HD-ZIP and NF-κB families, as well as the STATs (for full names, see Glossary). These TFs create a large number of dimers with distinct biological properties (over 500 dimers in humans and up to 2500 dimers when alternative splicing is accounted for) and form elaborate control circuits that are central to the evolution and generation of organismal complexity. Dimerizing TFs regulate a very wide range of processes such as the cell cycle, reproduction, development, homeostasis, metabolism, immunity, inflammation and programmed cell death [6–13].

Members of these dimerizing TF families mediate their DNA-binding and dimerization activities via highly conserved domains (Box 1). In all families, the DNA-binding

Box 1. DNA-binding and dimerization domains of the major eukaryotic TF families

The eukaryotic TFs contain several domains that direct their ability to bind DNA and to dimerize (Figure I).

- The bHLH TFs contain an alpha-helical basic region (BR) that binds E-box DNA elements (CANNTG) as a dimer. C-terminal to this region lies the helix–loop–helix (HLH) region that forms a four-helix bundle (as a dimer) and directs dimerization ([7] and references therein). Several bHLH phylogenetic groups contain an additional dimerization domain (e.g. the PAS, Orange, or leucine zipper (LZ) domains) C-terminal to the HLH; this domain confers specificity in the range of interactions.
- The bZIPs have a structure that is similar to that of the bHLH proteins, with the different parts' equivalent functions. The alpha-helical basic region recognizes (as a dimer) core hexanucleotide DNA elements. C-terminal to this region is the coiled-coil LZ domain that directs dimerization [39].
- In HD-ZIP TFs, DNA binding is mediated by the homeodomain. Like the bZIP family, the LZ domain (found C-terminal to the homeodomain) directs dimerization [54].
- In the NRs, DNA binding is mediated through zinc-finger (Zf) domains, whereas dimerization is mediated by both the Zf domain and a ligand-binding domain (LBD) [10].
- In MADS-box proteins, the MADS domain directs DNA binding. Domains that mediate dimerization are located C-terminal to the MADS domain. For example, in the MIKC type of MADS-box proteins (named after their specific domain architecture: MADS, I, K and C), the I domain is responsible for dimerization, whereas the coiled-coil K domain acts as an additional dimerization domain that confers specificity [42,67].
- In the NF-κB family, the Rel homology domain (RHD) is composed of two β-sheet immunoglobulin folds connected by a ten amino acid flexible linker [68]. The C-terminal immunoglobulin region (IG) is mainly responsible for dimerization, whereas DNA-binding is mediated by the ten flexible loops within the IGs.
- Dimer stabilization for STATs is achieved by specific and reciprocal interactions between the Src homology 2 (SH2) domain of one monomer and the region of a phosphorylated tyrosine within the other monomer. As with the NF-κB family, DNA binding is mediated by an immunoglobulin fold [69].

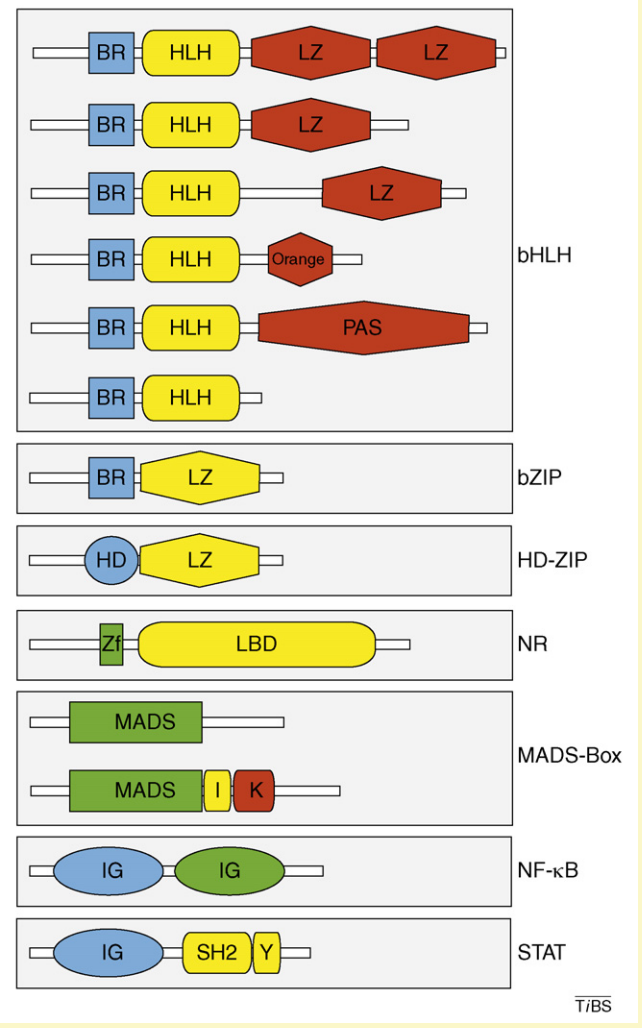

Figure I. Domain architecture of dimerizing TF families. The domain architecture is shown for dimerizing TF families, highlighting only the DNA-binding and dimerization domains or additional dimerization domains that influence interaction specificity. Blue, DNA-binding domains; yellow, dimerization domains; green, domains that have both DNA-binding and dimerization activity; and red, additional dimerization domains that confer specificity.

	Homo sapiens		Saccharomyces cerevisiae		Arabidopsis thaliana	
	Genes	Subfamilies	Genes	Subfamilies	Genes	Subfamilies
bHLH	118	44	8	?	162	12-21
bZIP	51	19	13-14	6-7	75	10
NR	48	19	-	-	-	-
MADS-Box	5	2	4	2	107	5
STAT	7	2	-	-	-	-
NF-κB	5	2	-	-	-	-
HD-ZIP	-	-	-	-	47	4

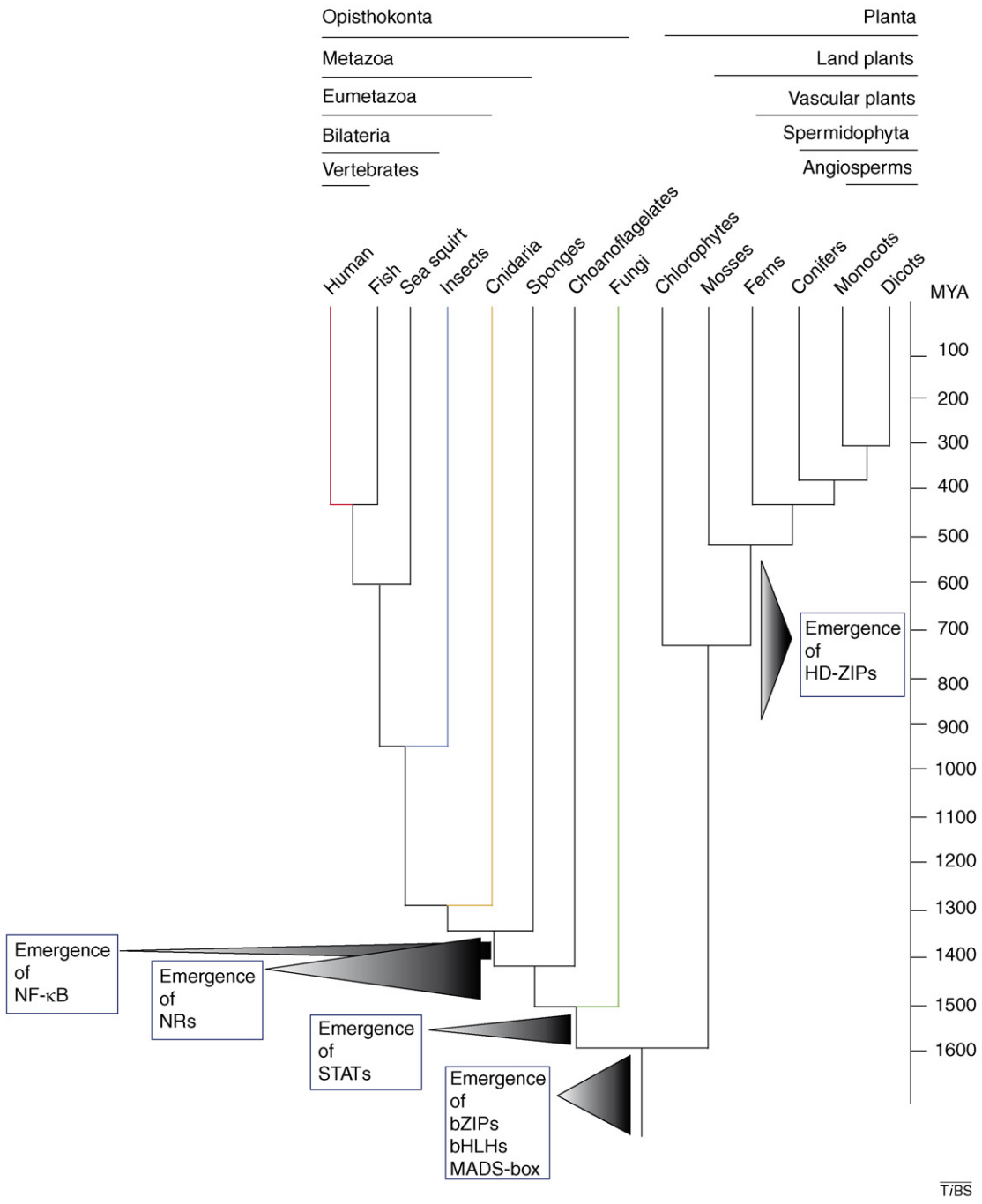

Figure 1. Origin and repertoire of TF dimerizing families in eukaryotes. The species tree reflects the order of evolution of the various eukaryotic phylogenetic lineages. The dates of divergence on the species tree are based on molecular clock studies [65,66] and constitute rough estimates only. The gene numbers and subfamilies and their evolution are based on published studies [7,14–25]. The bZIP, bHLH and MADS-box families emerged at the origin of eukaryotes and are shared by plants, fungi and

domain is the most conserved portion of the protein, whereas the dimerization domain (which usually lies downstream from the DNA-binding domain) is less conserved [7,14–25]. These domains are shared among all the members of each family and, therefore, are used in phylogenetic analyses. Usually, specific functions such as recognition of DNA elements and dimerization are tightly linked to the phylogenetic clustering. The flood of genomic data that has been generated in the past decade has allowed us, for the first time, to find the full complement of these proteins in a species and perform even more reliable phylogenetic analyses that cluster the sequences into subgroups (Figure 1), each with a specific domain architecture. Other regions of the proteins can contain transcriptional activation or repression domains, various functional domains or phosphorylation sites; usually these elements are not as highly conserved.

Functional implications of dimerization

Some of the most important implications of TF dimerization [1] are differential regulation by heterodimerization, facilitated proximity and orientation and enhancement of DNA-binding specificity. In addition, different combinations of heterodimers can recognize various DNA elements, whereas the monomer-to-dimer transition can become an additional point of regulation. Other implications of dimerization are sharp temporal responses to changes in protein concentration, allostery and synergy [9,10]; dampening of noise attributed to stochastic fluctuations in monomer concentrations [26,27]; and non-linear behaviors [28] such as multistability and oscillations.

One TF monomer can have multiple binding partners and, thus, form dimers that possess distinct properties and perform specific functions, thereby mediating differential gene regulation. In this case, the concentration of each monomer in the cell, its posttranslational modifications (e.g. phosphorylation) and its binding affinity for other monomers will determine which dimer will form and, consequently, which signaling process will prevail over the others. A very good example is the Myc–Max and Mad–Max heterodimerization system that defines whether a large number of targeted genes will be expressed or silenced [6,29]. Max is a ubiquitous protein that can heterodimerize with either Myc or Mad. Myc and Mad, in turn, can only heterodimerize with Max and not with each other. If the Myc–Max heterodimer forms, then this dimer recruits the SWI/SNF (switch/sucrose nonfermentable) nucleosome remodeling complex or histone acetyl transferases (HATs) at the promoters of target genes. The SWI/SNF complex disrupts the nucleosome structure, whereas the HAT acetylates highly conserved lysines on histones. In both instances the binding sites of other TFs are exposed on the promoter of the target gene. By contrast, the Mad–Max heterodimer silences the target genes by recruiting histone deacetylases (HDACs) onto the promoter that it binds. HDACs have the opposite effect to HAT recruitment.

Heterodimers combine different DNA-binding domains; thus, each protein complex has distinct DNA-binding specificities. The diverse heterodimeric complexes promote fine-tuning of gene expression through the combination of a limited number of TFs that recognize different promoter elements with various degrees of affinity. Such is the case, for instance, in the DNA recognition of the Jun–ATF2 (activating transcription factor 2) bZIP heterodimers, which have distinguishable binding activities from their parental homodimers [30]. The Jun–ATF2 heterodimer does not bind well to the activator protein-1 (AP-1) site (in contrast to the Jun homodimer), but it binds to the Enk-2 site from the proenkephalin regulatory region better than either homodimer.

Dimerization might represent an additional control point or even a switch in a regulatory pathway. Post-translational modification can determine the transition to a functional dimer, as is the case with the phosphorylation of the STAT proteins [8] or of RelB, which allows the formation of p100–RelB dimers [31]. Phosphorylation of the bHLH protein E47 blocks homodimer formation; this favors heterodimer formation with MyoD, the master regulatory switch of myogenesis [32]. The E47–MyoD heterodimer can then activate muscle-specific transcription.

Dimerization also can amplify the effect of alternative splicing on regulatory complexity [33]. For example, the four MADS-box proteins of the human myocyte enhancer factor 2 (MEF2) subfamily contain 13 alternatively spliced products; see The Ensembl Genome Browser (www.ensembl.org). The number of unique MEF2 isoform dimers is larger than 100, at least in theory, owing to every monomer's ability to form all possible dimers [33].

Specificity of interactions

TF dimerization theoretically could make a huge contribution to gene regulation flexibility and complexity, given the fact that there are ~2000 to ~3000 human-sequence-specific TFs [34,35]. In theory, N genes of a given TF family could generate N homodimers + ($N(N-1)/2$) unique heterodimers, assuming negligible binding specificity among the monomers, a lack of cell- or tissue-specific expression patterns and little alternative splicing. Therefore, for the 51 bZIPs, 118 bHLHs and 48 NRs in humans, there is the potential to form 1326, 7021 and 1176 unique dimers, respectively. In practice, the specificity of monomer–monomer interactions limits the available binding options. Protein-array experiments and reliable predictions based on biophysical constraints on leucine zipper (LZ) interactions lead to estimates of ~350 unique bZIP dimers [36–39]. Strong evidence also exists that indicates dimerization specificity in the bHLHs, NRs, HD-ZIPs, MADS box and plant bZIPs [7,12,40–43] (Box 2).

A very interesting finding is that the paralogs of any given phylogenetic subgroup in bZIPs, bHLHs, NRs, MADS box and HD-ZIPs share, to a high degree, their various dimerization partners [7,12,38,40,43]. This results

animals. The bZIPs and bHLHs underwent independent lineage-specific expansions in plants and animals. The MADS-box proteins underwent lineage-specific expansion in plants. Although no STATs have been identified so far in fungi, they are found in an opisthokont slime mould, *Dictyostelium discoideum*, and in animals. Nevertheless, this family did not undergo significant expansions. NRs emerged at the origin of animals and have undergone significant expansion, thus forming a complex dimerization network in humans. NF-κB is animal specific but did not undergo significant expansions. HD-ZIPs are plant specific and the family expanded significantly. In conclusion, five families expanded independently and significantly in animals and/or plants but not in fungi.

Box 2. Dimerization specificity via domain architecture and point mutations

LZs are amphipathic alpha-helices, and protein–protein interaction specificity is mediated primarily by the electrostatic forces of polar residues at the 'e' and 'g' positions of an LZ heptad repeat (termed 'abcdefg') [39]. In bHLH proteins, the inclusion of the additional dimerization domain (PAS, Orange or LZ) restricts the range of possible protein–protein interactions (reviewed in [7] and references therein). For example, the exclusion of the PAS domain of the protein dioxin receptor (which belongs to the bHLH phylogenetic group C) allows non-native interactions with other bHLH groups, such as upstream transcription factor (USF), that would not have been observed otherwise. Similarly, the presence of two additional LZ domains in the bHLH AP-4 protein inhibits its dimerization with the bHLH E2A hub. Furthermore, the differential spacing of the LZ from the HLH domain in the B-group TFs is a mechanism for restricting interactions in the transcription factor E3 (TFE3). In MIKC MADS-box TFs, the inclusion of the K coiled-coil domain restricts the interactions among members of this group [42,67]. It is also well-established that a small number of point mutations cause subtle structural alterations in the surface of dimerization domains. These mutations result in significant changes in dimerization specificity (e.g. LZs of the bHLH Myc–Max–Mad network; reviewed in [7] and references therein).

from the evolution of the TF families, as will be discussed further below. It appears that phylogenetic clustering, with a few exceptions, is a reliable method of assigning DNA-binding and dimerization specificities to novel TFs.

Evidently, the specificity in TF dimerization is a mechanism to avoid unwanted crosstalk between pathways: proteins from two different phylogenetic groups that are also the end points of two different signaling pathways could have the same spatiotemporal expression pattern and yet avoid unwanted crosstalk at the protein-interaction level because of dimer specificity.

Network topology

Owing to dimerization specificity, higher order structures and specific network architectures with interesting properties emerge at the protein–protein interaction level. Systems biologists have exploited undirected graphs to represent the ensemble of dimerizing interactions and reveal new levels of organization that would not have been evident from a reductionist (i.e. single-gene) approach alone. In these graphs, or 'dimerization networks', nodes represent monomers, whereas edges represent the interactions that bring together two monomers as a dimer. Interestingly, the metazoan bHLH and NR dimerization networks contain a small number of 'promiscuous' proteins (hubs), which have many dimeric partners. For example, the bHLH hub E2A has 38 dimeric partners, whereas the NR hubs retinoid acid receptor (RXR)-α and small heterodimer partner (SHP) have 24 and 22 partners, respectively [7,12]. Most of the other proteins have only few such partners and are termed 'peripheral members'. As a result, their distribution of connectivity (number of proteins with K interactions versus K interactions) decays in a power-law fashion, which is reminiscent of scale-free networks [44]. Although we are not aware of any similar statistical analysis of the plant MADS-box TFs, members from one (Mα) out of the five MADS-box subfamilies nonetheless interact with most of the other MADS-box proteins (including the MIKC subfamily), thus behaving as hubs [42].

Box 3. The importance of provenance in interaction data

Networks with a power-law distribution of connectivity, also termed scale-free networks, have been shown to occur repeatedly in both the natural and man-made worlds, for example, in protein–protein interaction or TF regulatory networks in cells, or on the internet [44]. Nevertheless, not all biological networks are scale free [70]. Moreover, rigorous analysis is required before assuming a scale-free topology and using it to explain the properties of biological networks [71–73]. A protein in a given network topology might falsely appear to be highly connected owing to experimental biases, for example, the high rate of false positives or low reproducibility in direct binary interaction assays using the yeast-two-hybrid (Y2H) method [71,74].

Nevertheless, careful examination of the quality of dimerization data provides confidence in the conclusions reached for these particular networks, thus highlighting the value of medium-scale analyses. One of the most comprehensively assembled dimerization datasets for NRs demonstrates that the topology of the network and the central role of its hubs are not artifactual [12]. Six independent datasets, which used different data-sampling methods, were integrated in this study ([12,75] and references therein). In most of the datasets, RXR and SHP emerge as the most connected nodes and, therefore, the data-sampling method does not alter this conclusion.

A dataset also was assembled by manually mining the text of the literature to study the bHLH dimerization network and its evolution [7]. The same conclusions regarding network topology were reached even when we used a limited dataset from the public database TRANSFAC (www.gene-regulation.com), which contained only 30% of the total interactions.

Although the dimerization network of human bZIP TFs is not scale free, its topology was reconstructed from high-quality interaction data. Protein-array technology, which is free from the biases that afflict Y2H assays, was used to monitor the bZIP interactions [38]. The interaction array showed high symmetry and reproducibility and was in good agreement with the literature. In addition, rules that were derived from this dataset for predicting specificity were in good agreement with rules derived from independent studies [39]. Notably, although these prediction rules apply well to metazoan bZIPs, they seem to have medium-to-poor performance in predicting the interactions between plant bZIPs [41], probably because of the idiosyncratic domain architecture (longer LZs) of these plant TFs.

From this research, several important questions arise about the quality of systems analysis data (Box 3). Potential biases could lead to erroneous conclusions, which would affect our understanding of the biological meaning of these network architectures and the role of evolution in shaping these particular networks.

Functional implications of dimer network architecture

There is widespread interest in network architectures because of their emergent properties. Analyses of high-throughput functional data demonstrate that scale-free networks are usually robust to perturbations of peripheral members but also vulnerable to the loss of hubs [44]. Particularly for the genome-wide protein–protein interaction and regulatory TF networks, the more connected the protein, the higher the probability that it is essential for cell viability [45]. In addition, in the yeast genome-wide protein-interaction network [45], hubs seem to be under tighter regulatory control than peripheral members. Nevertheless, hub proteins do not appear to evolve more slowly than peripheral members [45].

Fast information flow is another property that is proposed to characterize scale-free networks. A perturbation at one node of the network can propagate to any other node through a minimal number of steps (mainly via the hubs) [46]. Nevertheless, recent reports demonstrate that the introduction of certain quantitative parameters (e.g. protein concentrations and dimerization affinities) into a yeast dimerization network can minimize the effect of widespread propagation by dampening the perturbation and keeping it localized within a subnetwork [47]. So far, we are not aware of any study that has looked for these properties in TF dimerization networks. Owing to their small size – 500 nodes (proteins) – such global analyses should be applied with caution. In addition, peripheral members of the dimerizing networks could emerge as hubs at the TF-regulatory-network level. Therefore, the properties of a given protein are actually the result of the integration of various levels and networks of regulation, which probably complicate these analyses.

Apart from the general properties that are observed in genome-scale networks, several interesting properties have been observed for the dimerizing TF networks themselves. The hub-based topology of the bHLH dimerization network gives rise to a special mechanism of repression and creates high-dimensional switches [48]. A large number of peripheral bHLH (some with restricted cell-type- or tissue-specific expression patterns) compete to interact with the widely expressed hub protein to form functional heterodimers. In this way, the prevalence of one peripheral member over the others (and, therefore, the choice of one cellular fate over the others) is exerted by two mechanisms: monopolization of the much-needed hub and competition for the same DNA elements.

In the bZIP gene family, there is a link between the topology of the bZIP dimerization network and the reduction and oxidation (redox) control of dimer–DNA binding [20]. In particular, bZIP proteins that are (or are predicted to be) under redox control for DNA binding preferentially dimerize with other bZIPs that also are subject to redox control. This feature is robust, even when accounting for biases introduced by gene duplication, and was revealed only when a systems biology approach was employed [20].

Dimerization can drive transcriptional repression

Repression is as important as activation in transcriptional regulation; it is usually achieved in dimerization networks through the use of 'poison' subunits, also termed 'dominant negatives'. In such cases, one of the two monomers lacks a DNA-binding domain. The two monomers can still heterodimerize, but a complex is formed in which there is only one DNA-binding region, instead of the required two. Therefore, DNA binding (of the dimer) and subsequent target-gene activation or repression is disrupted. Dominant negatives can arise through either gene duplication followed by domain loss or alternative splicing.

The members of the Id subfamily of bHLH TFs are a well-studied example of dominant-negative repressors. Ids lack a DNA-binding region and, consequently, can disrupt the function of other bHLH TFs via dimerization [49]. Interestingly, Ids dimerize with the hub E2A and, therefore, have the potential to make the hub unavailable to all the other peripheral members that require an interaction with E2A to perform their function. This sequestration seems to disable the weakest peripheral elements [48] and probably functions as a dampener of transcriptional noise. Ids can dimerize with (and therefore repress) only a few peripheral members (~9), rather than all. Instead of evolving one repressor for every peripheral activator, this subnetwork only needs to express one of the four Id proteins.

The same mechanism is observed for the NR protein SHP, which lacks a DNA-binding domain but, in addition, harbors a transcriptional repression domain. SHP interacts with and disrupts the function of the ubiquitously expressed hub RXR [50]. Furthermore, SHP also dimerizes with most of the peripheral members within the network. This feature is important for widespread repression because the NR peripheral proteins do not rely solely on the hub RXR for being functional (in contrast to those of the bHLH network) but also can homodimerize or heterodimerize with other peripheral members. Again, repression is achieved in an economical way, owing to the hub-based topology of the network. Another interesting example is the PAS-domain protein Period (Per), which lacks a bHLH domain but can still dimerize with other bHLH-PAS proteins, including the hub protein aryl hydrocarbon receptor nuclear translocator (ARNT). In this way, Per functions as a dominant negative [51].

Apart from the dominant-negative mechanism of repression at the protein–protein interaction level, another method of repression is exhibited by the murine single-minded proteins (mSIMs), which contain both a bHLH and a PAS domain. mSIMs have a higher affinity for the hub ARNT than does the aryl hydrocarbon receptor (AHR) [52]. mSIMs also contain a transrepression domain. Thus, they not only sequester the hub away from AHR but also are active repressor proteins. The effect of mSIMs is similar to that of Mad in the Myc–Max–Mad bHLH-LZ subnetwork.

Dimerization origins

Eukaryotic TF dimerization emerged several times in protein families with otherwise structurally unrelated folds. Several studies [53–55] suggest that the TFs originally functioned as monomers. Their DNA-binding domains must have been able to bind DNA as monomers with sufficient strength to activate or repress transcription. Within several promoters, symmetrical palindromic repeats of the DNA-recognition motif probably brought two or more copies of the same TF molecule into close proximity. If an interaction domain with only one interaction surface appeared by chance, then this would strengthen the formation of the composite element on DNA because the TF complex would recognize a larger DNA motif [54]. Therefore, the evolutionary constraints on the TF DNA-binding domain (within a redundant duplicated gene) would be relaxed and permit the emergence of a DNA-binding domain that bound with less affinity but could still function. Once a duplicate gene diverges in such a way, there is no turning back and it must function as an obligate dimer. From that point further duplication and

changes in specificity probably led to the appearance and rapid diversification of the various dimerizing TF families.

The above evolutionary reconstruction is based on several experimental observations. Within several TF families, close or distant protein family members exist that are functional (even) as monomers (e.g. in NRs, HD-ZIPs, NF-κB and NFAT) [53–55]. Two to four point mutations in a DNA-binding domain are sufficient for the

Box 4. Evolution of the dimerization and DNA-binding of bZIPs in metazoa

Owing to the plethora of functional data available, the bZIPs are a well-studied case of how dimerization and DNA-binding evolved in animals. Current genomic data show that the genome of the last common ancestor of eumetazoa contained genes for many dimerizing bZIP subfamilies [11]. Most of these subfamilies must have emerged after the divergence of the fungi and before that of the cnidaria (Figure I). Only one of the 19 human bZIP subfamilies was shared with the fungi, whereas 13 are shared with cnidaria. In addition, these 13 subfamilies recognize all six DNA elements bound by the human bZIPs. Therefore, specificity of DNA binding mainly evolved during this period. Several of these ancient bZIP subfamilies subsequently duplicated and, while retaining their DNA-binding affinity for certain motifs, started to diverge at their dimerization domains, thus gaining and losing interactions with other bZIP subfamilies. This change in dimerization specificity could contribute to the evolution of key developmental circuits. By the time the common ancestor of bilateria arose (just before the hypothesized Cambrian explosion), 17 of the 19 bZIP subfamilies were present and formed a complex core-dimerization network, which is conserved in many vertebrate and invertebrate bilaterians. Until this time, most of the subfamilies must have consisted of only one gene.

At the origin of vertebrates, around 550 million years ago, all of the 19 bZIP subfamilies were present. The two rounds of whole-genome duplication (the 2R hypothesis) that the vertebrate ancestor underwent created more paralogs for each subfamily. These paralogs not only retained the DNA-binding specificity of their ancestral molecule but also kept most of its dimerizing interactions. The paralogs evidently diverged outside of the DNA-binding and dimerization domains, thus making new interactions with other signal-transduction molecules. Subsequent gene duplications and losses of TFs also occurred in the vertebrate lineage, though to a limited extent [64]. The evolution of bZIP dimerization in animals is based on [11].

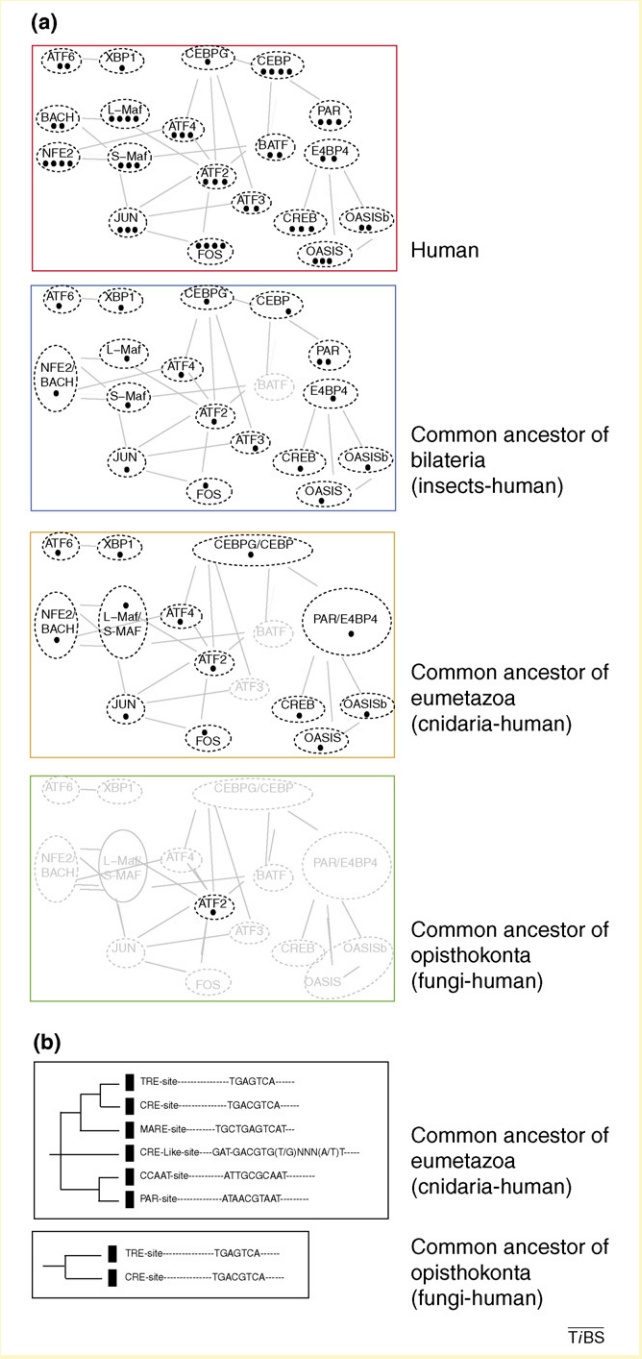

Figure I. Evolution of bZIP dimerization in metazoa. Evolution and expansion of the bZIP subfamilies during the major phases of animal macroevolution. **(a)** The dotted bubbles represent the subfamilies, whereas the black dots within them represent an estimation of how many genes each subfamily contained at the given evolutionary stage. The edges among the subfamilies represent the dimerizing interactions projected from the human interaction data. **(b)** The various DNA elements and their consensus sequences, which are recognized by bZIPs, are shown. The common ancestor of fungi and animals had bZIPs that recognized at least two DNA elements, whereas the common ancestor of cnidaria and bilateria had bZIPs that recognized all the DNA elements that the human bZIPs recognize today.

progressive transition from a functional monomer to an obligate dimer. In addition, the wide utilization [4] of the monomer pathway of DNA-dimer formation is probably a consequence of TF monomers (capable of DNA binding) that later evolved to become functional as obligate dimers. Furthermore, in many dimerizing families the dimerization domain contains an α-helical component that resembles a coiled-coil domain and directs dimerization. Coiled-coil domains, owing to their intrinsic structure and ubiquity, could easily appear by chance [56] or exon shuffling. Furthermore, several TFs can bind adjacent DNA-binding-sites and interact with each other, thus forming complex enhanceosomes [57]. These enhanceosomes could provide the opportunity for closely positioned TFs to evolve new interacting surfaces and begin to function as dimers.

Evolution of dimerization networks

Although some of the DNA-binding folds of dimerizing TFs are found also in prokaryotes (e.g. HTH), all of the TF families that we have discussed so far are specific to eukaryotes. Some are found in the metazoan, fungal and plant lineages (bHLH, bZIP and MADS box, whereas others are specific to plants (HD-ZIP), the metazoa and choanoflagellates (NF-κB and NR) or opisthokonta (STATs) (Figure 1) [7,14–25]. Although several ancient TF families are found in all three of these eukaryotic lineages, some have independently undergone significant lineage-specific expansion in only one (i.e. MADS-box TFs in plants) or two (i.e. bHLH and bZIP in metazoa and plants) of the lineages.

The integration of genomic and functional data for the three largest families of dimerizing TFs in metazoa – the bHLHs [7], NRs [12] and, especially, the bZIPs [11] – delineates, to some extent, the evolution of DNA-binding and dimerization specificity during the major phases of animal macroevolution (Box 4). The exact protein repertoire of the last common ancestor of metazoa remains elusive but will soon be revealed [14,18] because the genomes of a unicellular choanoflagellate (*Monosiga brevicolis*) and a multicellular sponge (*Amphimedon queenslandica*) are available now either as shotgun traces or even as an assembled genome (see the NCBI HomePage at http://www.ncbi.nlm.nih.gov/). Nevertheless, from the genomes of fungi, diploblastic cnidarians (*Nematostella vectensis*), invertebrates (insects) and vertebrates (fishes and humans), we can now understand more about the events that occurred during the emergence of bilaterian and vertebrate animals (see Figure 1 and Figure I in Box 4).

Briefly, major gene duplications, point mutations and domain rearrangements occurred at the origin of metazoa. These events shaped the repertoire of gene subfamilies and the interactions among them. By the time the urbilaterian ancestor arose, a core highly conserved dimerization network already had been formed, with most of the subfamilies present. Later, two rounds of whole-genome duplication occurred at the origin of vertebrates (2R event) and added more paralogs in each subfamily. These highly similar paralogs possessed very similar DNA-binding and dimerization specificities. The evolution of the bZIP, bHLH and NR networks strongly supports this picture [7,11,12].

This evolutionary reconstruction provides significant insights into the functionality of TF dimers and has the practical consequence that we can predict, with some accuracy, the dimerization specificity of a given TF, provided we have both a good knowledge of its phylogeny and interaction data for several of its paralogs [12]. TF dimerization specificity seems to be highly conserved [7,11,12], and this is in agreement with global analyses of protein–protein interaction networks in which the homotypic dimers are more conserved than the heterotypic dimers [58]. Nevertheless, we cannot extend this conclusion to dimerization affinities; more data are needed. Moreover, current knowledge does not permit us to extend our predictions confidently to other types of interactions, such as interactions among signal transduction molecules.

Dimerization: a therapeutic target

Recently, the disruption of TF dimerization has emerged as a novel and promising strategy for developing a new generation of drugs [59]. In many human cancers, for instance, a relatively small number of oncogenic TFs (mostly dimerizing proteins) are overactive [5] compared to the much larger number of unregulated oncoproteins that function in upstream signal-transduction pathways. Inhibition of these TFs is the subject of much current attention. So far, the disruption of dimerization in the bZIP and bHLH families has been achieved by using dominant-negative peptides. These synthetic peptides lack the canonical DNA-binding domain and, therefore, can disrupt endogenous TF dimerization and DNA binding [60]. Recently, several groups achieved disruption of protein–protein interactions by small non-peptide molecules, which are generally considered more 'druggable' than peptides [61]. This strategy seemed unlikely to be successful a few years ago because it was generally believed that interaction surfaces were too large to be disrupted by such small molecules. An encouraging degree of inhibition of cancer cell growth has been reported after the disruption of c-Myc and STAT3 dimerization by small non-peptide molecules [62,63]. Knowledge of the phylogeny of a dimeric TF family, targeted by these small molecules, would reveal which other close homologs the drug could potentially disrupt apart from the main targeted TF. This knowledge, when combined with coexpression and interaction data, could be useful in identifying potential side effects of the anticancer agent.

Conclusions

If all the genes in a genome are considered to be a dictionary for the language of life, then the syntactic and grammatical rules of that language would be equivalent to gene regulation. From a small set of genes, many different phenotypes can emerge, depending on the genes' spatiotemporal expression, the splicing of their transcripts and the posttranslational modifications of their protein products. Undoubtedly, TFs, as controllers of transcription initiation, influence such important biological functions as gene regulation and the complexity of development and differentiation [35,64]. Those TF families that can combine in homotypic and dimeric protein complexes create a large number of TF dimers with distinct properties, equaling or

even exceeding the actual number of sequence-specific TF genes in a genome. Owing to the specificity of dimerization, higher order network structures emerge in which the whole is more than the sum of its parts. Gene duplication and changes in specificity shape these networks and allow for widespread changes as well as fine-tuning. Undoubtedly, dimerizing TFs are at the core of many regulatory circuits that generate the complexity of organisms.

The systems biology approach has already started to bear fruit and, in the next few years, there should be a shift in focus from qualitative to quantitative data. A protein–protein interaction is not a binary event but follows the laws of mass balance. Reliable high-throughput technologies are needed to measure protein concentrations, binding affinities and posttranslational modifications. In this way, the static networks that we have started to unravel will eventually evolve into dynamic simulations. These simulations will be even more meaningful and will reveal, in full detail, the competition of molecules for common partners and the complex dynamics that emerge. It's all about choices, in the end.

Acknowledgements

We would like to thank Sarah Teichmann, Andrew Sharrocks, Elgar Pichler and Erich Bornberg-Bauer for useful discussions. G.D.A. is supported by a long-term European Molecular Biology Organization fellowship in the group of Y.V.d.P. Work on protein–protein interactions in the Manchester and Cambridge laboratories is supported by grants to D.L.R. and S.G.O. from the Biotechnology and Biological Sciences Research Council.

References

1 Klemm, J.D. *et al.* (1998) Dimerization as a regulatory mechanism in signal transduction. *Annu. Rev. Immunol.* 16, 569–592
2 Nooren, I.M. and Thornton, J.M. (2003) Diversity of protein–protein interactions. *EMBO J.* 22, 3486–3492
3 Marianayagam, N.J. *et al.* (2004) The power of two: protein dimerization in biology. *Trends Biochem. Sci.* 29, 618–625
4 Kohler, J.J. and Schepartz, A. (2001) Kinetic studies of Fos. Jun. DNA complex formation: DNA binding prior to dimerization. *Biochemistry* 40, 130–142
5 Darnell, J.E., Jr (2002) Transcription factors as targets for cancer therapy. *Nat. Rev. Cancer* 2, 740–749
6 Grandori, C. *et al.* (2000) The Myc/Max/Mad network and the transcriptional control of cell behavior. *Annu. Rev. Cell Dev. Biol.* 16, 653–699
7 Amoutzias, G.D. *et al.* (2004) Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. *EMBO Rep.* 5, 274–279
8 Levy, D.E. and Darnell, J.E., Jr (2002) Stats: transcriptional control and biological impact. *Nat. Rev. Mol. Cell Biol.* 3, 651–662
9 Desvergne, B. *et al.* (2006) Transcriptional regulation of metabolism. *Physiol. Rev.* 86, 465–514
10 Germain, P. *et al.* (2006) International Union of Pharmacology. LXIII. Retinoid X receptors. *Pharmacol. Rev.* 58, 760–772
11 Amoutzias, G.D. *et al.* (2007) One billion years of bZIP transcription factor evolution: conservation and change in dimerization and DNA-binding site specificity. *Mol. Biol. Evol.* 24, 827–835
12 Amoutzias, G.D. *et al.* (2007) A protein interaction atlas for the nuclear receptors: properties and quality of a hub-based dimerisation network. *BMC Syst. Biol.* 1, 34
13 Wietek, C. and O'Neill, L.A. (2007) Diversity and regulation in the NF-κB system. *Trends Biochem. Sci.* 32, 311–319
14 Simionato, E. *et al.* (2007) Origin and diversification of the basic helix–loop–helix gene family in metazoans: insights from comparative genomics. *BMC Evol. Biol.* 7, 33
15 Bertrand, S. *et al.* (2004) Evolutionary genomics of nuclear receptors: from twenty-five ancestral genes to derived endocrine systems. *Mol. Biol. Evol.* 21, 1923–1937
16 Alvarez-Buylla, E.R. *et al.* (2000) An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. *Proc. Natl. Acad. Sci. U. S. A.* 97, 5328–5333
17 Shore, P. and Sharrocks, A.D. (1995) The MADS-box family of transcription factors. *Eur. J. Biochem.* 229, 1–13
18 Gauthier, M. and Degnan, B.M. (2008) The transcription factor NF-κB in the demosponge *Amphimedon queenslandica*: insights on the evolutionary origin of the Rel homology domain. *Dev. Genes Evol.* 218, 23–32
19 Wang, Y. and Levy, D.E. (2006) *C. elegans* STAT: evolution of a regulatory switch. *FASEB J.* 20, 1641–1652
20 Amoutzias, G.D. *et al.* (2006) Reduction/oxidation-phosphorylation control of DNA binding in the bZIP dimerization network. *BMC Genomics* 7, 107
21 Deppmann, C.D. *et al.* (2006) Cross-species annotation of basic leucine zipper factor interactions: insight into the evolution of closed interaction networks. *Mol. Biol. Evol.* 23, 1480–1492
22 Bailey, P.C. *et al.* (2003) Update on the basic helix–loop–helix transcription factor gene family in *Arabidopsis thaliana*. *Plant Cell* 15, 2497–2502
23 Jakoby, M. *et al.* (2002) bZIP transcription factors in *Arabidopsis*. *Trends Plant Sci.* 7, 106–111
24 Nakamura, M. *et al.* (2006) Characterization of the class IV homeodomain-leucine zipper gene family in *Arabidopsis*. *Plant Physiol.* 141, 1363–1375
25 De Bodt, S. *et al.* (2003) And then there were many: MADS goes genomic. *Trends Plant Sci.* 8, 475–483
26 Palena, C.M. *et al.* (1999) A monomer-dimer equilibrium modulates the interaction of the sunflower homeodomain leucine-zipper protein Hahb-4 with DNA. *Biochem. J.* 341, 81–87
27 Bundschuh, R. *et al.* (2003) The role of dimerization in noise reduction of simple genetic networks. *J. Theor. Biol.* 220, 261–269
28 Smolen, P. *et al.* (2000) Mathematical modeling of gene networks. *Neuron* 26, 567–580
29 Luscher, B. (2001) Function and regulation of the transcription factors of the Myc/Max/Mad network. *Gene* 277, 1–14
30 Hai, T. and Curran, T. (1991) Cross-family dimerization of transcription factors Fos/Jun and ATF/CREB alters DNA binding specificity. *Proc. Natl. Acad. Sci. U. S. A.* 88, 3720–3724
31 Maier, H.J. *et al.* (2003) Critical role of RelB serine 368 for dimerization and p100 stabilization. *J. Biol. Chem.* 278, 39242–39250
32 Lluís, F. *et al.* (2005) E47 phosphorylation by p38 MAPK promotes MyoD/E47 association and muscle-specific gene transcription. *EMBO J.* 24, 974–984
33 Black, B.L. and Olson, E.N. (1998) Transcriptional control of muscle development by myocyte enhancer factor-2 (MEF2) proteins. *Annu. Rev. Cell Dev. Biol.* 14, 167–196
34 Kummerfeld, S.K. and Teichmann, S.A. (2006) DBD: a transcription factor prediction database. *Nucleic Acids Res.* 34 (Database issue), D74–D81
35 van Nimwegen, E. (2003) Scaling laws in the functional content of genomes. *Trends Genet.* 19, 479–484
36 Fong, J.H. *et al.* (2004) Predicting specificity in bZIP coiled-coil protein interactions. *Genome Biol.* 5, R11
37 Grigoryan, G. and Keating, A.E. (2006) Structure-based prediction of bZIP partnering specificity. *J. Mol. Biol.* 355, 1125–1142
38 Newman, J.R. and Keating, A.E. (2003) Comprehensive identification of human bZIP interactions with coiled-coil arrays. *Science* 300, 2097–2101
39 Vinson, C. *et al.* (2006) Deciphering B-ZIP transcription factor interactions *in vitro* and *in vivo*. *Biochim. Biophys. Acta* 1759, 4–12
40 Johannesson, H. *et al.* (2001) DNA-binding and dimerization preferences of *Arabidopsis* homeodomain-leucine zipper transcription factors *in vitro*. *Plant Mol. Biol.* 45, 63–73
41 Ehlert, A. *et al.* (2006) Two-hybrid protein-protein interaction analysis in *Arabidopsis* protoplasts: establishment of a heterodimerization map of group C and group S bZIP transcription factors. *Plant J.* 46, 890–900
42 de Folter, S. *et al.* (2005) Comprehensive interaction map of the *Arabidopsis* MADS Box transcription factors. *Plant Cell* 17, 1424–1433
43 Veron, A.S. *et al.* (2007) Evidence of interaction network evolution by whole-genome duplications: a case study in MADS-box proteins. *Mol. Biol. Evol.* 24, 670–678

44 Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113
45 Batada, N.N. *et al.* (2006) Evolutionary and physiological importance of hub proteins. *PLoS Comput. Biol.* 2, e88. DOI: 10.1371/journal.pcbi.0020088 (http://compbiol.plosjournal.org)
46 Wagner, A. and Fell, D.A. (2001) The small world inside large metabolic networks. *Proc. Biol. Sci.* 268, 1803–1810
47 Maslov, S. and Ispolatov, I. (2007) Propagation of large concentration changes in reversible protein-binding networks. *Proc. Natl. Acad. Sci. U. S. A.* 104, 13655–13660
48 Cinquin, O. and Page, K.M. (2007) Generalized, switch-like competitive heterodimerization networks. *Bull. Math. Biol.* 69, 483–494
49 Ruzinova, M.B. and Benezra, R. (2003) Id proteins in development, cell cycle and cancer. *Trends Cell Biol.* 13, 410–418
50 Seol, W. *et al.* (1997) Novel receptor interaction and repression domains in the orphan receptor SHP. *Mol. Cell. Biol.* 17, 7126–7131
51 Lindebro, M.C. *et al.* (1995) Protein-–protein interaction via PAS domains: role of the PAS domain in positive and negative regulation of the bHLH/PAS dioxin receptor-Arnt transcription factor complex. *EMBO J.* 14, 3528–3539
52 Ema, M. *et al.* (1996) Two new members of the murine Sim gene family are transcriptional repressors and show different expression patterns during mouse embryogenesis. *Mol. Cell. Biol.* 16, 5865–5875
53 Moraitis, A.N. and Giguere, V. (1999) Transition from monomeric to homodimeric DNA binding by nuclear receptors: identification of RevErbAα determinants required for RORα homodimer complex formation. *Mol. Endocrinol.* 13, 431–439
54 Tron, A.E. *et al.* (2004) Engineering the loop region of a homeodomain-leucine zipper protein promotes efficient binding to a monomeric DNA binding site. *Biochemistry* 43, 15845–15851
55 de Lumley, M. *et al.* (2004) A biophysical characterisation of factors controlling dimerisation and selectivity in the NF-κB and NFAT families. *J. Mol. Biol.* 339, 1059–1075
56 Brendel, V. and Karlin, S. (1989) Too many leucine zippers? *Nature* 341, 574–575
57 Panne, D. *et al.* (2007) An atomic model of the interferon-β enhanceosome. *Cell* 129, 1111–1123
58 Pereira-Leal, J.B. *et al.* (2007) Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol.* 8, R51
59 Wells, J.A. and McClendon, C.L. (2007) Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* 450, 1001–1009
60 Zhang, J.W. *et al.* (2004) Dominant-negative C/EBP disrupts mitotic clonal expansion and differentiation of 3T3-L1 preadipocytes. *Proc. Natl. Acad. Sci. U. S. A.* 101, 43–47
61 Pagliaro, L. *et al.* (2004) Emerging classes of protein–protein interaction inhibitors and new tools for their development. *Curr. Opin. Chem. Biol.* 8, 442–449
62 Kiessling, A. *et al.* (2006) Selective inhibition of c-Myc/Max dimerization and DNA binding by small molecules. *Chem. Biol.* 13, 745–751
63 Song, H. *et al.* (2005) A low-molecular-weight compound discovered through virtual database screening inhibits Stat3 function in breast cancer cells. *Proc. Natl. Acad. Sci. U. S. A.* 102, 4700–4705
64 Blomme, T. *et al.* (2006) The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.* 7, R43
65 Hedges, S.B. *et al.* (2004) A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol. Biol.* 4, 1–9
66 Zimmer, A. *et al.* (2007) Dating the early evolution of plants: detection and molecular clock analyses of orthologs. *Mol. Genet. Genomics* 278, 393–402
67 West, A.G. *et al.* (1998) DNA binding and dimerisation determinants of *Antirrhinum majus* MADS-box transcription factors. *Nucleic Acids Res.* 26, 5277–5287
68 Chen, F.E. and Ghosh, G. (1999) Regulation of DNA binding by Rel/NF-κB transcription factors: structural views. *Oncogene* 18, 6845–6852
69 Chen, X. *et al.* (1998) Crystal structure of a tyrosine phosphorylated STAT-1 dimer bound to DNA. *Cell* 93, 827–839
70 Pereira-Leal, J.B. *et al.* (2005) An exponential core in the heart of the yeast protein interaction network. *Mol. Biol. Evol.* 22, 421–425
71 Aloy, P. and Russell, R.B. (2002) Potential artefacts in protein-interaction networks. *FEBS Lett.* 530, 253–254
72 Coulomb, S. *et al.* (2005) Gene essentiality and the topology of protein interaction networks. *Proc. Biol. Sci.* 272, 1721–1725
73 Hakes, L. *et al.* (2008) Protein–protein interaction networks and biology – what's the connection? *Nat. Biotechnol.* 26, 69–72
74 von Mering, C. *et al.* (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 417, 399–403
75 Albers, M. *et al.* (2005) Automated yeast two-hybrid screening for nuclear receptor-interacting proteins. *Mol. Cell. Proteomics* 4, 205–213