Identification and analysis of evolutionarily cohesive functional modules in protein networks

Mónica Campillos, Christian von Mering, Lars Juhl Jensen and Peer Bork

Genome Res. 2006 16: 374-382; originally published online Jan 31, 2006;
Access the most recent version at doi:10.1101/gr.4336406

Letter

Identification and analysis of evolutionarily cohesive functional modules in protein networks

Mónica Campillos, Christian von Mering, Lars Juhl Jensen, and Peer Bork[1]

The European Molecular Biology Laboratory (EMBL), 69117 Heidelberg, Germany

The increasing number of sequenced genomes makes it possible to infer the evolutionary history of functional modules, i.e., groups of proteins that contribute jointly to the same cellular function in a given species. Here we identify and analyze those prokaryotic functional modules, whose composition remains largely unchanged during evolution, and study their properties. Such "cohesive" modules have a large number of internal functional connections, encode genes that tend to be in close proximity in prokaryotic genomes, and correspond to physical complexes or complex functional systems like the flagellar apparatus. Cohesive modules are enriched in processes such as energy and amino acid metabolism, cell motility, and intracellular trafficking, or secretion. By grouping genes into modules we achieve a more precise estimate of their age and find that the young modules are often horizontally transferred between species and are enriched in functions involved in interactions with the environment, implying that they play an important role in the adaptation of species to new environments.

[Supplemental material is available online at www.genome.org.]

Functional modules, groups of proteins that work together for the same cellular function, have been described in a variety of networks, e.g., as enzymatic pathways in metabolic networks (Ravasz et al. 2002), as groups of interconnected proteins in protein interaction networks (Ravasz et al. 2002; Rives and Galitski 2003), or as closely linked clusters in predicted in silico protein association networks (Snel et al. 2002b; von Mering et al. 2003a). Functionally linked proteins have been shown to evolve together (Pellegrini et al. 1999; Ettema et al. 2001, 2003) and proteins with similar phylogenetic distributions are often components of the same pathway (Huynen and Bork 1998; Marcotte et al. 1999; Pellegrini et al. 1999, 2001; Wu et al. 2003). Although algorithms that identify sets of genes with similar phylogenetic distributions are able to reconstruct many known pathways (Pellegrini et al. 1999; Date and Marcotte 2003; Wu et al. 2003), a recent study of the evolutionary modularity of several types of modules indicates that they show only limited conservation during evolution (Snel and Huynen 2004).

Nevertheless, some prokaryotic modules do show evolutionary cohesion (i.e., their components are frequently gained, transferred, or lost together); these are often conserved at the operon level and frequently encode biosynthetic pathways (Snel and Huynen 2004). Several hypotheses have been put forward to explain these observations, such as the notion of "selfish operons" (Lawrence 1997), although only a few operons are stable over very long evolutionary time scales (Itoh et al. 1999; Lathe III et al. 2000). There seem to be differences in the evolution of cohesive modules, as some prokaryotic metabolic pathways show a broad phylogenetic conservation (Peregrin-Alvarez et al. 2003), while others are more restricted to specific groups of bacteria (Martin et al. 2003). Some modules are more cohesive than others: Operons coding for physical complexes such as ribosomal proteins, proton ATPases, and ABC-type membrane transporters show conservation even at the level of gene order (Mushegian and Koonin 1996; Siefert et al. 1997; Dandekar et al. 1998; Wolf et al. 2001). However, others have been subjected to more dynamic evolution, with frequent losses and gains of genes in different phylogenetic lineages (Tanaka et al. 2005).

The lack of a quantitative measure of cohesiveness has so far prevented comparative analysis of the evolutionary properties of functional modules. Here, we identify modules that appear to be cohesive during evolution and perform a parsimony analysis to determine when and why these modules appeared. We first study topological and functional properties of these modules; we then classify them into *ancestral*, *intermediate*, and *young* age according to their inferred first appearance during evolution, and study functional characteristics of these age classes. Finally, we analyze the horizontal transfer of cohesive modules in extant species and the role of this transfer in the adaptation of species to environments.

Results and Discussion

Quantifying the evolutionary cohesiveness of functional modules

Defining functional modules

We have previously identified functional modules by clustering neighbors in protein interaction networks (von Mering et al. 2003a). We used interaction networks that cover multiple organisms at once, with each network node corresponding to an orthologous group of proteins (hereafter OG, see Methods). Edges between the nodes represent functional associations, derived by combining a variety of different protein interaction data including experimentally verified interactions, predicted interactions based on gene context methods such as gene neighborhood and fusion, as well as interactions derived from text-mining analysis (von Mering et al. 2005). As is generally the case for prokaryotes, the biggest contributors of association information are chromosomal neighborhood and text-mining, but the other association types contribute as well, sometimes even forming the majority of the interactions in a module (Table S4).

Functional modules derived this way have a high coverage

[1]Corresponding author.
E-mail bork@embl.de; fax +49 6221-387-517.
Article published online ahead of print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.4336406.

16:374–382 ©2006 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

and accuracy when benchmarked against manually curated *Escherichia coli* metabolic pathways (von Mering et al. 2003a) and they cover a broad range of cellular functions. As they are comprehensive and unsupervised (i.e., largely objective), they form a good basis for a systematic analysis of the evolution of prokaryotic functional modules. Applying the concept to 102 prokaryotic species with completely sequenced genomes, a total of 1161 functional modules were identified, containing 3812 out of the 9912 prokaryotic OGs included in these species.

Tracing the evolutionary history of modules

We inferred, for each functional module, the most plausible evolutionary history through parsimony analysis: In a phylogenetic tree containing 110 species (86 bacteria, 16 archaea, and eight eukaryotes), the presence or absence of each module component was inferred for all ancestral nodes in the tree (Fig. 1A). The evolutionary events that were modeled for this parsimony analysis were (1) gene birth, (2) gene loss, and (3) gene acquisition. We assigned relative costs to each type of event (see Methods) and computed for each module which evolutionary scenario incurred the lowest overall cost (using dynamic programming to screen scenarios capable of explaining the present-day distribution of the module). In the cost function, we assumed that multiple events happening at the same time incurred a somewhat lower cost than multiple independent events (as long as they were of the same type; see Methods). Our approach is based on two implicit assumptions: Proteins known to interact today are likely to have interacted also in the past, and a certain degree of evolutionary dependence between functional partners exists (hence the lower cost for events happening simultaneously). The latter assumption is not essential: All results reported below are also observed when full evolutionary independence is assumed (Fig. S6).

Scoring and defining evolutionarily cohesive modules

Given the most parsimonious scenario for the evolution of the genes in a module, we then asked to what extent the module was "cohesive," i.e., whether events involving one of the proteins had an influence on other proteins in the module (more so than what would be expected for random modules). We assessed how many events were "joined" (i.e., proteins lost or acquired together, at the same node in the tree). Together with the cost function, the "fraction of joined events" provides a measure describing the evolutionary history of each module. We compared both measures to values derived from a conservative randomization of modules (see Methods) and derived a single *P*-value for each module. We chose this approach (i.e., combining parsimony analysis with Monte Carlo *P*-value computation) because it explicitly models evolutionary events against the backdrop of the known species phylogeny, while at the same time it provides a quantitative measure of cohesiveness that can be used for ranking and comparing modules. At a cutoff of $P < 0.01$, we found 472 of the 1161 functional modules to be cohesive, in agreement with previous qualitative estimates of the modularity of functional modules (Snel and Huynen 2004) (Fig. 1C).

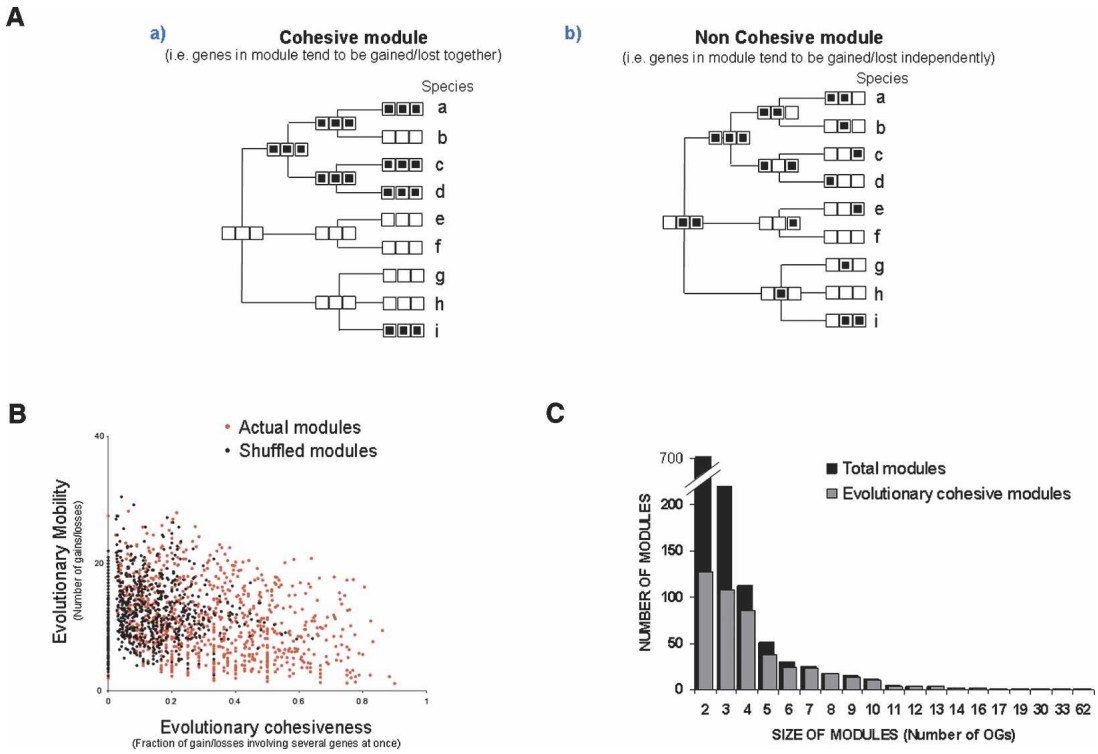

Figure 1. Quantification of evolutionary cohesiveness. (*A*) A simplified example of the ancestral states of a cohesive functional module (a) and a noncohesive module (b). The presence of a gene in a species or ancestral state is indicated by a black square. (*B*) The two evolutionary parameters are plotted for prokaryotic functional modules and random modules. The "normalized total cost" and the "fraction of joined events" for the cohesive (a) and non cohesive (b) modules of Figure 1*A* are indicated. (*C*) Distribution of total modules and evolutionarily cohesive modules ($P < 10^{-2}$), by module size.

Properties of evolutionarily cohesive functional modules

Cohesive modules are larger and more highly connected

Evolutionarily cohesive modules are frequently large (containing four or more components, represented here by distinct OGs; Fig. 1C). The size distribution of cohesive modules is globally shifted towards larger sizes—this is especially visible for module sizes two and three (Fig. 1C) but is also evident when considering only modules of size four and larger ($P < 0.05$; Kolmogorov-Smirnoff test). This may be partially due to the larger information content in the phylogenetic profiles of large modules, making any non-random behavior easier to detect. However, both large and small modules can be found among the most cohesive modules, indicating that size is not a dominating factor in determining cohesiveness. One possible reason for the cohesiveness of large modules could be an inherent resistance against "break-up" during evolution, as they tend to have a higher number of internal functional interactions. To test whether the number of interactions correlates with cohesiveness, we quantified the internal network connectivity (C) of modules, defined as the number of internal connections present, divided by the theoretical maximum of all possible internal connections between OGs in the module. Indeed, a positive correlation between the evolutionary cohesiveness of the modules and internal network connectivity was observed (Fig. 2, top). This correlation is stronger when only considering neighborhood associations between OGs for the connectivity measure (Fig. 2, top green). Thus, conserved operons in particular, or parts of "uber-operons" (i.e., a limited number of operons that rearrange in a restricted way by using the same pool of genes [Lathe III et al. 2000]), frequently form cohesive modules. Accordingly, the analysis of associations in relation to cohesiveness reveals that cohesive modules tend to have a higher percentage of genome neighborhood associations (although other types of associations are present as well, Fig. S4), while noncohesive modules are slightly enriched in text-mining interactions.

Genes encoding cohesive module components rarely duplicate

The high internal connectivity of evolutionarily cohesive modules also indicates a higher functional dependency between the genes of such modules; i.e., when one of the genes is disrupted, the functionality of the others may also be compromised. Likewise, gene duplications in a cohesive module should have a lower chance of survival, particularly when the other genes in the modules are not duplicated alongside, as has been observed for protein complexes in eukaryotic genomes (dosage sensitivity) (Veitia 2002; Papp et al. 2003; Yang et al. 2003). We would thus expect to find a relatively low level of paralogy for cohesive modules (i.e., such modules should have a tendency to keep a low gene copy number, maintaining their "gene stochiometry"). Indeed, we find that cohesive modules have significantly fewer duplicated genes than noncohesive modules (Figs. 2 and S7). Cohesive modules are significantly enriched in protein complexes (based on keywords such as "subunit," "chain," "complex," and "component," data not shown), but even modules that do not appear to encode protein complexes show fewer gene duplications when they are cohesive. Thus, apart from protein complexes (Yang et al. 2003) and single copy orthologs (Ciccarelli et al. 2005), cohe-

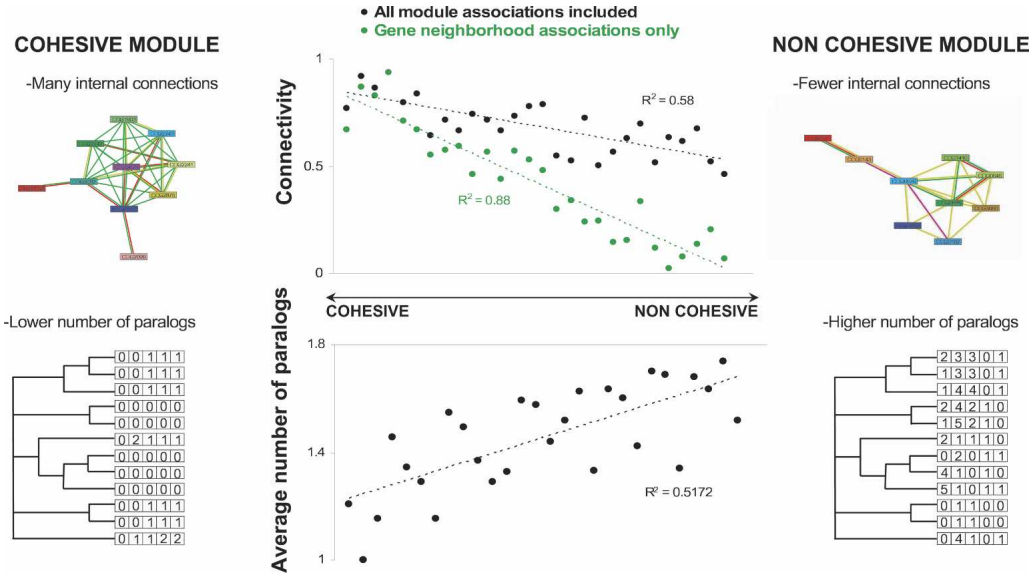

Figure 2. Properties of evolutionarily cohesive modules. *Top*: internal connectivity of modules. The internal connectivity is defined here as the number of actual connections between OGs divided by the number of possible connections. The connectivity of modules is plotted against their cohesiveness ranking (black), and for comparison the plot is repeated for the same modules but limiting the connections to the neighborhood associations (green, association score >700). Two examples that illustrate differences in internal connectivity of a cohesive and a noncohesive module are shown. The cohesive module ($P = 10^{-8}$) contains genes for cobalamin (vitamin B12) biosynthesis and the genes of noncohesive module ($P = 0.34$) are implicated in methionine biosynthesis. Color code for connections: neighborhood, green; gene fusion, red; text-mining, yellow; experiments, violet. *Bottom*: average number of paralogs in a module. An example of a cohesive module is shown (plotting the number of paralogous genes for a subset of species; the module encodes a metabolic pathway involved in the conversion of succinate to propionate: COG1272, COG0427, COG1703, COG1884, $P = 7 \cdot 10^{-7}$), as well as an uncharacterized module with a lower cohesiveness (COG1305, COG0714, COG1721, COG1001, COG2252, $P = 3 \cdot 10^{-3}$). For both plots, modules were ranked according to evolutionary cohesiveness (*P*-value, only considering modules of size four or larger) and binned into groups of 10 modules.

sive functional modules may also display a degree of dosage sensitivity.

Cohesive modules are frequently found in processes interacting with the environment

To obtain an overview of the functional capacities of cohesive network modules, we classified the proteins forming these modules according to broad functional categories annotated for the respective OGs (Tatusov et al. 2001). The vast majority of proteins that are singletons (i.e., not grouping into modules) are uncharacterized (Fig. 3A). However, apart from uncharacterized modules, the relative coverage of major functional categories (information storage and processing, metabolism, other cellular processes including signaling) remains roughly the same between singletons and modules, i.e., the latter seem not to be biased toward certain processes in the cell. For example, there are about twice as many functional modules implicated in metabolism than in translation, both in singletons and in functional modules (Fig. 3A).

When comparing cohesive and noncohesive modules, those involved in metabolism and some other cellular processes (e.g., cell motility or secretion) are clearly enriched (Fig. 3B). Among the metabolic categories, the production and conversion of energy, as well as the transport and metabolism of amino acids are overrepresented in cohesive modules (Fig. 3B) while within the other cellular processes, cell motility and intracellular trafficking, secretion, as well as vesicular transport are found more frequently than expected. Overall, modules in information processing are slightly less cohesive than those in other cellular processes. This is somewhat surprising given the well-known conservation of translational and transcriptional processes, and could be indicative of differences in cohesiveness of modules of different evolutionary ages. The fact that both conserved functions and functions of more recent origin are represented in cohesive modules leads us to perform a deeper study of the evolutionary age of cohesive modules and their contribution to ongoing species adaptation.

Distinct properties of ancestral and young modules

Defining the age of modules

We define ancestral modules as having emerged before the split of bacteria, archaea, and eukaryotes. Young modules have emerged in only a single species or in the ancestor of a set of species so closely related that they still retain significant synteny. Intermediate modules appeared between the ancestral and young modules. We assigned these age categories to all 472 cohesive modules by identifying the most ancestral node in the species phylogeny for which at least 70% of the genes of the module were found to be present during parsimony analysis (Table S2). One hundred and twenty-six of these cohesive modules were classified as ancient, 124 as intermediate, and 151 as young (for the rest, no clear assignment was possible, see Methods). This use of modules for age classification—as opposed to classifying genes individually—has the advantage of potentially guiding the classification of a gene by its functional partners. Accordingly, we observe that the functional properties of the age classes (see below) are defined more sharply, i.e., the differences between age classes are larger than when classifying genes individually. The signal is also clearer when comparing genes in cohesive modules versus genes in noncohesive modules (Fig. S5).

Functional differences between ancestral and young modules

Many of the observed differences in module function (with respect to module age) reflect general trends that are also evident from the genomic background of individual genes: Modules involved in information processing are frequently of ancient origin, while uncharacterized modules often appear to be young. However, significant functional enrichments within certain age

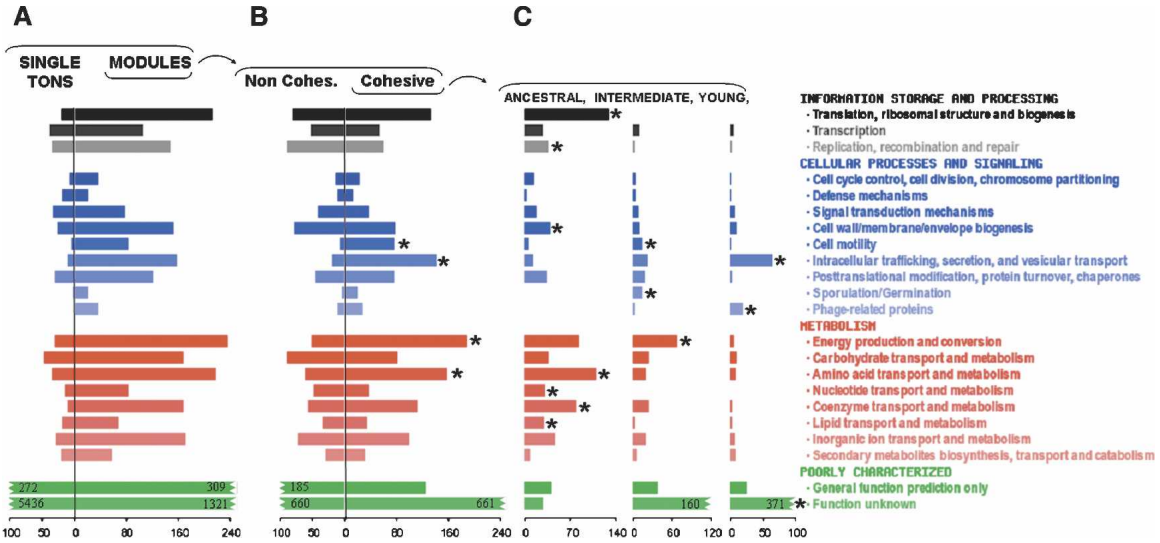

Figure 3. Functional characteristics of modules. The high-level categories defined for COGs (Tatusov et al. 2001) were used to assess functional differences between (*A*) OGs in modules versus unconnected singleton OGs, (*B*) cohesive and noncohesive modules, and (*C*) cohesive modules of different age classes. The total counts of each set are shown at the *bottom*. Colors reflect the hierarchical classification of functional categories, as in the COG database (Tatusov et al. 2001). Statistically significant functional enrichments of cohesive over noncohesive modules are marked with asterisks (*P*-values $<10^{-2}$, calculated using a hypergeometric distribution with Bonferroni correction). In part (*C*), the statistical significance of functional enrichments was computed for each age class (asterisks), using the hypergeometric distribution as *above*.

classes are visible exclusively within cohesive modules (Figs. 3C and S5)—these we discuss below.

Ancient and intermediate cohesive modules are often involved in metabolism (Fig. 3C). Large, ancestral metabolic modules include widespread energetic complexes like ATPases, oxidoreductases, and dehydrogenase complexes (Table S3). The rarer formylmethanofuran dehydrogenase and Acetyl-CoA decarbonylase/synthase complexes of the Euryarchaeota and the Methanosarcina subgroup of Archeabacteria appear to belong to the intermediate age category, together with the photosynthetic complexes of Cyanobacteria (Table S3). Other ancestral metabolic modules include the majority of amino acid biosynthesis pathways and amino acid transporters of the ABC type, although some of the latter are also found in the intermediate age group. In general, the ancestral metabolic modules are among the best conserved during evolution (Kunin and Ouzounis 2003).

When considering cell motility, around half of the components of bacterial flagella are ancestral, while pilus, secretion systems of type II, III, and IV, and conjugation systems appear to be of intermediate or young age. Although functionally different, a phylogenetic relationship between flagellar and a type III secretion system has been suggested (Galan and Collmer 1999; Macnab 1999; Nguyen et al. 2000) based on structural and sequence similarities. These data suggest that young modules may have evolved from older ones by a whole module duplication mechanism.

In general, young modules are responsible for the majority of the functions related to communication with the environment. Although we cannot exclude that some of these modules might turn out to be older when more genomes become sequenced, some of these functions have likely evolved recently and help to adapt to new environments.

Enrichment of uncharacterized younger modules

The high number of uncharacterized or poorly characterized cohesive modules, most of them of intermediate or young age, is striking and hints at a variety of complex functions still to be discovered. It includes a few large and widespread modules, e.g., one with as many as 16 components conserved in many bacteria (Table S3). It is probably involved in the reaction to external factors as one of the proteins is predicted to be coregulated with hemolysin and another is involved in temperature-dependent secretion. Most of the cohesive modules are, however, restricted to a limited number of species or environments and thus might be horizontally transferred between organisms.

Horizontal transfer of modules

To study the horizontal transfer of modules and how it contributes to the evolution of species, we used our parsimony reconstruction to quantify the number of modules likely transferred recently (i.e., classified as "gained" in an extant species, but not in any of its next relatives in the species phylogeny). As many as 447 instances of recent module gain were observed (assumed to be horizontal acquisitions; see Methods for details). Similarly, 480 module losses were observed and the respective cohesive modules were assigned to functional categories (Tatusov et al. 2001).

Frequent horizontal transfer of younger cohesive modules

Horizontal transfer of intermediate and young cohesive modules seems a likely scenario as they usually have a conserved operon structure (Table S2), whereas ancestral modules are often encoded by several operons that are subject to frequent rearrangement and gene exchange (so-called "uber-operons"; [Lathe III et al. 2000]). The uber-operon structure of ancient modules may be explained by the fact that they contain, on average, more proteins than intermediate and young modules (Fig. 4B), which is in accordance with finding that longer amino acid synthesis pathways exhibit lower rates of changes in pathway structure than shorter ones (Rutter and Zufall 2004). We find that only very few ancestral operons have been retained, corresponding to physically interacting products, as has been observed before (Itoh et al. 1999). The closer genome proximity of the genes constituting young modules should simplify the horizontal transfer of entire processes.

Transferred cohesive modules are selected positively in big genomes

In order to assess whether gains and losses of cohesive modules are due to positive selection or neutral evolution, we studied

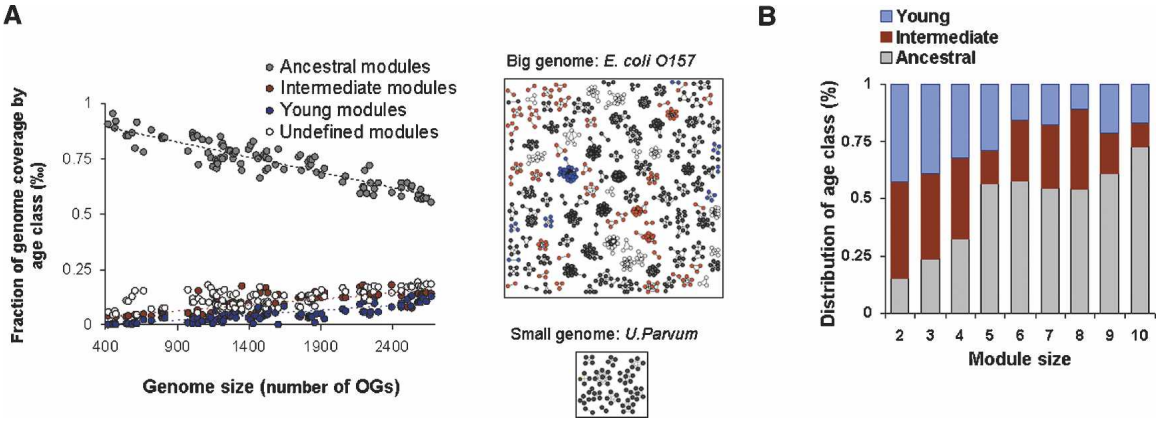

Figure 4. Properties of ancestral, intermediate, and young modules. (*A*) Correlation of genome size with proportion of ancestral modules. For each prokaryotic species, the proportion of OGs in ancestral, intermediate, young, and undefined age modules are plotted against genome size (R^2 values of the ancestral, intermediate, young, and undefined age classes are 0.85, 0.71, 0.41, and 0.69, respectively). The genome size is measured as number of OGs per species. An example of the number of modules in each age class in a large species (*Escherichia coli O157: H7*) and in a small species (*Ureaplasma Parvum*) is depicted. (*B*) Distribution of module size per age class.

whether gain or loss correlates with genome size and/or elapsed time and examined the functions of the corresponding genes. We found that the number of modules lost correlates with time (Fig. S3). The functions of the 480 modules that have been lost are very similar to the range of functions represented in cohesive modules in general (compare Fig. 3B and Fig. 5A) with the exception of genes involved in information storage and processing, which are almost never lost. Together, this indicates a rather clock-like behavior and hints at module loss being a largely neutral process. Indeed, many reduced genomes correspond to obligate intracellular parasites, which have gradually lost a vast variety of modules that are not needed in their highly specialized environment.

In contrast, the number of modules gained correlates with genome size (Fig. S3), suggesting that these modules are selected for adaptation purposes as has been observed for individual genes in archaeal and proteobacterial genomes (Snel et al. 2002a). It is primarily young and intermediate modules that are gained, whereas the number of ancestral modules increases only slightly with genome size (Fig. 4A). Three functional categories are overrepresented among the modules gained, when compared against all genes gained: energy production and conversion, protein secretion systems, and unknown functions. These functional biases of transferred modules agree with previous studies of transferred genes (Rivera et al. 1998; Daubin et al. 2003; Nakamura et al. 2004). Certain functions enriched in large genomes and among transferred genes, such as regulatory functions (van Nimwegen 2003; Konstantinidis and Tiedje 2004; Nakamura et al. 2004), are not found among the frequently transferred modules detected here, in accordance with the lower evolutionary cohesiveness of this type of modules (Snel and Huynen 2004).

Transferred modules contribute to complex phenotypes

To analyze the benefits of module gain, we studied the phenotypic properties of the respective species. For this purpose, we used terms describing phenotypic characteristics for 91 bacteria that were extracted from the literature (Korbel et al. 2005). They were assigned to the 20 species that have gained the most modules. Seventy percent of the most significant 30 terms associated with these species could be classified into two main general processes: degradation and communication/defense (Fig. 5C). The same procedure applied randomly to 20 species does not report any term related to the two categories (data not shown). Terms such as "biofilm," "lactamases," "degrader," and "cefoxitin" suggest that bacteria that acquired the most modules live in competitive environments (high species density, antibiotics, etc.) with varying and stressful external conditions (e.g., limited water supply, presence of xenobiotics). In fact, a high level of conjugative gene transfer has been reported in biofilms (Hausner and Wuertz 1999) and several cohesive modules have been reported to be involved in the degradation of xenobiotic compounds (van der Meer et al. 1998; Smejkal et al. 2001; Johnson et al. 2002). Taken together, both the gain and the loss of entire functional modules seem to contribute to bacterial diversification, enabling fast adaptation to new niches.

Conclusions

In summary, we quantified the degree of evolutionary cohesiveness of functional modules in protein interaction networks. Although there is a continuum from extremely conserved to rapidly changing modules, we have been able to detect largely cohesive modules by tracing the evolution of their components

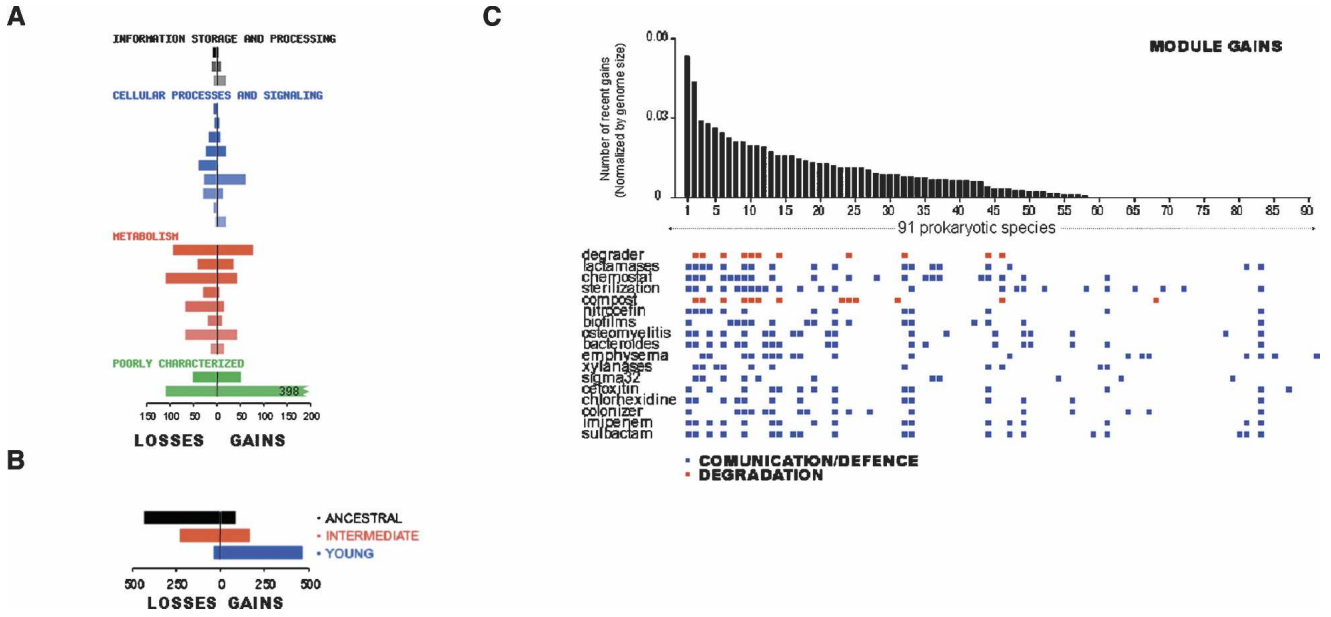

Figure 5. Relationship between module gain and specific phenotypes. (*A*) The functional categories of the OGs in modules that have been lost or gained in the 102 prokaryotic species are shown. A module is considered to be lost or gained when 70% or more of its OGs are lost or gained. Statistically significant functional enrichments of genes in cohesive modules over all genes are marked with asterisks (P-values $< 10^{-2}$, calculated using a hypergeometric distribution with Bonferroni correction). (*B*) The number of OGs in the three age categories of the modules that are gained or lost is represented. (*C*) Phenotypic properties of species that have gained a higher number of modules. Species were sorted by the number of modules that have been gained since speciation from the closest relative in the tree, divided by total number of OGs in the species (see Table S5 for the species list). Phenotypic terms associated to the 20 species that have gained more modules among the 30 first terms significantly associated ($P < 10^{-3}$) (one-tailed probability of the χ^2 distribution, one degree of freedom), to these species are shown.

and by developing a statistical scoring function. We consider ~40% of the 1161 modules analyzed as evolutionarily cohesive; they represent a mix between ancient, large, often ubiquitous modules involved in information storage processes, and younger modules that perform functions which enable adaptation to diverse environments. Cohesive modules are frequently acquired as units and contribute considerably to the phenotypic diversity of extant species.

Methods

Orthologous groups and functional modules

Groups of orthologous genes (OGs) with at least one gene present in prokaryotic genomes were extracted from STRING, version 5.1 (von Mering et al. 2003b, 2005), a database that integrates and extends the widely used clusters of orthologous groups (COG) database (Tatusov et al. 2001). Within the STRING framework, novel orthologous groups have been automatically created for gene families not yet covered in the COG database; these are called NOGs (non-supervised orthologous groups). Where these appeared in modules of interest, we classified and annotated them manually, using the same classification scheme as in the COGs database. In order to avoid artifacts introduced by poorly resolved orthologous groups, we excluded groups (COGs or NOGs) that contained, on average, more than four distinct proteins per species (this excluded less than 3% of OGs). In total, 9913 groups in 102 prokaryotic species were considered. From the STRING database, we also imported functional association data, linking these groups into a network; the associations were based on conserved gene neighborhood, gene fusion, experimental data (such as affinity purifications/yeast-two-hybrid screens), knowledge databases (such as the MIPS collection of annotated complexes), and text-mining. We chose not to consider associations based on microarray data (which are sparse and sometimes noisy in prokaryotes) and associations based on co-occurrence (the latter assume evolutionary cohesiveness and could possibly lead to circular reasoning). Functional modules were defined by clustering the association network using UPGMA (unweighed pair group method with arithmetic mean) using a cutoff association score of 0.400 as described before (von Mering et al. 2003a).

Parsimony analysis: Inferring ancestral module states

We based the parsimony analysis on a phylogenetic species tree consisting of 86 bacteria, 16 archaea, and eight eukaryotes, all of which have been completely sequenced. The tree is based on a manual integration of a variety of phylogenies: gene order trees (Blanchette et al. 1999; Snel et al. 1999), genome content trees (Snel et al. 1999), and the NCBI taxonomy.

In order to infer ancestral states of a module, we first recorded the presence or absence of each gene (i.e., orthologous group) in each of the 110 extant species. Then, the most parsimonious scenario (Fitch 1971; Hartigan 1973) of presence/absence of all the genes at all ancestral nodes of the tree was predicted, using an in-house C++ program that calculates the scenario with the lowest evolutionary cost. In cases where several equivalent scenarios were possible, one of them was chosen randomly. The relative costs of the evolutionary events were estimated after parameter exploration as follows: two cost units for gene birth or gene acquisition, and one cost unit for gene loss. This 2:1 ratio of costs (gain penalty) has been estimated by others (Snel et al. 2002a; Kunin and Ouzounis 2003) and results in a higher enrichment of modules having genes in the same functional category than is the case for gain penalty 1, a parameter that has been suggested as well (Mirkin et al. 2003) (results reported in this study are reproducible with both settings, Fig. S6). For cases in which several genes underwent the same change at the same time (gain or loss), all but the first event were assigned only half the cost, reflecting the known tendency of functionally linked systems to evolve together (Pellegrini et al. 1999; Ettema et al. 2001, 2003). We decided to add this initial assumption of evolutionary cohesiveness into the scoring scheme, because the main goal of this study is to reliably rank and study the highly cohesive modules, not to prove that cohesiveness as such exists. Due to the high computational cost of predicting ancestral states of modules of size 16 or larger, we applied an approximation algorithm for these. The approximation consisted of collapsing equal phylogenetic profiles, assuming that identical phylogenetic profiles have identical evolutionary histories, and also partitioning the module in smaller submodules when needed. For these cases, the ancestral states of all module genes were then reconstructed manually from the simplified scenarios.

It should be noted that our approach is distinct from a traditional phylogenetic profile analysis, because it explicitly takes into consideration the known species phylogeny and models the past states of a module. As a result, it can better determine which profiles contain the most evolutionary "signal," i.e., require the largest number of evolutionary events to explain the present-day patterns.

Classifying functional modules by evolutionary age

All internal nodes in the phylogenetic species tree were classified into one of three age categories: Ancestral nodes are (1) the root of the tree, and the respective last common ancestors of (2) eubacteria, (3) archaea, and (4) eukaryotes. Nodes in the young category are those whose descendents still have retained significant genome synteny (syntenic groups were delineated manually using STRING). All remaining nodes were classified as intermediate (Fig. S2). To assign the putative age of a given module, we asked at what time point in evolution the majority of its components had appeared (based on the parsimony analysis). Modules for which 70% or more of the OGs appeared in a single age category were assigned to that category; modules whose genes appeared in several categories, without a clear majority, were assigned into the category *undefined*.

Scoring scheme for quantifying cohesiveness of modules

The evolutionary history of each module is represented by two parameters: the summed cost of the most parsimonious evolutionary events (total cost) and the relative fraction of events affecting more than one gene at a time (fraction of joined events). The cost measure is higher for larger modules (simply because they contain more genes), so we also introduced a normalized measure which corrects for module size (normalized total cost, i.e., total cost divided by the number of OGs in the module). Because the currently available set of completely sequenced genomes is highly biased (e.g., for pathogens), and because some areas of phylogeny have very few completed genome sequences while others have very many, it is difficult to directly assess the significance of any given cohesiveness measurement. We, therefore, performed a randomization of the modules, in order to estimate the likelihood of observing a particular cohesive behavior by chance. For each module size class, 10^7 random modules were generated by randomizing the OGs membership in modules (we simply populated the modules by repeatedly drawing, with replacement, from the total set of OGs; this approach is very conservative as we did not shuffle the species profiles of the OGs

themselves, nor did we change the size distribution of OGs). We recorded the distribution of the two cohesiveness measures for the randomized data, by plotting their densities in a two-dimensional plane defined by both measures. A visual inspection of this density showed that, with the exception of modules of size two and three, the density can be modeled with high accuracy by the two dimensional multivariate normal distribution (Fig. S1) where $x1$ is the normalized total cost and $x2$ the fraction of joined events:

$$f(x1,x2) = \frac{1}{2\pi\sigma1\sigma2\sqrt{1-\rho^2}} \exp\left[\frac{-1}{2(1-\rho^2)}\left[\frac{(x1-\mu1)^2}{\sigma1^2} + \frac{(x2-\mu2)^2}{\sigma2^2} - 2\rho\frac{(x1-\mu1)(x2-\mu2)}{\sigma1\sigma2}\right]\right]$$

The center of the covariance ellipse ($\mu1$, $\mu2$), angle (ρ) and the standard deviations ($\sigma1$, $\sigma2$) were adjusted independently for each module size (see Table S1) by minimizing the sum of the difference between the density values and the values predicted by the function for points of the evolutionary history parameters using the quasi-Newton method. For modules of size two and three, we directly used the empirical density where available and only used the approximation function (multivariate distribution) for areas in which the density was zero. For modules of this small size, our measurements of cohesiveness are probably underestimated, because small modules may, to some extent, appear cohesive by pure chance. This will be rectified as soon as more fully sequenced genomes become available—currently our cohesiveness estimate is a lower limit of the actual cohesiveness.

The evolutionary cohesiveness of a real module is then calculated using the following cumulative equation, where N is the module size and $x1$ and $x2$ are values of total cost and fraction of joined events whose density is lower than the one of the module.

$$Score = \int_{x1}^{\infty}\int_{x2}^{\infty} f_N(x1,x2)dx1dx2$$

Acknowledgments

We thank members of the Bork group for helpful discussions, in particular Jan Korbel for invaluable input and suggestions, Eoghan Harrington for comments on the manuscript, and Tobias Doerks and Sean Hooper for technical assistance. We also thank three referees for their constructive comments on the manuscript. M.C. is a recipient of a FEBS long-term fellowship. The work was supported by the EU grants LSHG-CT-2004-503568 and LSHG-CT-2003-503265.

References

Blanchette, M., Kunisawa, T., and Sankoff, D. 1999. Gene order breakpoint evidence in animal mitochondrial phylogeny. *J. Mol. Evol.* **49:** 193–203.

Ciccarelli, F.D., von Mering, C., Suyama, M., Harrington, E.D., Izaurralde, E., and Bork, P. 2005. Complex genomic rearrangements lead to novel primate gene function. *Genome Res.* **15:** 343–351.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. 1998. Conservation of gene order: A fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23:** 324–328.

Date, S.V. and Marcotte, E.M. 2003. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol.* **21:** 1055–1062.

Daubin, V., Moran, N.A., and Ochman, H. 2003. Phylogenetics and the cohesion of bacterial genomes. *Science* **301:** 829–832.

Ettema, T., van der Oost, J., and Huynen, M. 2001. Modularity in the gain and loss of genes: Applications for function prediction. *Trends Genet.* **17:** 485–487.

Ettema, T.J., Huynen, M.A., de Vos, W.M., and van der Oost, J. 2003. TRASH: a novel metal-binding domain predicted to be involved in heavy-metal sensing, trafficking and resistance. *Trends Biochem. Sci.* **28:** 170–173.

Fitch, W.M. 1971. Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.* **20:** 406–416.

Galan, J.E. and Collmer, A. 1999. Type III secretion machines: Bacterial devices for protein delivery into host cells. *Science* **284:** 1322–1328.

Hartigan, J.A. 1973. Minimum mutation fits to a given tree. *Biometrics* **29:** 53–65.

Hausner, M. and Wuertz, S. 1999. High rates of conjugation in bacterial biofilms as determined by quantitative in situ analysis. *Appl. Environ. Microbiol.* **65:** 3710–3713.

Huynen, M.A. and Bork, P. 1998. Measuring genome evolution. *Proc. Natl. Acad. Sci.* **95:** 5849–5856.

Itoh, T., Takemoto, K., Mori, H., and Gojobori, T. 1999. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol. Biol. Evol.* **16:** 332–346.

Johnson, G.R., Jain, R.K., and Spain, J.C. 2002. Origins of the 2,4-dinitrotoluene pathway. *J. Bacteriol.* **184:** 4219–4232.

Konstantinidis, K.T. and Tiedje, J.M. 2004. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc. Natl. Acad. Sci.* **101:** 3160–3165.

Korbel, J.O., Doerks, T., Jensen, L.J., Perez-Iratxeta, C., Kaczanowski, S., Hooper, S.D., Andrade, M.A., and Bork, P. 2005. Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol.* **3:** e134.

Kunin, V. and Ouzounis, C.A. 2003. The balance of driving forces during genome evolution in prokaryotes. *Genome Res.* **13:** 1589–1594.

Lathe III, W.C., Snel, B., and Bork, P. 2000. Gene context conservation of a higher order than operons. *Trends Biochem. Sci.* **25:** 474–479.

Lawrence, J.G. 1997. Selfish operons and speciation by gene transfer. *Trends Microbiol.* **5:** 355–359.

Macnab, R.M. 1999. The bacterial flagellum: Reversible rotary propellor and type III export apparatus. *J. Bacteriol.* **181:** 7149–7153.

Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., and Eisenberg, D. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature* **402:** 83–86.

Martin, M.J., Herrero, J., Mateos, A., and Dopazo, J. 2003. Comparing bacterial genomes through conservation profiles. *Genome Res.* **13:** 991–998.

Mirkin, B.G., Fenner, T.I., Galperin, M.Y., and Koonin, E.V. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* **3:** 2.

Mushegian, A.R. and Koonin, E.V. 1996. Gene order is not conserved in bacterial evolution. *Trends Genet.* **12:** 289–290.

Nakamura, Y., Itoh, T., Matsuda, H., and Gojobori, T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat. Genet.* **36:** 760–766.

Nguyen, L., Paulsen, I.T., Tchieu, J., Hueck, C.J., and Saier Jr., M.H. 2000. Phylogenetic analyses of the constituents of Type III protein secretion systems. *J. Mol. Microbiol. Biotechnol.* **2:** 125–144.

Papp, B., Pal, C., and Hurst, L.D. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424:** 194–197.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci.* **96:** 4285–4288.

Pellegrini, M., Thompson, M., Fierro, J., and Bowers, P. 2001. Computational method to assign microbial genes to pathways. *J. Cell. Biochem. Suppl.* **Suppl 37:** 106–109.

Peregrin-Alvarez, J.M., Tsoka, S., and Ouzounis, C.A. 2003. The phylogenetic extent of metabolic enzymes and pathways. *Genome Res.* **13:** 422–427.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., and Barabasi, A.L. 2002. Hierarchical organization of modularity in metabolic networks. *Science* **297:** 1551–1555.

Rivera, M.C., Jain, R., Moore, J.E., and Lake, J.A. 1998. Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci.* **95:** 6239–6244.

Rives, A.W. and Galitski, T. 2003. Modular organization of cellular networks. *Proc. Natl. Acad. Sci.* **100:** 1128–1133.

Rutter, M.T. and Zufall, R.A. 2004. Pathway length and evolutionary constraint in amino acid biosynthesis. *J. Mol. Evol.* **58:** 218–224.

Siefert, J.L., Martin, K.A., Abdi, F., Widger, W.R., and Fox, G.E. 1997. Conserved gene clusters in bacterial genomes provide further

support for the primacy of RNA. *J. Mol. Evol.* **45:** 467–472.
Smejkal, C.W., Vallaeys, T., Seymour, F.A., Burton, S.K., and Lappin-Scott, H.M. 2001. Characterization of (*R/S*)-mecoprop [2-(2-methyl-4-chlorophenoxy) propionic acid]-degrading *Alcaligenes* sp. CS1 and *Ralstonia* sp. CS2 isolated from agricultural soils. *Environ. Microbiol.* **3:** 288–293.
Snel, B. and Huynen, M.A. 2004. Quantifying modularity in the evolution of biomolecular systems. *Genome Res.* **14:** 391–397.
Snel, B., Bork, P., and Huynen, M.A. 1999. Genome phylogeny based on gene content. *Nat. Genet.* **21:** 108–110.
———. 2002a. Genomes in flux: The evolution of archaeal and proteobacterial gene content. *Genome Res.* **12:** 17–25.
———. 2002b. The identification of functional modules from the genomic association of genes. *Proc. Natl. Acad. Sci.* **99:** 5890–5895.
Tanaka, T., Tateno, Y., and Gojobori, T. 2005. Evolution of vitamin B6 (pyridoxine) metabolism by gain and loss of genes. *Mol. Biol. Evol.* **22:** 243–250.
Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. 2001. The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29:** 22–28.
van der Meer, J.R., Werlen, C., Nishino, S.F., and Spain, J.C. 1998. Evolution of a pathway for chlorobenzene metabolism leads to natural attenuation in contaminated groundwater. *Appl. Environ. Microbiol.* **64:** 4185–4193.
van Nimwegen, E. 2003. Scaling laws in the functional content of genomes. *Trends Genet.* **19:** 479–484.
Veitia, R.A. 2002. Exploring the etiology of haploinsufficiency. *Bioessays* **24:** 175–184.
von Mering, C., Zdobnov, E.M., Tsoka, S., Ciccarelli, F.D., Pereira-Leal, J.B., Ouzounis, C.A., and Bork, P. 2003a. Genome evolution reveals biochemical networks and functional modules. *Proc. Natl. Acad. Sci.* **100:** 15428–15433.
von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. 2003b. STRING: A database of predicted functional associations between proteins. *Nucleic Acids Res.* **31:** 258–261.
von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., and Bork, P. 2005. STRING: Known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **33:** D433–D437.
Wolf, Y.I., Rogozin, I.B., Kondrashov, A.S., and Koonin, E.V. 2001. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.* **11:** 356–372.
Wu, J., Kasif, S., and DeLisi, C. 2003. Identification of functional links between genes using phylogenetic profiles. *Bioinformatics* **19:** 1524–1530.
Yang, J., Lusk, R., and Li, W.H. 2003. Organismal complexity, protein complexity, and gene duplicability. *Proc. Natl. Acad. Sci.* **100:** 15661–15665.

Received June 24, 2005; accepted in revised form November 30, 2005.