Phylosystemics: Merging Phylogenomics, Systems Biology, and Ecology to Study Evolution

a K Watson, M Habib, E Bapteste

CellPress
REVIEWS

Opinion

Phylosystemics: Merging Phylogenomics, Systems Biology, and Ecology to Study Evolution

A.K. Watson,[1] M. Habib,[2] and E. Bapteste[1,*]

We define phylosystemics, a multidisciplinary strategy uniting short timescale interaction studies from systems biologists and ecologists with the longer timescale studies familiar to evolutionary biologists, taking advantage of methods from network sciences. Phylosystemics superimposes evolutionary information on entities/edges forming interaction networks produced by systems biology and ecology. At the molecular level, phylosystemics could provide evidence to infer and to time the evolution of molecular processes within a single branch of a phylogeny, in particular between the first and last common ancestors of a group arising during a major evolutionary transition. At the ecosystemic level, phylosystemics could culminate with the development of multilayer temporal networks encompassing biotic and abiotic interactions, whose analyses could unravel ecological interactions with evolutionary consequences.

Highlights

To fully understand biological interactions and evolution, they must be studied together.

There are few generic models and approaches able to unite the short timescales of interactions unraveled by systems biologists and ecologists, with the longer timescale of evolution.

We propose an approach to incorporate short timescale interaction networks with longer timescale evolutionary studies, by superimposing evolutionary information on interaction networks, which we name phylosystemics.

Applications of phylosystemics have the potential to enhance knowledge of major evolutionary transitions and the evolution of ecosystems.

Interaction Networks in Biological Studies

If nothing in biology makes sense except in the light of evolution, nothing in the biological world exists in isolation. At all levels of biological organization, from molecules to ecosystems, interactions shaping biological organization and processes at a given time contribute to further evolutionary dynamics. Therefore, interactions and evolution must be studied together.

Interaction networks (see Glossary) are commonly used in diverse biological studies (e.g., transcriptomics, proteomics, metagenomics, microbiology, protistology, developmental biology, ecology, and systems biology), and thus are of interest to many biologists. These interaction networks model organizations and processes by representing a diversity of interactions by edges between nodes corresponding to biotic and sometimes to abiotic components [1]. Because such interaction networks feature short-timescale dynamics, their typical questions are not primarily questions related to evolutionary biology.

For example, **cellular gene coexpression networks (GCNs)** are undirected graphs composed of nodes and edges, which represent genes and mutual coexpression relationships, respectively [2]. GCNs are used for various purposes, including the identification of regulatory genes, candidate disease gene prioritization, and functional gene annotation [3], occasionally strengthened by the conservation of a given coexpression profile across several species [4,5]. Other examples: **gene regulatory networks (GRNs)**, and in particular transcriptional regulatory networks, consist of nodes representing regulatory components, typically transcription factors, which are connected to regulated target genes, themselves under the influence of the DNA-binding sites of the transcription factors (and some works include kinases and additional layers of regulation of transcription factors by noncoding RNAs [6]). GRN studies firstly seek to understand how cells respond to internal and external condition changes. **Protein–protein interaction networks (PPIs)** illustrate yet another kind of interaction network. PPIs connect proteins (nodes) that interact with edges, and are mostly used to decipher cellular processes [7]. Similarly, at a larger spatial scale, **co-occurrence networks (CNs)** represent individual microbes [**operational taxonomic units (OTUs)**, or sequences] as nodes, connected by edges when significant correlations are found in the distributions of these taxa across samples. CNs are currently inferred for a wide range of microbial communities, such as soils, oceans, or hosts. CNs are used to investigate various types of ecological interaction (parasitism, mutualism, etc.) between taxa [8], to determine their drivers [9], and to gain insights into the organization of communities, notably by identifying keystone species and modules with niche-specific communities [10–12]. All of these networks are also commonly studied by systems biologists or ecologists to identify their principles of biological organization [1,10–12]. Here, we propose (i) to formalize phylosystemics as an

[1]Institut de Systématique, Evolution, Biodiversité (ISYEB), Sorbonne Université, CNRS, Museum National d'Histoire Naturelle, EPHE, Université des Antilles, Paris, France

[2]CNRS-IRIF, Paris University, Paris, France

*Correspondence: eric.bapteste@upmc.fr

 https://doi.org/10.1016/j.tim.2019.10.011

emergent field aiming to track the evolution of biological processes by general approaches of graph theory, then (ii) we demonstrate that phylosystemics is timely and actionable, and (iii) illustrate some major pay-offs expected from this approach.

Interaction Networks in Evolutionary Inferences

There are few generic models and approaches able to unite the short timescales of interactions, unraveled by systems biologists and ecologists, with the longest timescale of evolution familiar to evolutionary biologists. However, a novel interdisciplinary approach, that we call phylosystemics, can generalize evolutionary analyses of interactions using networks, and is arguably already in the making. Phylosystemics is comprised of two terms: 'phylo-', which means 'lineage' (and for this reason echoes with the established fields of phylogenetics and phylogenomics, possibly attracting evolutionary biologists with this particular background), and '-systemics', which stands for 'systems' (which echoes with systems biology and could attract systems biologists and ecologists toward this new field).

More precisely, phylosystemics consists of adding evolutionary knowledge (typically, but not only, from orthology and homology) onto nodes of interaction networks (Box 1) to analyze the evolution of processes involving biological entities (type 1 phylosystemics), including the evolution of evolutionary processes (type 2 phylosystemics). Importantly, phylosystemics can be applied throughout all scales of biological organization because interaction networks are now constructed at all scales, from the molecular level to the ecosystemic level. Therefore, phylosystemics provides a general framework to tackle multiple, distinct questions related to the evolution of life on Earth. To do so in a generalized way, phylosystemics can exploit the information contained in increasingly abundant networks, by focusing on two (complementary) problems of graph theory.

First, phylosystemics can use the formalism of a general covering problem to identify subgraphs in which nodes share specific evolutionary labels within individual interaction networks. The covering approach was notably pioneered by Qin *et al*. [13]. As early as 2003, these authors used an unorthodox type of phylogenetic profile for orthologous proteins called 'isotemporal categories', describing the distribution of orthologs across six broad taxonomic categories (of which only some were monophyletic.). Qin *et al*. mapped these unorthodox evolutionary labels onto the yeast protein interaction network in order to test for the presence of preferential connections between nodes with similar

Box 1. Evolutionary Labelling of Interaction Networks

In molecular networks, typically, each node of the network represents a molecule (e.g., a gene, a protein) that can be associated with a gene or protein family, for example, by sequence similarity network (SSN) analyses [63]. This clustering produces a first evolutionary label, for each node of interaction networks, for example, its belonging to 'homologous family *x*'. Next, a relative dating of the origin of each homologous family can be achieved. This first requires testing whether the gene family has been laterally transferred. There are many approaches for LGT detection, including quick SSN-based approaches [64] and state-of-the-art maximum likelihood (ML) phylogenetic trees and network approaches. The identification of LGT events for these families provides additional evolutionary labels for the nodes of the interaction networks: transferability of the family, and duplicability of the family, since homologous genes present in multiple copies in one species or lineage, but not as a result of LGT, can be further distinguished as (in/out)-paralogs.

Combining the definition of homologous families with the LGT detection approach allows one to associate the origin of each gene family with a certain phylogenetic depth given a reference species tree. Typically, for gene families unaffected by LGT, this 'dating' is often realized with the Dollo parsimony approach implemented in Count, using an accepted reference species tree at the time of the analysis, as a backbone. By contrast, gene families affected by LGT for which donors and hosts can be inferred should be dated separately after careful human inspection to remove possible confounding effects on their phylogenetic depth (i.e., to avoid considering a taxonomically widespread family as phylogenetically old, when some distant lineages have in fact recently acquired the genes via LGT).

Glossary

Betweenness: a centrality measure for a node in a graph. In the normalized form, this is the proportion of shortest paths between all possible pairs of nodes in a connected component that pass through this node. A betweenness close to 1 is indicative of a central gene, whereas close to 0 is more peripheral.
Co-occurrence networks (CNs): represent individual microbes (OTUs or sequences) as nodes, connected by edges when significant correlations are found in the distributions of these taxa across samples.
Constraint satisfaction problem (CSP): Any problem that can be formalized using a finite number of variables, each of them having a finite set of values (its domain) and a finite set of constraints within these variables (e.g., 'being different', 'having the same parity'). A Sudoku puzzle is a simple example in which a set of variables (empty squares) that store solutions can be filled using a set of possible discrete values (the numbers 1–9), where the choice of solution for a variable is based on a given set of limitations (the rules of the game, numbers in predetermined squares).
Degree: the number of edges connected to a given node in a graph. In a directed graph these this can be separated to in-degree and out-degree, for edges directed into the node or out of the node respectively.
Endosymbiont gene transfer (EGT): the transfer of genes from an endosymbiont to its host genome.
Gene coexpression networks (GCNs): are undirected graphs composed of nodes and edges, which represent genes and mutual coexpression relationships, respectively.
Gene regulatory networks (GRNs): consist of nodes representing regulatory components.
Interaction networks: networks representing a diversity of interactions by edges between nodes corresponding to biotic and sometimes to abiotic components, in order to model organizations and processes.
ITSNTS hypothesis: the 'It's the song, not the singers' hypothesis

phylogenetic profiles – that were expected if the network had recorded some evolutionary signals rather than evolved via ahistorical connections. Qin *et al.* then used these isotemporal category labels to retrace the evolution of the yeast PPI network since the last universal common ancestor (LUCA) as a series of temporal stages of additions of clusters of connected nodes. They polarized these stages by assuming that taxonomically broader profiles indicated sets of the more ancient nodes connected by the more ancient edges, according to the accepted phylogeny of species at the time of their analysis. They concluded that the growth pattern of the yeast PPI network had registered a series of major evolutionary events, in particular the endosymbiotic origins of eukaryotes.

Second, phylosystemics can use the general concept of graph homomorphism to identify evolutionary conserved subgraphs, corresponding to shared biological processes, found across sets of interaction networks (Figure 1 and Box 2).

By our definition, phylosystemics expands over two other kinds of network studies related to evolution. Firstly, phylosystemics encompasses studies that focused on the impact of general evolutionary processes on interaction network evolution (e.g., how gene duplication affects the structure of gene regulatory networks, etc. [14,15]). Such studies typically belong to type 2 phylosystemics, of which Wittkopp *et al.* [16] represents an early example. These authors investigated the general evolutionary processes of *cis*- and *trans*-regulatory changes involved in the evolution of divergent gene expression in two closely related species of *Drosophila*. Accordingly, their work provided general insights into the evolution of gene regulatory networks, rather than explaining the evolution of specific biological pathways. Likewise, in 2005, He and Zhang produced another notable type 2 phylosystemics analysis [17]. It consisted of mapping evolutionary information (paralogy) and functional information onto the yeast PPI network to analyze the topological distribution of these labels in order to test which of three possible general models of molecular evolution (subfunctionalization, neofunctionalization, or subneofunctionalization) would better explain the yeast PPI network topology. He and Zhang concluded that the general process of subneofunctionalization better accounted for these data. Whereas this early work described a general evolutionary process, it did not track the evolution of specific biological processes/pathways performed by the yeast proteins, nor did it study the distribution of these biological processes and their resulting phenotypes across the tree of life. More recently, in 2017, Yang and Wittkopp reported that the architecture of regulatory networks, in particular the connectivity of genes in transcriptional regulatory networks, influenced regulatory evolution in *Drosophila* [18]. This study focused on centralities, comparing gene **in-degree** distributions and gene **out-degree** distributions across the networks of closely related species, in order to uncover general principles governing the evolution of interactions. It concluded that genes regulated by larger numbers of transcription factors tend to have fewer and smaller changes in expression both within and between *Drosophila* species than genes regulated by smaller numbers of transcription factors. Likewise, yet uncommonly ambitious in terms of taxonomy for a phylosystemic type 2 study, Zitnik *et al.* investigated the evolution of resilience in 1840 protein interactomes across the tree of life. They used strategies of randomized node removals, followed by analyses of the extent of the resulting network fragmentation [19]. They concluded that interactomes become more resilient (more robust to network failures) over evolutionary time, and reported correlations between this resilience and ecological properties of the organisms under study. Zitnik *et al.* also mapped orthology labels on the protein interactomes to compare the local connectivity of orthologous pairs of proteins across networks (the connectivity between neighbor nodes of a focal node, using a 2-hop subnetwork). They found that protein **neighborhoods** around orthologs rewire and become increasingly different as the evolutionary distance between species increases, typically becoming more interconnected in the species having undergone more genetic changes since their last common ancestors.

Even though type 2 phylosystemic studies are increasingly popular, they remain far less often performed than phylogenomics studies. Moreover, it is important to stress that, by definition, the scope of phylosystemics, which is comprised of both type 1 and type 2 studies, is broader than the identification of general processes of network evolution as a result of molecular evolution. Phylosystemics intends to track the evolution of all sorts of biological processes modelled by interaction networks,

proposed by Doolittle and Inkpen suggests that patterns of interaction (songs) can be selected as units, even when the entities in interaction (singers) change.
Lateral gene transfer (LGT): the transfer of genes between genetic entities.
Neighborhood: the neighborhood of a node in a graph is the subgraph induced by all vertices adjacent to that node.
Operational taxonomic units (OTUs): A group of inidivudals that are similar based on chosen criteria, not nessicarily limited to taxonomic categories (e.g., organisms clustered based on small subunit 16s rRNA sequence similarity above 97% identity).
Protein–protein interaction networks (PPIs): connect proteins (nodes) that interact by edges and are mostly used to decipher cellular processes.

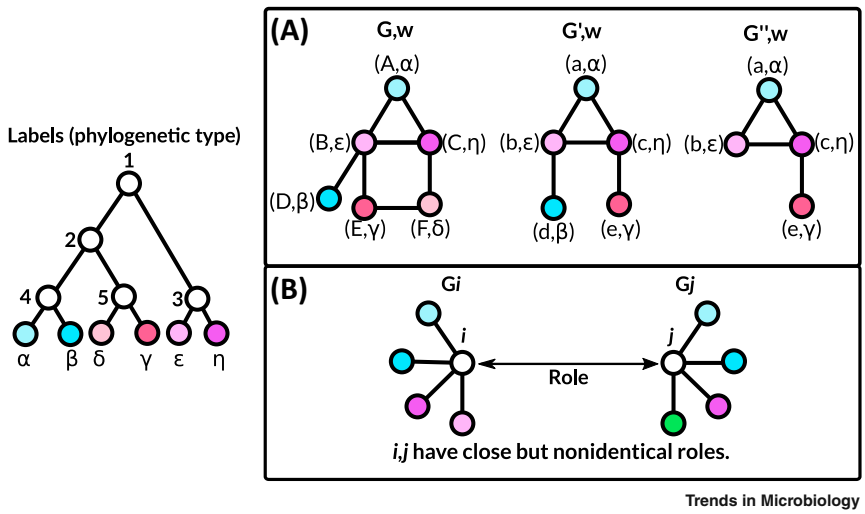


Figure 1. Matching in Graph Comparison.

(A) Detecting common maximal subgraphs using a hierarchical type of label, while minimizing phylogenetic distance. Evolutionary labels are at the left. Graphs G, G′ and G″, which could be extracts of co-occurrence networks from microbial communities, are boxed at the right, with node names indicated in Latin letters, and associated phylogenetic labels indicated by Greek letters. The distance between two phylogenetic types, for instance α and β, is called w(α,β) and is defined as the number of branches (or the sum of the branch lengths) between α and β on the reference phylogenetic tree. The concept of subtype can be used to identify common subgraphs of different graphs (sets or subsets of identical connections between G*i* and G*j*) based on node labels. G cannot be embedded into G′, but two subgraphs with a distance less than a fixed k = 3 in the label phylogeny: G(H1) and G(H2) subgraphs of G, with H1 = (A, B, C, D) and H2 = (A, B, C, F), can be embedded into G′. With '→' indicating mapping between nodes in G and G′: (i) Mapping for G(H1) shows a distance of zero to G′ because: for nodes (A → a) w(α, α) = 0, nodes (B → b) w(ε, ε) = 0, nodes (C → c) w(η, η) = 0, and nodes (D → d) w(β, β) = 0, summing to a total distance of zero. (ii) Mapping for G(H2) shows a distance of two to G′ because: for nodes (A → a) w(α, α) = 0, nodes (B → b) w(ε, ε) = 0, nodes (C → c) w(η, η) = 0, and nodes (F → e) w(δ, γ) = 2, summing to a total distance of two. Because H1 minimizes the phylogenetic distance between subgraphs, it will be preferred to H2. In that toy example, an isomorph co-occurrence pattern is detected between microbial communities, described by G and G′. The best embedding of G in G″ is found for G(H3) with H3 = (A, B, C, F), although not a perfect embedding because it shows a phylogenetic distance of two. (iii) Mapping for G(H3) shows a distance of two to G″ because: for nodes (A → a) w(α, α) = 0, nodes (B → b) w(ε, ε) = 0, nodes (C → c) w(η, η) = 0, and nodes (F→ e) w(δ, γ) = 2, summing to a total distance of two. (B) The concept of role in graph comparison. This concept can be applied to co-occurrence networks, where nodes represent operational taxonomic units (OTUs), and are labelled based on their taxonomy. The concept of a role can identify pairs of nodes in distinct networks, and define the proportion of their shared neighbors with identical evolutionary labels. In G*i*, node *i* has neighbors with lilac, purple, blue, and cyan labels, while in G*j* node *j* exchanges the lilac neighbor for a green neighbor, meaning that nodes *i* and *j* have a similar but nonidentical role. The total number of neighbors does not affect this comparison, so the removal of the lilac node would also mean nodes *i* and *j* had similar but nonidentical roles. This concept can be applied to any kind of labelled graph, including functional labels.

whereas clearly 'general processes of network evolution' (type 2) constitute a subset of the biological processes that evolved on Earth.

Furthermore, phylosystemics differs from a recent series of remarkable studies that used evolutionary inferences to improve the construction of interaction networks. For instance, Castro *et al.* (2019) introduced a learning approach for joint network inference using closely related datasets (i.e., gene regulatory networks from bacteria on the one hand and from yeasts on the other hand) to improve the reconstruction of gene regulatory networks [20]. Their strategy was based on the reasoning that gene regulatory networks are in part composed of evolutionarily shared subnetworks, conserved across datasets. Thus, evolutionary conservation was typically and primarily a means towards better

Box 2. Using the Constraint Satisfaction Problem (CSP)

CSP identifies matchings (similar sets of nodes and edges) between pairs of graphs (see Figure 1 in main text). Nodes harbor labels which can be numbers (e.g., % of duplicated genes or of transferred genes within a gene family), or also types (e.g., homology, orthology, paralogy classes, taxa). Types means that nodes can be equipped with a partial order or simply a hierarchy, as for example with a phylogeny. The main operation is graph homomorphism, which solves a typical problem: given two labelled graphs G and G′, and a phylogenetic distance w(α,β), one must find a homomorphism from G to G′ compatible with the labels, that is, a mapping θ from the nodes of V(G) to V(G′) satisfying:

(i) If x y is an edge in G, then θ(x) θ(y) is an edge of G′

(ii) w(x,y) less than or equal to a given threshold k

If the edges are weighted (e.g., by the relative binding strength of a transcription factor in the case of GRNs), weights should be preserved by θ.

The same approach applies for the discovery of subgraphs: given two labelled graphs (G, w), (G′, w′), CSP can find an homomorphism from a subgraph of G into a subgraph of G′, by maximizing some criterium such as the size (e.g., number of shared labelled nodes in the subgraph), or minimizing some distance (e.g., the phylogenetic distance between subgraphs, or the difference in edge weights).

Increasingly general types of matching can be applied. First, the concept of role [65] can identify pairs of nodes in distinct networks (one node in G*i* of species *i*, another node in G*j* of species *j*), and define, for this pair of nodes, the proportion of their shared neighbors presenting identical evolutionary labels. A role score reaches 1 when node *i* in G*i* and node *j* in G*j* are surrounded by sets of nodes from the same gene families. Computing the role scores for all pairs of nodes for all pairs of evolutionarily labelled GCN and PPI can readily identify conserved interactions, such as pairs of nodes from two networks with high role scores, which significance can be assessed by separate random shuffling of the evolutionary labels in each of the compared networks. Importantly, nodes with the same role could themselves not be homologous: node *i* in G*i* could belong to a different gene family than node *j* in G*j*, if there has been a nonhomologous replacement of that node in one of the species. Thus, significantly high role scores can also be used to identify nonhomologous replacements in interaction networks from different species. Analyses of directed graphs, such as GRN, may warrant specific local optimization techniques, typically a node in graph *i* surrounded by similar neighbors than another node in graph *j* will not necessarily constitute a pair of nodes with similar roles, if the edges connecting these nodes to their neighbors present different rather than identical directions in G*i* and G*j*.

More broadly, the notion of subtype defines cases where labelled patterns of G*j* are nested into or equivalent to labelled patterns of G*i*. Subtypes can be used to identify common subgraphs (sets or subsets of identical connections between G*i* and G*j*), for any pairs of evolutionarily labelled GCN, GRN, and PPI, respectively. Importantly, CSP approaches can be used not only to detect isomorphisms (subgraphs with identical patterns of interactions), but also homomorphisms (subgraphs with significantly similar patterns of interactions and sets of nodes) between networks. Focusing on homomorphisms rather than isomorphisms can be especially useful when local rewiring has affected the topology of a network around focal nodes (as is the case for example, when rapid *cis*-regulatory evolution rewires network structures across short evolutionary time scales). Hence, subtypes identify identical or partly identical processes, and some of these subgraphs may correspond to functional units [66], especially for phylogenetically broadly conserved subgraphs in agreement with [67–70], etc. Using CSP formalism is realistic because it is very practical to deal with graph homomorphisms, as shown, for instance, in applications to chemistry [71]. Software dealing with CSP, thanks to efficient filtering algorithms, is available (e.g., Ilog at IBM). Moreover, good implementations (e.g., Cogitant: http://cogitant.sourceforge.net.) are also available for such analyses on bipartite networks [39]. Another advantage of CSP is the ability to reason on the set of all equivalent solutions, for example, by identifying common points shared between these solutions that may be more robust, or to emphasize the sources of variation between solutions.

network construction, but not, in itself, an object of study. Evolutionary biology hypotheses were used to enhance inferences of systems biology. Such studies might be considered by some as a third type of phylosystemics (type 3 studies), but according to us their goals (improving networks) is still too contrasted with that of phylosystemics (tracking the evolution of biological processes) to necessarily fall under the same general umbrella. For the time being, phylosystemics appears complementary to these network approaches, but, if further developed, it might, with them, achieve a virtuous circle

by importing knowledge from systems biology to enhance evolutionary biology. The door for such a fruitful cooperation is probably starting to open. For example, the work by Koch *et al*. conjugated the three kinds of researches on interaction networks and evolutionary studies [21]. It aimed at introducing multispecies regulatory network learning, using phylogenetic information and a probabilistic graphical model to improve regulatory network inference from transcriptomics data (type 3 phylosystemics study). Yet, Koch *et al*. also compared the predicted regulatory networks from different species to identify general properties of network evolution (type 2 phylosystemics), unravelling correlations between gene duplication with edge losses and edge gains rates. Finally, Koch *et al*. performed a type 1 phylosystemics study on the evolution of stress-specific regulatory networks along the phylogeny of six yeast species. Specifically, they identified regulators with evolutionarily conserved roles, featuring as conserved hubs in the most repressed and most induced modules. We believe it is time to encourage these ambitious approaches at all evolutionary scales by recognizing phylosystemics as a new field, whose operationality will next be demonstrated.

Phylosystemics Analyses Are Actionable

Phylosystemic Analyses Are Timely

The data are there. Moreover, available interaction networks appear suited even for macro- and megascale evolutionary analyses. Macroevolutionary comparisons of gene expression become more realistic every day, since large-scale expression data are increasingly available for a diversity of species. Because expression profiles only cover a subset of all possible cellular conditions [4], at first sight, gene coexpression networks generated for different species and for different conditions might seem difficult to exploit in a comparative analysis [4]. However, early analyses of moderately and distantly related organisms have already shown the opposite [4,5]. First, coexpression of functionally linked genes is often conserved among organisms, as highly connected genes tend to be essential and conserved [4,22]. Second, GCNs are modular. Most of the relations between modules vary between organisms, and the more recently evolved modules are associated with organism-specific functions [4]. Altogether, the above properties of GCNs open the possibility of a sequential reconstruction of ancient GCNs for different phylogenetic depths, one module at a time, based on distinct sets of partial (high quality) GCNs. The same is true for PPIs, as they are also modular [23]. They largely evolve by duplication and divergence, with very slow evolutionary rates for protein–protein interactions (2.6 ± 1.6 changes $\times 10^{-10}$ per PPI per year) [24], which preserves clues of ancient interactions between proteins (since some of these ancient interactions will still be observable for some paralogs, at least). These properties encourage sequential inferences of ancient PPIs [13,23] and graph comparisons of PPIs [7], even at a large evolutionary timescale, especially in eukaryotes. Likewise, macro- and megaevolutionary analyses of gene regulatory networks appear promising, for other reasons. Although GRNs (typically that of *Escherichia coli*) [25] present a highly interconnected, nested structure, complicating the identification of functional modules, and although GRN analyses usually focus on motifs (significantly over-represented local network architectures) [25,26], evolutionarily conserved regulatory mechanisms have been described [27]. Inferences of robust yet necessarily partial ancient GRNs appear possible.

Interestingly, all of these types of interaction network present variations, even between the Domains of life [6,28,29], strongly suggesting that interaction networks contain some ancient evolutionary evidence about the history of life. For example, the global architectures of PPIs and GCNs seem to be universal across life forms, whereas their local network structures are much less constrained, and differ even among closely related organisms [30]. The local wiring of GCNs would be constrained by selection, whereas their global properties are not, which would reflect stochastic, nonselective processes [30,31]. For this reason, phylosystemics can focus on local architectures, typically identified by homomorphism and graph-covering approaches. Similarly, at the ecological scale, some evolutionary signal seems to be reflected in microbial interactions between taxa [11]. Co-occurrence networks, often defined at various phylogenetic depths, appear to be shaped by biotic and abiotic factors [8,32,33]. Environmental factors alone, while historically considered as very strong determinants of community structure (as famously stated in the Baas-Becking theory from 1934: 'everything is everywhere, but the environment selects'), now appear to be incomplete predictors of community structures [32]. Consistently, it has been proposed that phylogeny shapes some aspects of the global plankton interactome [32], and that, in nature, co-occurring taxa are more closely

related than expected by chance [10,11]. Thus, CNs also likely encompass some underexploited evolutionary information.

In addition, all the bioinformatic tools needed to implement this general strategy are currently available, therefore the evolutionary labelling of interaction networks is operational. Critical analyses are possible because confidence metrics are commonly associated with edges of interaction networks. For instance, RegulonDB [34] provides levels of confidence associated with its gene regulatory networks, and STRING [35] provides confidence scores associated with PPIs, etc. These metrics are often used to identify the most reliable inferences. For example, Castro *et al.* considered edges in GRNs as valid within a 0.5 precision cut-off [20], Yang and Wittkopp considered the presence or absence of statistically significant [false discovery rate (FDR) = 0.05] differences in gene expression [18], etc.

Graph Covering Allows Separate Network Analyses for Super-Interaction Network Inferences

The topology of evolutionarily labelled networks can be analyzed, in particular to infer past interaction networks and specific types of change in these networks. As in Qin *et al.* [13], yet using phylogenetic assignations matching clades, a given interaction network, for example, a gene regulatory network from a given species (even when partial, but high-quality) can be subjected to a decomposition into rough temporal slices. This decomposition is achieved by (i) attributing a relative phylogenetic depth to all nodes (e.g., the node will receive the age of apparition inferred for its homologous family), and then (ii) by assuming that nodes of a similar phylogenetic depth correspond to entities that, if directly connected in present day networks, may have also interacted as early as the time at which these entities appeared together [36,37]. This strategy will define, for each interaction network, more or less disconnected phylogenetically dated subgraphs representing putative interactions at given phylogenetic depths.

This first decomposition strategy has the potential to infer ancestral molecular interaction networks at various evolutionary depths (e.g., in ancestral fungi) in a sequential fashion by aggregating inferences separately generated from the networks of all available species (e.g., in various species of fungi). In the same spirit that super-trees or super-networks are built in phylogenomics by aggregating phylogenetic information from independent phylogenetic markers, for example, aggregating information from gene family trees which may not all overlap with each other in terms of their host taxa, phylosystemics proposes to aggregate inferences gathered from independent interaction networks to infer super-interaction networks. This aggregative process goes in two steps. First, interactions between nodes of individual networks are assigned a phylogenetic depth, based on the phylogenetic distribution of their components. For example, a set of interacting genes reported in the yeast PPI, with detectable homologs exclusively present in all fungal lineages, and for which no **lateral gene transfer (LGT)/endosymbiotic gene transfer (EGT)** is suspected, would be considered as encoding (a portion of) a process potentially present in ancestral fungi.

Second, interactions between nodes independently inferred to be of the same phylogenetic depth in separate analyses of interaction networks are aggregated to produce super-interaction networks for that phylogenetic depth. This approach to super-interaction networks is trivial, when one considers that networks are lists of edges that can be concatenated, which provides further support (i.e., the number of independent interaction networks recovering a particular interaction) for each edge in the super-interaction network.

Of note, in the case of prokaryotic molecular networks, LGT will impact the topology of interaction networks and complicate inferences, either by introducing groups of interacting entities (promoters, genes, proteins), therefore grafting subgraphs within host interaction networks, or by introducing single genes. At small evolutionary scales, such LGT may thus be detectable through network comparisons. At a larger evolutionary scale, eventually for cross-domain network comparisons, components and edges of interaction networks originated from LGT will not be immediately distinguishable from components and edges evolving vertically within a host lineage, if the evolutionary labels used in

phylosystemics are limited to homology between network components. In cases of undetected LGT between Archaea and Bacteria, a naive network comparison for species from these two Domains might lead to the belief that some biological processes are shared and interpreted as if these processes were ancestral, even though the shared network topology would here reflect LGT between these lineages. Precisely because LGT is expected to complicate phylosystemic analyses, in cases involving prokaryotic networks a good practice is to use additional evolutionary labels, namely where nodes correspond to transferred entities. Standard phylogenomic approaches should be used first to identify LGT, to map the LGT labels onto interaction networks to distinguish nodes and edges that may be affected by LGT. Indeed, interaction networks can be labelled with a diversity of labels reflecting all *a priori* knowledge on evolutionary processes affecting their components rather than only homology/orthology information.

Homomorphism Approaches for Phylosystemic Comparative Analyses

The sequential inference of ancestral networks can be further completed and tested via a second approach of graph comparison, namely homomorphism analyses, taking advantage of recent development from the **constraint satisfaction problem (CSP** [38]), or a more general concept of bipartite graph comparison, the notion of subtype, developed in CSP [39] (Box 2). Common local network patterns identified in comparisons may describe interactions inherited from the last common ancestor of the compared species. Such inferences based on conserved sets of nodes and edges across interaction networks are likely to be robust to missing data, in the sense that they will point to real, even if partial, processes, since the patterns are observed in multiple species. Moreover, the statistical significance of the subgraphs can be tested by randomly shuffling labels in the compared networks, to implement a *P* value, associated with a particular shared subgraph in the two networks. Let us now explain what the payoff of these bioinformatically realistic approaches might be.

Phylosystemics Can Enhance Knowledge in Evolution

To support our call for phylosystemics, let us develop two theoretical examples of the unique potential of ambitious phylosystemics studies, conducted at a taxonomically broad level, and exploiting both homology and orthology relationships. Simply put, we propose to embrace phylosystemics and apply it, systematically, using the above kinds of interaction network as evidence *per se* to tackle major evolutionary questions such as the evolution of: major groups, endosymbioses, turn-over of protein content in organelles, and more generally of the evolution of protein interactions within the cells, of community structures, of emergent metabolic functions and of ecosystems (see Outstanding Questions). In this way information from systems biology would be systematically put at the service of evolutionary studies addressing major evolutionary issues.

Improved Understanding of Major Evolutionary Transitions and Endosymbioses

Phylosystemic analysis of molecular interaction networks could provide new, alternative evidence to address simple, yet pressing, issues: what molecular processes evolved along with new major groups (e.g., from the first eukaryotic common ancestor, FECA, to the last eukaryotic common ancestor, LECA [40])? What extant lineages are more ancestral-like in terms of their constitutive processes? None of these questions can be readily answered by phylogenetic analyses alone, whereas phylosystemics can provide a partial inference of the cellular processes present in the LUCA or in the progenotes, in early eukaryotes, Archaea, Bacteria, and in major groups within these three Domains.

Consider eukaryogenesis. The phylogenetic dating of interaction networks offers a way to infer the local architecture of *early* eukaryotic GCNs, GRNs and PPIs by focusing on nodes and edges found in the last common ancestors of eukaryotes, including gene families that may have been contributed by the ancestral endosymbiotic bacteria and their archaeal hosts. Phylogenomics alone is not informative about the temporal order in which the different early eukaryotic specific genes evolved during eukaryogenesis: their relative timing of appearance is not decidable, since such genes may have appeared as early as FECA and as late as LECA, or anytime in between [40–42]. Yet, phylosystemic analysis provides a new strategy to address this relative timing issue, which could propose a relative dating of the time at which early lineage-specific genes evolved, under a certainly simple yet

apparently reasonable assumption [4,43–45]. On average, older genes have had more time than younger genes to form novel interactions with molecular partners, and therefore older genes might be more highly connected than younger genes (with similar functions) in early eukaryotic GCNs, GRNs, and PPIs. Under the testable logic that, for nodes with similar phylogenetic depth (and especially for genes with similar functions), higher degree nodes preceded nodes with less neighbors in interaction networks in time, phylosystemics could be tried out to propose original hypotheses that phylogeny alone could not support [40,41]: the relative ordering of appearance of processes 'within' a single (e.g., basal) branch of the tree, using the topology of the interaction network as a time machine (Figure 2, Key Figure). Thus, centrality analyses (degrees, etc.) of eukaryotic genes (exclusive to eukaryotes, inherited from archaea, or inherited from bacteria) in interaction networks could be used to enlighten changes in molecular processes during the FECA to LECA transition. Because there are fewer core gene families than there are ancient gene families, and because interactions are retained in some lineages but can be lost in other lineages, separate phylosystemic analyses are likely to identify the largest set of early nodes and interactions. Typically, gene families conserved in 'unikonts' and 'bikonts' (or 'opimoda' and 'diphoda') [46] could be traced back to the branch from FECA to LECA even if these families are not still present in all eukaryotic taxa. Similar approaches could be conducted on the poorly known transitions from the first to the last archaeal common ancestors (the FACA to LACA transition [42,47]) and from the first to the last bacterial common ancestors (the FBCA to LBCA transition) [42] to propose a relative timing of evolution of the exclusive archaeal/bacterial genes and processes, respectively, although for prokaryotes great care must be taken to avoid false positives due to LGT.

Key Figure

Phylosystemic Study of the Origins of New Groups

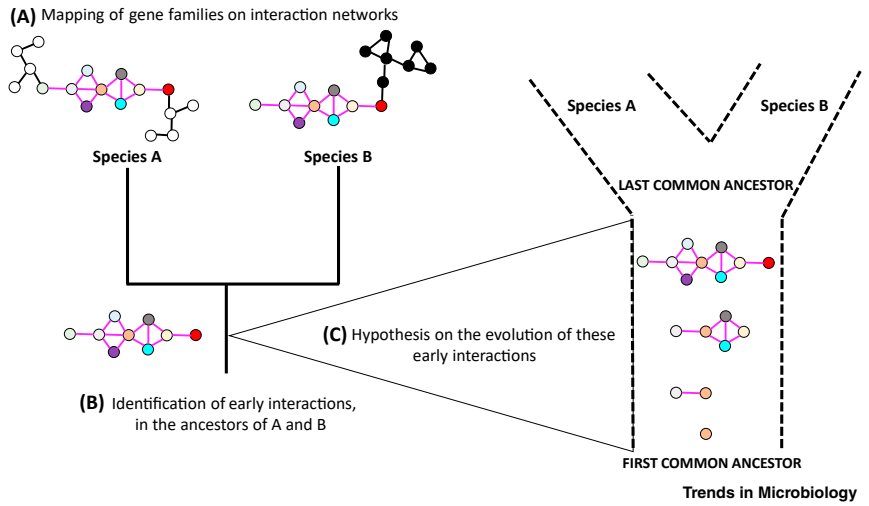


Figure 2. Three typical steps of a phylosystemic analysis, oriented towards the inference of the evolution of early processes at the base of a clade. (A) Left panel. Top part, interaction networks of two related species, with nodes labelled with an identical color for homologous entities (e.g., genes), or left white or black otherwise. Black edges correspond to connections exclusive to each species. Pink edges indicate shared interactions between homologous families. (B) Bottom part, a phylogenetic logic would predict that the shared architecture, identified using algorithms to find the maximal common labelled subgraph, dates from the base of the (A,B) clade. (C) Right panel. Phylosystemics proposes to refine this inference based on hypotheses considering topological information from the interaction networks. Here, for example, nodes with higher degrees (in actual networks) are considered to be more ancient than nodes with lower degree, which allows us to decompose the evolution of the processes between the first and last common ancestors of the clades.

Once the ancestral interaction networks are inferred, using subtype analyses for pairs of networks within each Domain of life, phylosystemics can determine what extant major group/taxa hosts the highest proportion of ancient interactions inherited from the last common ancestor of its group. Interaction networks from extant taxa can be sorted by their proportions of preserved ancient interactions, using the proportion of 'early nodes + edges' these contemporary networks embed. This original phylosystemic measure thus quantifies the similarity of processes within a taxon with the processes inferred to have been present in its common ancestor(s). For example, this measure can show what extant groups retained the highest proportion of ancient interactions within each Domain of life. Looking ancient does not mean being ancient, but phylosystemics can easily single out taxa with processes that are largely conserved. It can thus offer a new kind of evidence on the nature (in terms of processes) of first and last common ancestors, for example, by identifying the most LUCA/progenote-like of the extant taxa, which could give new insights to the lasting debates on the nature of the last common ancestors of cells ([48–52] and references therein), or by identifying the most FECA/LECA-like extant archaeal group, providing alternative evidence in a currently hot issue [40,41]. For example, searching for subtypes between inferred early eukaryotic interaction networks (GCN/GRN or PPI) and the interaction networks from contemporary archaea would show what archaeal lineages (e.g., the Asgard lineage) harbor the significantly highest proportion of inferred early eukaryotic interactions.

Another exemplar application of phylosystemics relates to endosymbiosis studies. Typically, phylogenomicists agree that primary, secondary, and tertiary endosymbioses impact the gene content of nuclear eukaryotic genomes as a result of EGT (from the genomes of the endosymbionts to the nuclear genomes of their hosts [53]). Yet, they fiercely debate on the amount of additional laterally transferred genes that may have made their way inside eukaryotic genomes [54–57]. Phylosystemics can provide independent evidence on the reality and the biological consequences of proposed EGT/LGT in protists since it can describe how candidate EGT/LGT genes and their products functionally integrate in their host cells. For example, in case of negative correlation between transferability and node degree, within a given functional category, ancient gene families may typically have higher degrees [4,43–45] in the interaction networks of their eukaryotic host species than gene families recently acquired by long distance LGT. Therefore, phylosystemics provides an additional test to distinguish some bona fide LGT genes (expected to show significantly lower degrees) from ancient, repeatedly lost, gene families (expected to show significantly higher degrees) in evolutionarily labelled interaction networks from protists. Overall, phylosystemics appears as a relevant strategy to provide more evidence for or against introgressive events in eukaryotes.

Thus, phylosystemics could provide novel ways to sort out events of molecular change within a given branch of the species phylogeny, identify early-looking taxa, and, in protists, validate some LGT/EGT (detected as significantly underconnected nodes with homology to bacterial genes/proteins).

Improved Understanding of Ecosystems Evolution, while Embracing Diversity

To paraphrase a famous biological saying, '*Omnis reticulum e reticulum*' (all networks arise from other networks). Indeed, all contemporary biological organizations and processes certainly succeeded to former organizations and processes (a constraint called 'contingency of network evolution on the pre-existing network structure' by [30], or 'string of historical events' by [13]). Yet, so far, evolutionary biologists have very few methods to formally model this fundamental intuition, and to account for the contribution of a diversity of modes of interactions to explain the evolution of biodiversity. Phylosystemic analyses at the ecosystem level could unravel ecological interactions with strong evolutionary relevance.

Consider microbial interactions described by co-occurrence networks with OTUs as nodes, inferred from time series, and eventually from ecological transects from a diversity of environments (soils, hosts, waters). A CN can be evolutionarily labelled by adding the taxonomy of the sequences, represented by its nodes. CNs produced (i) from replicates at different times for a given environment (microdynamics studies), or (ii) for different environments (soil, lakes, etc.), can be compared to identify (i)

nodes with similar roles and (ii) subtypes corresponding to subgraphs with similar architectures of microbial interaction between CNs. Such comparisons would detect stable patterns of interactions within the extant biodiversity [58] (Figure 3).

Moreover, phylosystemic studies could enhance these descriptions by looking for potentially ancient stable interactions, providing each CN is transformed by aggregating the nodes corresponding to given taxa (e.g., a yeast and a *Penicillium* species) into metanodes corresponding to broader taxonomic categories (e.g., fungi). Transforming a contemporary original CN into an evolutionary pooled CN allows one to make an iterative use of CSP approaches to compare the induced CN generated across various levels of the taxonomy (from species to phyla) to identify, for each level of the taxonomy, (i) metanodes with similar roles and (ii) subgraphs of metanodes with similar architectures between evolutionary pooled CNs from different samples (Figure 3). This strategy would typically detect recurring types of syntrophic associations, for example, between distinct methanogenic archaeal lineages and distinct sulfoxidizing bacterial lineages in different environments. Phylosystemics could thus unravel microbial interactions which may be homologous across environments, ancient/persistent kinds of interactions, and determine what spatially separated communities on the planet share the highest proportions of such homologous interactions, and in that regard rely upon 'related' structures. Furthermore, phylosystemics could help to better understand the dynamics of microbial interactions over increasingly large evolutionary timescales, by detecting nodes and metanodes with significantly high role scores (Box 2) between CNs from different environments. Such a comparison would show whether alternative microbial interactions arose by the replacement of some taxa by others (e.g., if Oomycetes replaced fungi in some environmental interactions).

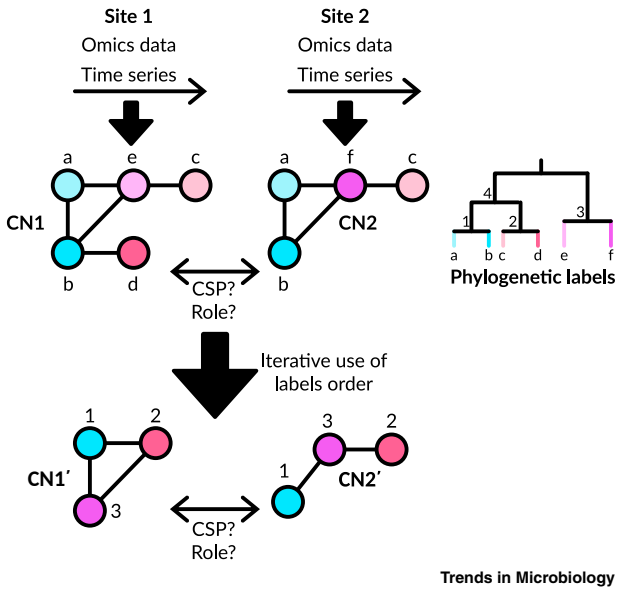
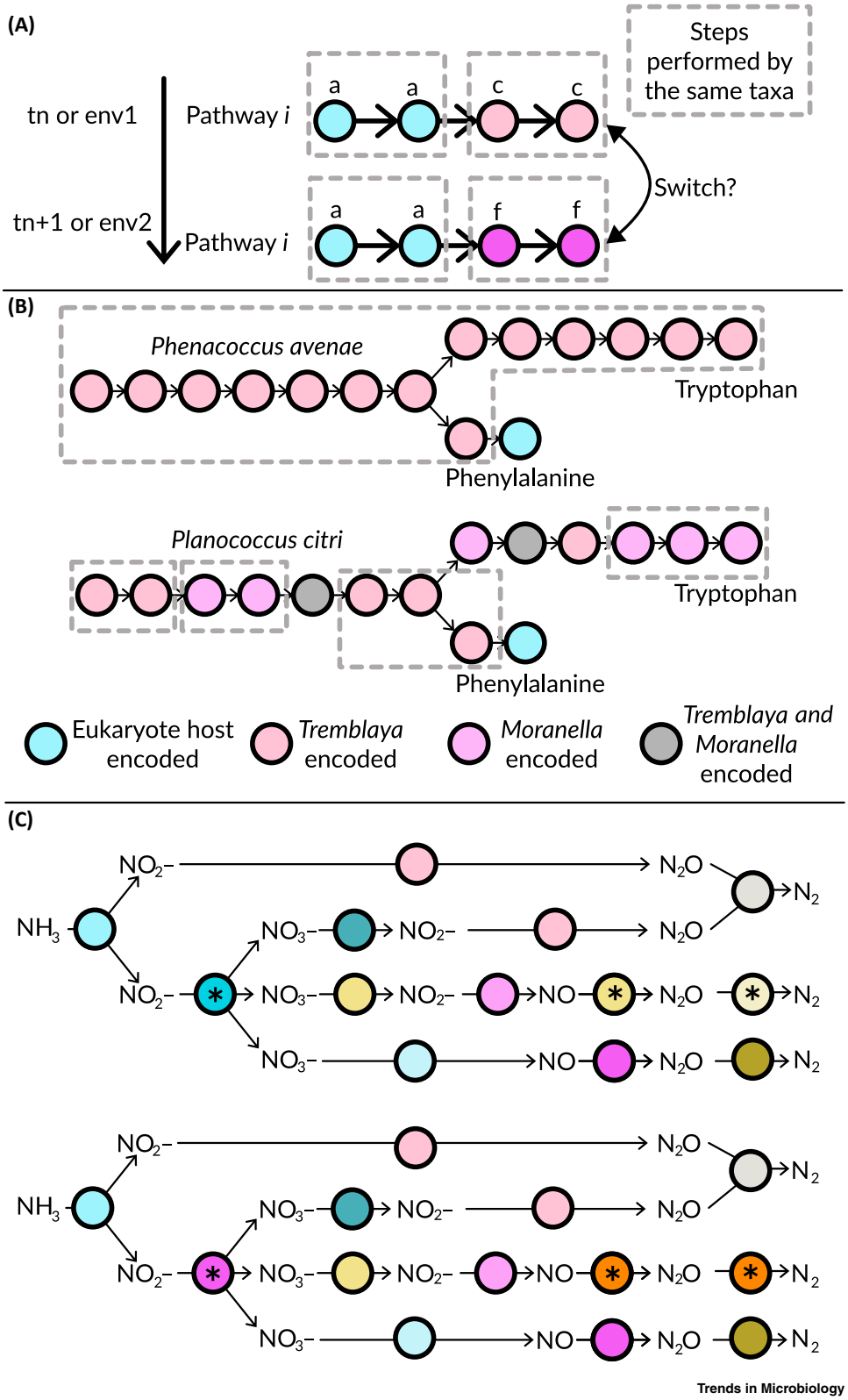


Figure 3. Iterative Aggregation of Nodes in Co-occurrence Networks for Constraint Satisfaction Problem (CSP)-Based Comparison.

In this example, co-occurrence networks (CN)1 and CN2 are built from omics data gathered at two different sites. Nodes are operational taxonomic units (OTUs) and edges represent significant co-occurrence of these OTUs at each site. These networks can be compared using CSP and role scores. In the initial comparison of CN1 and CN2, the common subgraph is 'a-b', indicating that these two related labels (e.g., two fungal species) interact at both sites, but no nodes share the same role. Aggregating the nodes based on taxonomy generates the induced networks CN1′ and CN2′. The induced networks can also be compared using CSP and role scores. At this taxonomic depth, node '3', the internal node on the branch to labels e and f, has the same role in CN1′ and CN2′ and '1-3-2' represents a common subgraph between CN1′ and CN2′, indicative of a conserved interaction at a deeper phylogenetic level.



(Figure legend at the bottom of the next page)

More fundamentally, phylosystemics could shed light upon the evolution of metabolic networks and biogeochemical cycles by representing the abundance of reads/enzymes/host species associated with particular KEGG pathways in a given host or environment. For example, when different taxa alternatively encode different steps of the reactions, as in the case for essential amino acid pathways in the symbioses between *Tremblaya PCIT* and *Moranella endobia* within mealybugs [59], or nitrogen fixation in microbial communities by metabolic hand-off [60], phylosystemics can detect candidate patterns of metabolic complementarity between microbial lineages by decomposing taxonomically labelled metabolic pathways into monochromatic connected components (Figure 4). If a given pathway is encoded by the same taxa, this operation of decomposition produces one connected component per pathway, otherwise that decomposition produces multiple connected components. This latter type of pattern is of major, fundamental interest since it suggests instances where the collective function performed by the metabolic pathway or by the geochemical cycles may be under selection, in agreement with the **ITSNTS hypothesis** [61]. The proposed approach aims at analyzing the topology of the network, not its quantitative behavior. Thus, CSP analyses of labelled metabolic networks could show, for each environment for which samples from different time points or spatial locations are available, whether and which (parts of) pathways persist (since enzymes are present), whereas taxa encoding their genes change [61]. These network comparisons would typically detect switches, that is, subgraphs corresponding to nodes and edges for which the (relative) abundance of taxa encoding the genes significantly change in the community over time (or over space for ecological transects). Additionally, phylosystemic analyses of the centralities (degree, eccentricity, **betweenness**, and presence on a cycle) in such labelled metabolic networks can show whether nodes with particular centralities are associated with one or multiple host taxa (e.g., whether reactions corresponding to high betweenness nodes in the metabolic networks are highly constrained or potentially interchangeable in terms of taxa).

Concluding Remarks

Merging results and methods from systems biology and from ecology with those of phylogenomics appears timely and promising. We propose to call the interdisciplinary outcome of such a merging 'phylosystemics'. Naming this field will hopefully be important for its development, as it can enhance the visibility of this research program into which a diversity of scientists can invest, ease scientific collaborations, and encourage the development of novel tools/approaches for a nascent community. Thus, in the late 1990s to early 2000s, the introduction of the term 'phylogenomics' supported the development of cutting-edge approaches to handle the wealth of molecular data, and probably contributed to a successful transition from phylogenetics, whose practices were traditionally centered on studies of single or a few gene families, towards broader, multimarker analyses. Phylosystemics aims at filling a function comparable with the

Figure 4. Patterns in Taxonomically Labelled Metabolic Networks.

(A) A theoretical example. In this network, nodes represent enzymes in a KEGG pathway *i*, labelled based on the most abundant taxa encoding these enzymes (e.g., a blue lineage, a green lineage, and a pink lineage). Evidence for metabolic complementarity between different lineages could be found when a single pathway includes steps requiring enzymes that are predominantly found in different taxa. A switch in label of nodes in this pathway at different sample sites or at different timepoints (e.g., from green at t_n to pink at t_{n+1}) could indicate a switch in the taxa predominantly performing this metabolic function, in agreement with the ITSNTS hypothesis, which proposes that genes, rather than host taxa, matter for the completion of a given pathway. (B) Taxa switching in the mealybug pathway for tryptophan and phenylalanine biosynthesis adapted from [59], with nodes representing enzymes in that metabolic pathway colored by their taxonomic origin. Upper. In *Planococcus avenue*, tryptophan and phenylalanine biosynthesis are predominantly encoded by the *Moranella endobia* endosymbiont. Lower. In *Phenacoccus citri* parts of the same pathway are replaced by a second endosymbiont, *Tremblaya PCIT*. (C) Environmental taxa switching in the denitrification pathway, adapted from [60,62], with nodes representing enzymes, and colors representing the predominant taxa inferred as encoding that enzyme in a given environment. Upper. Taxonomic distribution of enzymes in the denitrification pathway identified in aquifer metagenome samples in [62], where no single genome encoded a complete denitrification pathway, suggesting 'metabolic hand-off' between community members, adapted from [60]. Lower. Hypothetical second environment in which the denitrification pathway is also present, but different taxa encode the enzymes responsible for particular steps of the pathway (noted with *).

one that phylogenomics did in the past, because the prevalence of interaction networks shows that there is a common ontology throughout the biological world, and a common problem: understanding the origins and evolution of biological processes. As phylosystemics prolongs the transition from phylogenetics to phylogenomics, we hope that introducing this term will further encourage and generalize the development of analyses of biological processes through the combination of practices of comparative genomics and phylogenomics with that of systems biology and of ecology. However, the term 'evosystemics' may be preferred as an alternative since it would be a more inclusive name for the field studying the evolution of biological processes using networks.

Acknowledgments

We thank Professors Philippe Lopez, Debashish Bhattacharya, Christopher Lane, Laura Hug, Hervé Le Guyader, Dominique Higuet, and François-Joseph Lapointe for critical comments, and three anonymous reviewers for their constructive feedback. A.K.W. and E.B. were funded by the European Research Council (ERC) (FP7/2007-2013 Grant Agreement # 615274, category LS8).

References

1. Alon, U. (2006) *An Introduction to Systems Biology: Design Principles of Biological Circuits*, Chapman & Hall
2. Aoki, K. *et al.* (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant Cell Physiol.* 48, 381–390
3. van Dam, S. *et al.* (2017) Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinformat.* 19, 575–592
4. Bergmann, S. *et al.* (2004) Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol.* 2, E9
5. Stuart, J.M. *et al.* (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302, 249–255
6. Martinez-Pastor, M. *et al.* (2017) Transcriptional regulation in Archaea: from individual genes to global regulatory networks. *Annu. Rev. Genet.* 51, 143–170
7. Gerke, M. *et al.* (2007) Finding common protein interaction patterns across organisms. *Evol. Bioinform. Online* 2, 45–52
8. Faust, K. *et al.* (2015) Cross-biome comparison of microbial association networks. *Front. Microbiol.* 6, 1200
9. Weiss, S. *et al.* (2016) Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* 10, 1669–1681
10. Chaffron, S. *et al.* (2010) A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.* 20, 947–959
11. Faust, K. and Raes, J. (2012) Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* 10, 538–550
12. von Mering, C. *et al.* (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* 315, 1126–1130
13. Qin, H. *et al.* (2003) Evolution of the yeast protein interaction network. *Proc. Natl. Acad. Sci. U. S. A.* 100, 12820–12824
14. Price, M.N. *et al.* (2008) Horizontal gene transfer and the evolution of transcriptional regulation in *Escherichia coli*. *Genome Biol.* 9, R4
15. Waltman, P. *et al.* (2010) Multi-species integrative biclustering. *Genome Biol.* 11, R96
16. Wittkopp, P.J. *et al.* (2004) Evolutionary changes in cis and trans gene regulation. *Nature* 430, 85–88
17. He, X. and Zhang, J. (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169, 1157–1164
18. Yang, B. and Wittkopp, P.J. (2017) Structure of the transcriptional regulatory network correlates with regulatory divergence in *Drosophila*. *Mol. Biol. Evol.* 34, 1352–1362
19. Zitnik, M. *et al.* (2019) Evolution of resilience in protein interactomes across the tree of life. *Proc. Natl. Acad. Sci. U. S. A.* 116, 4426–4433
20. Castro, D.M. *et al.* (2019) Multi-study inference of regulatory networks for more accurate models of gene regulation. *PLoS Comput. Biol.* 15, e1006591
21. Koch, C. *et al.* (2017) Inference and evolutionary analysis of genome-scale regulatory networks in large phylogenies. *Cell Systems* 4, 543–558.e8
22. McCarroll, S.A. *et al.* (2004) Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat. Genet.* 36, 197–204
23. Jin, Y. *et al.* (2013) The evolutionary dynamics of protein–protein interaction networks inferred from the reconstruction of ancient networks. *PLoS One* 8, e58134
24. Qian, W. *et al.* (2011) Measuring the evolutionary rate of protein–protein interaction. *Proc. Natl. Acad. Sci. U. S. A.* 108, 8725–8730
25. Seshasayee, A.S.N. *et al.* (2006) Transcriptional regulatory networks in bacteria: from input signals to output responses. *Curr. Opin. Microbiol.* 9, 511–519
26. Shen-Orr, S.S. *et al.* (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* 31, 64–68
27. Ludwig, M.Z. *et al.* (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* 403, 564–567
28. Darnell, C.L. *et al.* (2017) Systematic discovery of archaeal transcription factor functions in regulatory networks through quantitative phenotyping analysis. *mSystems* 2, e00032-17.
29. Faria, J.P. *et al.* (2014) Genome-scale bacterial transcriptional regulatory networks: reconstruction and integrated analysis with metabolic models. *Brief Bioinformat.* 15, 592–611
30. Koonin, E.V. and Wolf, Y.I. (2010) Constraints and plasticity in genome and molecular-phenome evolution. *Nat. Rev. Genet.* 11, 487–498
31. Jordan, I.K. *et al.* (2008) Natural selection governs local, but not global, evolutionary gene coexpression networks in *Caenorhabditis elegans*. *BMC Systems Biol.* 2, 96
32. Lima-Mendez, G. *et al.* (2015) Ocean plankton. Determinants of community structure in the global plankton interactome. *Science* 348, 1262073

Outstanding Questions

Can we learn more about the origins of new groups, including the transitions from the first common ancestor and last common ancestor of these groups (e.g.,. the transitions from the first eukaryotic common ancestor to the last eukaryotic common ancestor, or those for other domains of life)?

What extant archaeal lineages are more similar, in terms of their processes, to the ancestral host cell that acquired a mitochondrial endosymbiont at the origin of eukaryotes?

In cases where genes have a patchy distribution, can we better distinguish between examples of recent LGT/EGT as compared with ancestral acquisition followed by gene losses?

Can we systematically identify examples of nonorthologous gene replacement, based on pairs of nonhomologous genes with similar roles in interaction networks?

Can microbial interactions be identified that are homologous across environments, and potentially ancient/persistent in nature?

Can spatially separate communities be identified that share high proportions of homologous interactions?

Is it possible to systematically identify evidence for metabolic complementarity between different lineages in communities, where a single pathway includes steps requiring enzymes that are predominantly found in different taxa?

Can we identify taxa that switched between communities based on their similar roles in interaction networks?

33. Raes, J. *et al.* (2011) Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol. Systems Biol.* 7, 473
34. Gama-Castro, S. *et al.* (2016) RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.* 44 (Database issue), D133–D143
35. von Mering, C. *et al.* (2005) STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* 33 (Database issue), D433–D437
36. Liang, C. *et al.* (2014) Network simulation reveals significant contribution of network motifs to the age-dependency of yeast protein-protein interaction networks. *Mol. Biosyst.* 10, 2277–2288
37. Ruprecht, C. *et al.* (2017) Phylogenomic analysis of gene co-expression networks reveals the evolution of functional modules. *Plant J.* 90, 447–465
38. Grohe, M. (2007) The complexity of homomorphism and constraint satisfaction problems seen from the other side. *J. A. C. M.* 54, 1–24
39. Chein, M. and Mugnier, M. (2009) *Graph-based Knowledge Representation – Computational Foundations of Conceptual Graphs*, Springer
40. Eme, L. *et al.* (2017) Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol.* 15, 711–723
41. Dacks, J.B. *et al.* (2016) The changing view of eukaryogenesis – fossils, cells, lineages and how they all come together. *J. Cell Sci.* 129, 3695–3703
42. Makarova, K.S. *et al.* (2005) Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. *Nucleic Acids Res.* 33, 4626–4638
43. Pawlowski, P.H. *et al.* (2013) A kinetic model of the evolution of a protein interaction network. *BMC Genom.* 14, 172
44. Peterson, G.J. *et al.* (2012) Simulated evolution of protein–protein interaction networks with realistic topology. *PLoS One* 7, e39052
45. Zhong, Q. *et al.* (2016) An inter-species protein–protein interaction network across vast evolutionary distance. *Mol. Syst. Biol.* 12, 865
46. Derelle, R. *et al.* (2015) Bacterial proteins pinpoint a single eukaryotic root. *Proc. Natl. Acad. Sci. U. S. A.* 112, E693–E699
47. Makarova, K.S. *et al.* (2015) Comparative genomic analysis of evolutionarily conserved but functionally uncharacterized membrane proteins in archaea: Prediction of novel components of secretion, membrane remodeling and glycosylation systems. *Biochimie* 118, 302–312
48. Giulio, M.D. (2011) The last universal common ancestor (LUCA) and the ancestors of Archaea and Bacteria were progenotes. *J. Mol. Evol.* 72, 119–126
49. Gogarten, J.P. and Deamer, D. (2016) Is LUCA a thermophilic progenote? *Nat. Microbiol.* 1, 16229
50. Koonin, E.V. and Martin, W. (2005) On the origin of genomes and cells within inorganic compartments. *Trends Genet.* 21, 647–654
51. Mat, W.K. *et al.* (2008) The genomics of LUCA. *Front. Biosci.* 13, 5605–5613
52. Weiss, M.C. *et al.* (2016) The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* 1, 16116
53. Timmis, J.N. *et al.* (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* 5, 123–135
54. Archibald, J.M. (2015) Evolution: gene transfer in complex cells. *Nature* 524, 423–424
55. Ku, C. *et al.* (2015) Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* 524, 427–432
56. Martin, W.F. (2018) Eukaryote lateral gene transfer is Lamarckian. *Nat. Ecol. Evol.* 2, 754
57. Roger, A.J. (2018) Reply to 'Eukaryote lateral gene transfer is Lamarckian. *Nat. Ecol. Evol.* 2, 755
58. Bapteste, E. and Huneman, P. (2018) Towards a dynamic interaction network of life to unify and expand the evolutionary theory. *BMC Biol.* 16, 56
59. Husnik, F. *et al.* (2013) Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell* 153, 1567–1578
60. Castelle, C.J. and Banfield, J.F. (2018) Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell* 172, 1181–1197
61. Doolittle, W.F. and Inkpen, S.A. (2018) Processes and patterns of interaction as units of selection: An introduction to ITSNTS thinking. *Proc. Natl. Acad. Sci. U. S. A.* 115, 4006–4014
62. Anantharaman, K. *et al.* (2016) Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* 7, 13219
63. Corel, E. *et al.* (2016) Network-thinking: graphs to analyze microbial complexity and evolution. *Trends Microbiol.* 24, 224–237
64. Corel, E. *et al.* (2017) Bipartite network analysis of gene sharings in the microbial world. *Mol. Biol. Evol.* 35, 899–913
65. Fiala, J. and Paulusma, D. (2005) A complete complexity classification of the role assignment problem. *Theoret. Comput. Sci.* 349, 67–81
66. Zhao, Y. and Mooney, S.D. (2012) Functional organization and its implication in evolution of the human protein–protein interaction network. *BMC Genom.* 13, 150
67. Masalia, R.R. *et al.* (2017) Connectivity in gene coexpression networks negatively correlates with rates of molecular evolution in flowering plants. *PLoS One* 12, e0182289
68. Netotea, S. *et al.* (2014) ComPlEx: conservation and divergence of co-expression networks in *A. thaliana, Populus and O. sativa. BMC Genom.* 15, 106
69. Kim, H.S. *et al.* (2006) MANET: tracing evolution of protein architecture in metabolic networks. *BMC Bioinformat.* 7, 351
70. Hase, T. *et al.* (2010) Difference in gene duplicability may explain the difference in overall structure of protein–protein interaction networks among eukaryotes. *BMC Evol. Biol.* 10, 358
71. Regin, J.-C. (1994) A filtering algorithm for constraints of difference in CSPs. Proceedings of the Twelfth National Conference on Artificial Intelligence (vol. 1) pp. 362–367, American Association for Artificial Intelligence