BioEssays
www.bioessays-journal.com

Inferring the Evolutionary History of Your Favorite Protein: A Guide for Molecular Biologists

*Jolien J.E. van Hooff, Eelco Tromer, Teunis J.P. van Dam, Geert J.P.L. Kops, and Berend Snel**

Comparative genomics has proven a fruitful approach to acquire many functional and evolutionary insights into core cellular processes. Here it is argued that in order to perform accurate and interesting comparative genomics, one first and foremost has to be able to recognize, postulate, and revise different evolutionary scenarios. After all, these studies lack a simple protocol, due to different proteins having different evolutionary dynamics and demanding different approaches. The authors here discuss this challenge from a practical (what are the observations?) and conceptual (how do these indicate a specific evolutionary scenario?) viewpoint, with the aim to guide investigators who want to analyze the evolution of their protein(s) of interest. By sharing how the authors draft, test, and update such a scenario and how it directs their investigations, the authors hope to illuminate how to execute molecular evolution studies and how to interpret them. Also see the video abstract here https://youtu.be/VCt3l2pbdbQ.

Dr. J. J. E. van Hooff, Dr. E. Tromer, Dr. T. J. P. van Dam, Dr. B. Snel
Theoretical Biology and Bioinformatics, Biology
Science Faculty
Utrecht University
Padualaan 8
3584 CH, Utrecht, The Netherlands
E-mail: b.snel@uu.nl

Dr. J. J. E. van Hooff, Dr. E. Tromer, Dr. G. J. P. L. Kops
Oncode Institute, Hubrecht Institute—KNAW (Royal Netherlands Academy of Arts and Sciences)
Uppsalalaan 8
3584 CT, Utrecht, The Netherlands

Dr. G. J. P. L. Kops
Molecular Cancer Research
University Medical Centre Utrecht
Heidelberglaan 100
3584 CX, Utrecht, The Netherlands

Dr. E. Tromer
Department of Biochemistry
University of Cambridge
Hopkins Building, Tennis Court Road
Cambridge CB2 1QW, UK

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/bies.201900006

DOI: 10.1002/bies.201900006

1. Introduction

Historically, comparative and evolutionary analysis of genes and proteins has been a versatile and useful approach to guide experiments that investigate the cellular role of human proteins. Typically, a protein's evolutionary interrogation answers questions like: What is the function of characterized homologs in other species, such as budding yeast? For example, the telomere function of the human protein Rap1 was predicted from its yeast ortholog.[1] This way, evolutionary analyses also contribute to human health, since they may characterize disease-associated proteins through their ortholog in model organisms such as *Caenorhabditis elegans* (*C. elegans*). For example, the worm ortholog of human protein KIAA0556, associated with Joubert syndrome, was shown to operate in the cilium,[2] allowing in-depth molecular characterization of this protein's cellular function. Another frequently asked question is which residues are conserved and therefore likely functionally important? The multiple sequence alignment (MSA) of S6 kinases uncovered the TOS motif, which is essential for TOR signaling,[3] a pathway that regulates cell growth and proliferation in response to nutrient availability. Some evolutionary analyses have more advanced applications (**Table 1**), like predicting protein function based on coevolution. Other analyses do not focus on protein function but generate findings that resonate in the field of molecular evolution, e.g., by revising the importance of events like horizontal gene transfer or gene fission.

Evolutionary analyses of protein sequences generally have the highest quality if directed manually, that is, by evaluating results obtained by automated algorithms (such as BLAST) and databases through in-depth visual inspection. In fact, the great majority of the examples we discuss here resulted from such manual analysis. Automated approaches often have certain biases that make them perform well only in specific cases.[25] Resources such as Pfam[26] are suitable for identifying homologs, not orthologs (see Box 2). Others, such as Orthologous MAtrix (OMA),[27] focus on identifying pairwise (orthology) relationships between two proteins, while many studies aim for collecting all proteins from an orthologous group (OG) (Box 2). Algorithms that infer such groups occasionally split a single group into multiple ones, due to sequence divergence and/or domain dynamics.[28,29] In general, automatic orthology and homology methods often do not suffice because proteins differ in their evolutionary dynamics and therefore require specific

Table 1. Evolutionary examination of a protein serves a variety of purposes.

Purpose	Example studies	Key challenges	Key observations
Function prediction of (human) proteins through projection	•The human protein Rap1 operates at telomeres[1] •KIAA0556, a Joubert syndrome-related protein, has a role in the cilium[2]	Finding (pairs of) orthologs (Box 2).	Orthologs have biologically equivalent functions.
Identifying conserved functional domains/motifs	•S6 kinase—TOS motif[3] •Nbs1/Ku80/ATRIP—ATM interaction motif[4]	Using correct sequences for the MSA.	Functional sequence motifs are often conserved patches in unconserved regions.
Understanding the function and evolution of biological features	•CenH3 and holocentricity[5] •Effector proteins, horizontal gene transfer and pathogenicity[6] •Ketone biosythesis and brain size evolution[7]	Retrieving information on biological (e.g., cellular) features.	Protein loss and gain (e.g., via horizontal gene transfer) correlate to feature loss and gain; when they anticorrelate new hypotheses are required.
Reconstructing evolution of function by ancestral sequence reconstruction	•GK_{PID}'s ability to bind Pins evolved via a single amino acid substitution[8]	Obtaining a high-quality gene tree.	New vs old protein functions can be distinguished based on sequence information.
Function prediction by coevolution or phylogenetic profiling	•BBS proteins are involved in the cilium[9] •MCU and MICU1 are ancient uniporter interactors[10]	Correct identification of presences as well as absences across species.	Nonclassical phylogenetic profiles predict analogy and bifunctionality.
Detecting covariance in a MSA to make or improve protein structure models	•ABCG5 and ABCG8 undergo conformational changes that affect their interaction sites[11]	Identifying homologous proteins with sufficient variance while still allowing for high-quality MSA.	Not only amino acid conservation, but also amino acid variation can be exploited for structural/functional insight.
Understanding functional divergence after gene duplication	•MadBub duplicates diverged into a Bub-like and Mad-like protein through reciprocal domain/motif loss[12,13]	Allowing for loss of domains in homology searches, obtaining a high-quality gene tree.	Duplicate subfunctionalization might be more common than neofunctionalization.
Predicting gene functions or cellular make-up of nonmodel organisms	•CAMSAP tracks the minus ends of microtubules in *Trichomonas vaginalis* and *Tetrahymena thermophila*[14]	Finding (pairs of) orthologs.	Orthologs have biologically equivalent functions.
Phylogenetically classifying proteins that are part of a large protein family	Kinesins,[15] histones,[16] ATPases,[17] myosins,[18,19] Rab GTPases[20]	Distinguishing different OGs and identifying ancient duplications.	Large-scale duplications before LECA.
Uncovering ancient origins of eukaryotic cellular features	Prokaryotic origins of various "eukaryotic signature proteins"[21–24]	Detecting remote homology.	Characteristic eukaryotic features are encoded by genes shared with Archaea.

BBS, Bardet–Biedl syndrome; MCU, mitochondrial calcium uniporter.

approaches. Combined with manual analysis, one can compose a customized approach for the protein of interest.[30] For example, refined homology searches combined with visual analysis aided researchers to identify the "missing" eukaryotic subunit B of TopoVI type II topoisomerase involved in meiotic recombination.[31,32] This and many other examples highlighted in this paper would not have generated their detailed insights if not partially carried out by hand.

Molecular biologists can greatly benefit from manually studying the evolution of their protein of interest, which is why we outline what we think are the most important principles of such analyses. Due to their intimate knowledge of proteins and cellular processes, these biologists can provide invaluable insights into evolutionary interpretations posed by evolutionary biologists. In collaboration with molecular biologists, we studied diverse eukaryotic proteins, varying from chromatin remodelers and bZIP transcription factors to flagellum and kinetochore components.[2,33–35] We experienced that for non-bioinformaticians various challenges arise regarding the use of tools, the knowledge of genome evolution and of the species tree, and, importantly, how complicated evolutionary histories are reflected in a myriad of computational tools and sequence databases. Such challenges may cause flawed evolutionary inferences, which in turn may cause incorrect, or incomplete, function prediction.[30] We here detail principles that explain "best practices" from the field, while we also draw attention to what we consider neglected. We argue that the most crucial yet overlooked skill is to be able to recognize, postulate, and revise different evolutionary scenarios, a process that is typically difficult to automate.

2. Inferring the Evolution of a Protein Means to Search for Its Orthologs

Although evolutionary analysis of a protein can be applied in different ways (Table 1), it typically involves answering these two main questions:

Box 1
What can happen to a protein during evolution?

Most evolutionary interrogations are focused on, or at least include, determining what happened to a protein during its evolution across different lineages. By "this protein," we here refer to the protein sequences that constitute an OG (see Box 2). The list of lineages and proteins that have an elevated event frequency is likely incomplete, since some eukaryotic lineages are much better genomically characterized than others.

Event	Definition	Frequency in eukaryotes compared to prokaryotes	Eukaryotic clades with elevated frequency	Proteins with elevated frequency
Genesis/de novo gene origin	Birth of a coding sequence from noncoding sequence	High	None, continuous, widespread	E.g., assembly factors, disease effectors
Duplication	Duplication of a whole or partial coding sequence, either part of a small-scale or of a whole-genome duplication	High	Land plants (whole-genome duplications), animals	Regulation (signal transduction, transcription factors), metabolic enzymes, host–pathogen interaction
Loss	Pseudogenization and/or elimination of coding sequence	Low	Parasitic lineages, obligatory symbionts	Genes with few interaction partners, lower expression, higher mutation rates
Horizontal gene transfer	Transfer of coding sequence whereby donor and acceptor are not in a parent–offspring relationship	Low	Fungi	Plant pathogenicity effectors, metabolic enzyme clusters
Fusion	Joining of two coding sequences into a single open reading frame	Higher than fission	Metazoa	Signal transduction
Sequence divergence	Mutations in coding sequence that alter the amino acid sequence	High	Intracellular parasites, *C. elegans*	Genes with fewer interaction partners, lower expression rates, host–pathogen interaction

1) When did this protein originate?
2) Which other current-day species have this protein?

These questions come down to asking which events occurred during the evolutionary history of a protein, such as origination, loss, and duplication (see Box 1). Note that, although these two main questions are seemingly straightforward, the devil is in the semantics. What do we mean with "origin," or with "this protein"? As illustrated below, we sometimes actually redefine the "origin" and "this protein" during the analysis.

Unraveling the evolutionary history of a protein entails searching and finding the same protein in other species. When doing so, we search for its orthologs. Orthologs are homologous proteins in two species that arose from a single protein in the common ancestor of those two species. For those not aware of the distinction between different types of homologs, we recommend Box 2. Why are we, and many others, often specifically interested in the orthologs, rather than all homologs? The answer is found in the "ortholog conjecture," which states that orthologs have biologically equivalent functions in different organisms, whereas paralogs have different functions due to divergence after gene duplication.[36,37] After all, if paralogs did not diverge functionally after gene duplication, they are redundant and therefore often one copy is lost.[38] As stated in Section 1, evolutionary analysis often does not limit itself to finding an ortholog of protein X from species A only in species B. To infer the presence of a protein across multiple species, one has to collect a set of proteins, termed "the orthologous group." Proteins constitute an OG if they descended from a single ancestral protein in the common ancestor of all species considered (see Box 2). For example, proteins form OG if they descended from a single protein in the last eukaryotic common ancestor (LECA). In this example, the OG was defined on the level of "eukaryotes," but one might want to choose a narrower level, like "animals" (single protein in the last common ancestor of animals), or a wider level, like "all cellular life" (single protein in the last universal common ancestor [LUCA]). See Box 2 for more details on the OG.

In the following sections, we exemplify protein evolutionary analysis by assuming we aim to unravel the evolution of a human protein involved in a eukaryotic cellular process. We assume we search for orthologs (the OG) in other eukaryotes and aim to infer the evolutionary events that happened since LECA. However, the same principles hold when studying, e.g., a budding yeast protein, or studying the evolution of a protein since the last common ancestor of animals. We assume readers have general knowledge of the eukaryotic species tree (Box 3), sufficient to use detailed resources like NCBI taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy) or the Tree of Life Web Project (http://tolweb.org/tree/).[39,48] Otherwise, we strongly advise to gather this knowledge. Also, we assume readers have some basic experience with computational analyses, including similarity searches with NCBI's BLASTP[39] or HMMER's phmmer and generating MSAs with tools like Clustal Omega at the EMBL–EBI web interface.[49,50] Such skills provide a solid base to execute the research strategies that we propose. We outline the initial steps of the analysis, their potential results, and how these results point to an evolutionary scenario. The latter will guide subsequent steps in

the analysis. We elaborate on how we iterate from initial results to a final scenario using different approaches, which we select based on the specific challenges posed by our protein.

3. A Quick (and Dirty) Guide to Inferring the Evolutionary History of a Protein

A standard evolutionary analysis ideally consists of searching for homologous sequences, aligning them, using the alignment to infer a gene tree, and interpreting this tree in light of the species tree to reconstruct the evolutionary history of the protein ("tree reconciliation"). While method sections in scientific papers read like this, the analysis often requires more steps and not-so-straightforward decisions, which are not explicitly reported. In order to shed light on this process, we here discuss how we start and proceed our analysis. In this process, we continuously take into account a scenario (or hypothesis or model) for the evolution of the protein of interest. We do not provide a guide on how to use specific bioinformatics tools, but aim to supply practical stepping stones and their conceptual underpinnings.

3.1. Collecting Observations to Generate an Initial Scenario

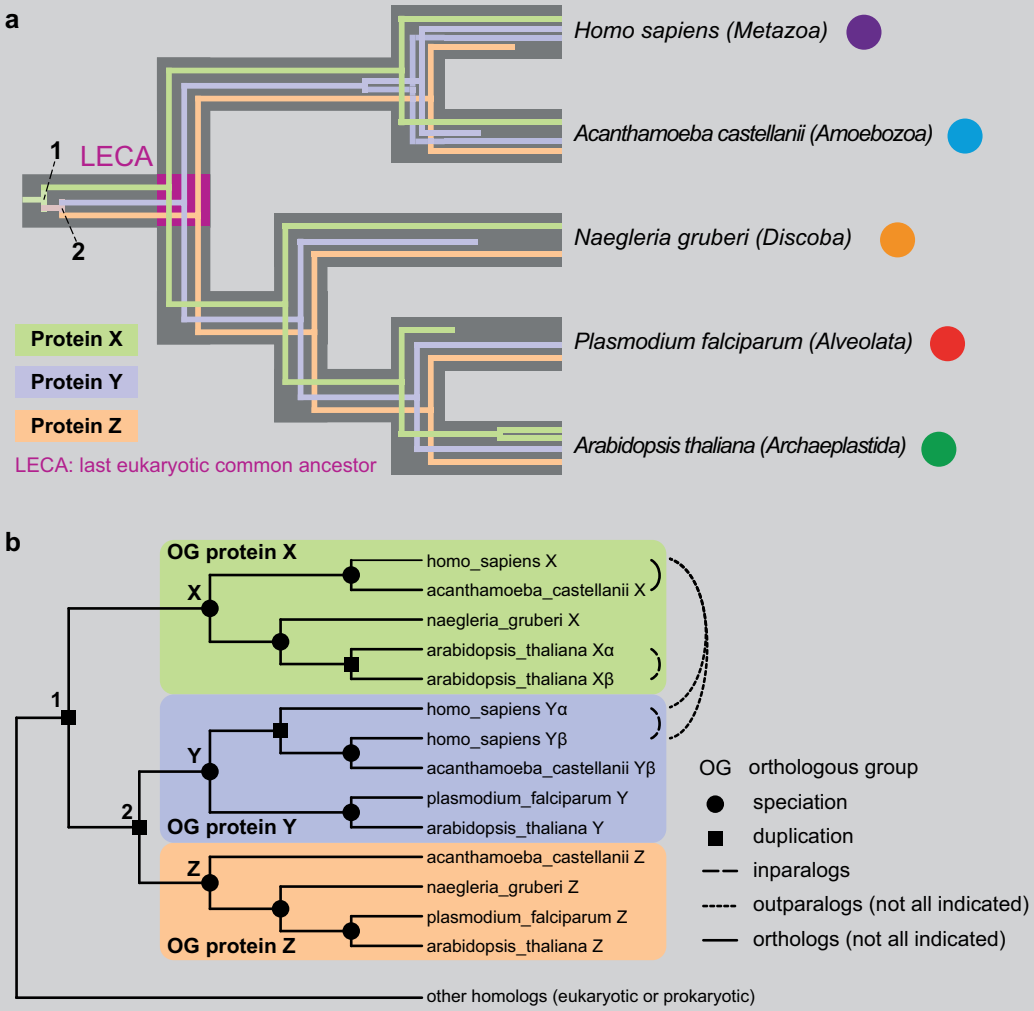
Inferring the evolutionary history of a protein starts with collecting some observations in order to draft the evolutionary scenario. Such a scenario should account for potential technical problems, which might prevent us from directly deducing the true evolutionary history of a protein from the results. We acquire the first basic observations about the protein of interest, such as its domain architecture, functional sequence motifs, and (candidate) orthologs from experiments in other species or precomputed comparative genomics databases, such as OrthoDB, InParanoid, PANTHER, or Ensembl Compara.[51–54] Then, we perform a few similarity searches with the protein sequence of interest ("the query") using BLASTP and phmmer[49,55] (**Figure 1**a), varying the significance cut-off and/or the search database. In order to obtain interpretable results, we recommend selecting a subset of species, e.g., 20 diverse eukaryotic species including human (Box 3). The hits we obtain are statistically likely homologous to our protein if these hits have a low *E* (expectation)-value (e.g., 0.001 or 0.0001; see also Section 4.2).[56] A low *E*-value allows one to reject the null hypothesis, which means that the query-hit alignment score has no biological significance. The homologous hits tell us 1) which organisms contain homologs and which do not, and 2) how many homologs are present in human (and other vertebrates or animals, if included in the search set). Together, the external information and these preliminary results enable us to draft an initial scenario for the evolutionary history of our protein. The initial scenario can broadly be categorized into four classes and directs the next steps in our examination.

3.2. Scenario 1 ("Easy"): Clear Initial Observations That Indicate a Simple Scenario

The first scenario is straightforward and comes with clear-cut observations. If the initial searches yield highly significant single hits across a range of species, these results indicate that the protein is relatively well conserved (Figure 1b). Moreover, the protein has a clear single point of origin: since likely no ancient duplications occurred, the protein can be safely inferred to have originated in or just before the last common ancestor of the species containing the homologs. As a result, all hits are by definition orthologs. If the protein is present in eukaryotes that are most distantly related to humans, such as *Plasmodium falciparum* (*P. falciparum*) or plants (see Box 3), this implies that it originated in or before LECA. If the protein is short (e.g.,<150 amino acids) or if its sequence divergence is somewhat high (as observed by the first searches), we often use a single iteration of an iterative searching tool like PSI-BLAST or jackhammer.[49,55] Such an iterative search may help in finding some additional

Box 2
Key definitions: homology, orthology, OGs, inparalogs, and outparalogs

Homology refers to an evolutionary relationship between two biological features. Features are homologous if they descent from a single common ancestral origin, regardless of whether they are morphological structures or molecules. Hence, proteins that do not share a common ancestor but that do share a function should not be referred to as "functional homologs": those proteins are better named "analogs." Homology is a qualitative trait, not a quantitative one, such that saying that "sequences A and B are 60% homologous" does not make sense. In addition, it is a transitive trait, which means that if sequences (in the narrow meaning of: a sequence of nucleotides or amino acids, not necessarily a full-length protein or gene) A and B are homologous and B and C are homologous, A and C are homologous as well. In molecular evolution of eukaryotes, the two most important ways in which two proteins can be related are by speciation (orthology) or by duplication (paralogy). Hence, proteins are orthologs if they result from a species divergence, and they are paralogs if they result from a gene duplication in a given species. The terms "homologs," "orthologs," and "paralogs" are used to describe the relationship between two genes or proteins. However, comparative genomics is often about studying many species, and therefore many proteins, at the same time. Therefore, we use the concept of the OG. An OG encircles the set of proteins that diverged from a single protein in the last common ancestral species, such as "the last eukaryotic common" (LECA) or "the last common ancestor of all animals." Within the OG, two proteins can be paralogs if they resulted from a gene duplication that occurred in a more recent lineage than the ancestral species we chose to define the OG for. These proteins are called "inparalogs." Two proteins that result from a gene duplication prior to the last common ancestor of choice are called "outparalogs"; they are in different OGs.



Box Figure a) The evolution of hypothetical proteins X (green), Y (violet), and Z (orange), before and after LECA, projected onto the species tree. Before LECA (indicated by the purple box), an ancestral protein duplicated twice to give rise to three proteins in LECA, as indicated by "1" and "2." These duplications gave rise to the OGs of proteins X, Y, and Z. After LECA, protein X duplicated in the lineage leading to *A. thaliana*, and protein Y duplicated before the last common ancestor of *Acanthamoeba castellanii* (*A. castellanii*) and *Homo sapiens* (*H. sapiens*). Various eukaryotic lineages lost proteins. For example, humans lost protein Z. The bullets behind the species names indicate the supergroups to which they belong, as in the figure of Box 3. b) Gene tree of hypothetical proteins X, Y, and Z, rooted on other, further related homologous sequences. Regularly, the nodes in the tree that unite sequences from the same species are inferred to be duplication nodes, and the nodes in the tree that unite sequences from different species are mostly speciation nodes. However, when the protein topology differs from the species topology, ancient duplication nodes lineage-specific losses may have to be inferred, although there are no sequences from the same species on both sides of these nodes.[42] Duplication nodes prior to LECA separate outparalogs and duplication nodes after LECA separate inparalogs, as indicated with dotted lines. Proteins that belong to the same OG descent from a single protein in LECA. These LECA proteins are reflected by the internal nodes "X," "Y," and "Z." Note that within this figure, not all pairs of outparalogs and pairs of orthologs are indicated with curved lines. For example, protein Y of *A. thaliana* and protein Z of *A. thaliana* are also outparalogs. In fact, because these proteins are split by a duplication node, protein Y of *A. thaliana* and protein Z. of *H. sapiens* are also outparalogs. Also, the protein Z of *A. castellanii* and *P. falciparum* form another pair of orthologs, so do proteins Yα of *H. sapiens* and protein Y of *P. falciparum*. Since protein Yβ (*H. sapiens*) is also orthologous to protein Y of *P. falciparum*, human Yα and Yβ are called "co-orthologs" to this *P. falciparum protein*.

orthologs in species that were not present in the first search. Iterative searching tools make a "sequence profile" of the protein, based on the MSA of the hits found in the first search. These tools then use this profile to again search through the database. Because this profile contains much more information than a single query sequence, it is more sensitive to detect divergent orthologous sequences. With such a profile-based iteration, we might obtain a similar result as with the longer or less diverged sequence in the single search, implying the same, simple evolutionary scenario. An example that adheres to this straightforward scenario is cytochrome *b*, a mitochondrial protein operating in the electron transport system.[57]

3.3. Scenario 2 ("Taxonomically Limited?"): Limited Taxonomic Distribution, Protein Recently Invented or Orthologs That Are Difficult to Detect?

We hypothesize the second scenario if we retrieve hits only in closely related organisms, with low sequence similarity (e.g., a <75% similarity hit in mouse when searching with a human protein sequence). At first sight this might indicate a recent origin. However, it more likely reflects the evolutionary history of an older protein that diverged rapidly and therefore does not allow us to detect the more distant orthologs in the first search (Figure 1c). Proteins that are typically prone to such detection failures are those having many nonglobular regions (mainly coiled-coil or intrinsically disordered). In fact, many proteins are much older than suggested by simple similarity searches such as BLASTP. Under closer bioinformatic or biochemical scrutiny, these proteins often can be shown to have orthologs in more distantly related species than observed before.[58,59] Lineage-specific proteins are the exception, especially for proteins involved in core cellular processes. Advanced search strategies to cope with this scenario are discussed in Section 4.2.

3.4. Scenario 3 ("Lineage-Specific Duplication"): Multiple Human Sequences Are Highly Similar and Point to Lineage-Specific Duplication(s)

We encounter the third scenario if we search with a human query protein and find hits in human with a higher similarity than hits in more distantly related species. Such results indicate that the protein underwent recent lineage-specific duplications, e.g., in the ancestor of vertebrates (Figure 1d). Although this evolutionary scenario is easy to comprehend, it does not necessarily give an easy answer to the question "when did this protein originate?". If the functional differentiation between the duplicates is minimal, both likely conserved the function of the pre-duplication ancestor. Hits in species that diverged before this duplication (such as a hit in the choanoflagellate *Salpingoeca rosetta* (*S. rosetta*) in the example of a duplication in the metazoan ancestor, see Box 3) are also expected to perform this ancestral function. Therefore, we would consider these pre-duplication hits to be the "same" protein as well. However, if there was functional innovation in the lineage leading to the query protein ("neofunctionalization"),[60] it makes more sense to consider pre-duplication hits not to be the "same protein." For example, hemoglobin evolved after duplication of an ancestral globin in the vertebrate lineage.[61] Hence, while globins clearly exist in other animals and eukaryotes, it is not very meaningful to posit that plants contain hemoglobin even though they are technically orthologous (Box 2). In this case, we might opt to define the OG not on the level of all eukaryotes, but on the level of vertebrates. To distinguish between these two alternative histories, we assess whether the protein of interest neofunctionalized by using functional information about this protein, its (human) paralog, and the pre-duplication orthologs. Furthermore, in order to pinpoint the timing of the duplication (was it indeed in the common ancestor of vertebrates, or maybe already in the ancestor of all chordates?), we make a gene tree.

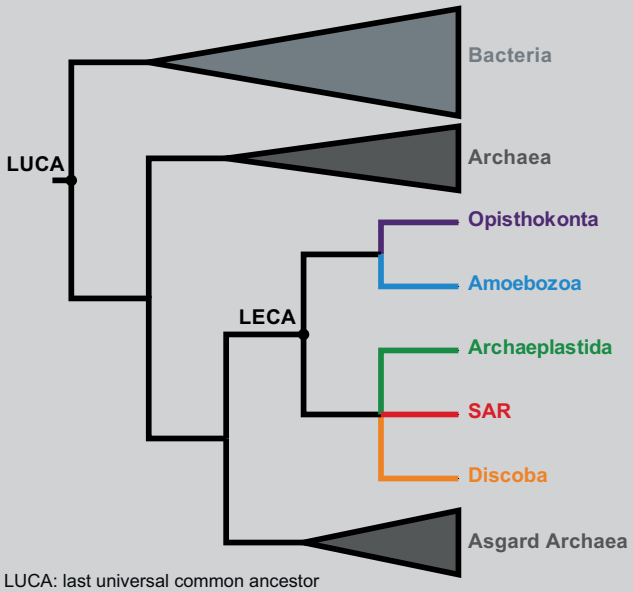
3.5. Scenario 4 ("Ancient Duplication"): Similar Proteins in All Eukaryotes Likely Are Paralogs That Resulted from an Ancient Duplication

In the fourth scenario, the search result contains homologs in human, but many hits in more distant eukaryotes are more similar to the query than this human sequence (Figure 1e). This

Box 3
A species selection for eukaryotes

We consider it useful to work with a selection of genomes as an initial search database, because this often gives more comprehensible results. Using such a subset prevents that only closely related species are visible in the sequence similarity search output, which may be the case in large databases such as "nr," the nonredundant protein database from NCBI.[39] A selection also facilitates checking for species that lack hits. We suggest including the species of the query in this database, because the search results then can reveal lineage-specific or ancient duplications. Moreover, we suggest including species that represent the diversity of the relevant taxon, such as eukaryotes, and make sure that all major clades are represented by multiple species (see the figure). Ideally, one includes for each of these major clades those species that do not have strongly reduced genomes. For this reason we, e.g., recommend to avoid having only pathogenic species for a given clade. Such species often lack orthologs of the protein of interest and may therefore, e.g., lead to erroneous conclusions about the age of a protein ("When and from where did my protein originate?"). Moreover, it may be helpful to include model organisms if these harbor additional experimental data, which often applies to budding and fission yeast, and sometimes to *Arabidopsis thaliana* (*A. thaliana*). Experimental data may give hits about conservation of function, and/or indicate candidate orthologs ("When and how to consider gray zone hits"). Although for specific proteins other subsets are more useful ("Using a tailor-made database yields more comprehensible and more interesting results") as an initial search database we suggest the species in this table. Of course, if any species not in this list is likely to be relevant for the protein of interest, one may add it or let it replace closely related species. Species suggested can be selected when using BLASTP or phmmer online, or their proteome sequences can be downloaded from NCBI when running similarity searches locally. The species are colored according to the supergroup to which they belong (see the figure).

Species	Taxon	Eukaryotic supergroup	Taxonomy ID
Homo sapiens	Chordata	Opisthokonta	9606
Xenopus tropicalis	Chordata	Opisthokonta	8364
Drosophila melanogaster	Arthropoda	Opisthokonta	7227
Salpingoeca rosetta	Choanoflagellida	Opisthokonta	946362
Saccharomyces cerevisiae	Ascomycota	Opisthokonta	4932
Spizellomyces punctatus	Chytridiomycota	Opisthokonta	109760
Thecamonas trahens		Apusozoa	529818
Acanthamoeba castellanii	Longamoebia	Amoebozoa	5755
Dictyostelium discoideum	Mycetozoa	Amoebozoa	44689
Amborella trichopoda	Streptophyta	Archaeplastida	13333
Klebsormidium flaccidum	Streptophyta	Archaeplastida	3175
Chlamydomonas reinhardtii	Chlorophyta	Archaeplastida	3055
Bathycoccus prasinos	Chlorophyta	Archaeplastida	41875
Cyanidioschyzon merolae	Rhodophyta	Archaeplastida	45157
Ectocarpus siliculosus	Stramenopila	SAR	2880
Phytophthora infestans	Stramenopila	SAR	403677
Plasmodium falciparum	Alveolata	SAR	5833
Paramecium tetraurelia	Alveolata	SAR	5888
Plasmodiophora brassicae	Rhizaria	SAR	37360
Naegleria gruberi	Percolozoa	Discoba	5762
Bodo saltans	Euglenozoa	Discoba	75058
Giardia intestinalis	Metamonada	Discoba	5741
Trichomonas vaginalis	Metamonada	Discoba	5722



Box Figure Species tree showing the position of the eukaryotes in the tree of life. According to the two primary domains model, the LUCA diverged to give rise to Bacteria and Archaea. Eukaryotes evolved from within the Archaea,[43,44] likely from the newly identified Asgard superphylum.[21,45] After divergence from these archaeal ancestors, the pre-eukaryotic lineage engulfed an Alphaproteobacterium via endosymbiosis, which later evolved into the mitochondrion.[46] Current-day eukaryotes descend from a single common ancestor, called the LECA. After LECA, eukaryotes diverged quickly into the major lineages that we refer to as "supergroups": Opisthokonta, Amoebozoa, Archaeplastida, SAR (Stramenopila-Alveolata-Rhizaria), and Discoba.[47] Note that in addition to these major lineages, there are apparently smaller lineages. For some of these, the position in the tree of life is still unresolved, such as Cryptophyta (e.g., *Guillardia theta*), Centrohelida (e.g., *Raphidiophrys contractilis*), and Haptophyta (e.g., *Emiliania huxleyi*).

suggests that before LECA (a) gene duplication(s) resulted in two or more proteins that probably functionally diverged, and that each gave rise to separate OGs (Box 2). After LECA, these proteins likely conserved their functions in different eukaryotic lineages. This scenario applies to, e.g., tubulins and Rab GTPases.[20,62,63] In case of broadly conserved protein domains, such as kinases, protein sequences that descended from different proteins in LECA are often mingled in the similarity search output. In order to assign these sequences to the correct OG, and hence to establish the OG of the protein of interest, we have to make a gene tree.

3.6. Making and Using Gene Trees under Different Evolutionary Scenario

Although particularly important for the third and fourth scenario, we also make and interpret ("reconcile") gene trees in case of the other scenarios. Note that while we here use the term "gene trees" - this term highlights that they model the evolution of a gene, as opposed to the "species tree" - we typically use protein sequences to make a gene tree for longer evolutionary distances. We make gene trees for two reasons. First, we use them to test whether the proposed scenario was the correct one. The lineage-specific duplication scenario will yield a gene tree with a topology very different from that of the ancient duplication scenario (Figure 1d,e, right-sided panels). Second, the gene tree uncovers the evolutionary history of the protein into greater detail. When making the gene tree, we should be sure that the sequences that we input are in fact homologous to one another: the algorithms that align proteins or reconstruct phylogenies assume that the sequences are homologous. Using nonhomologous sequences as input gives meaningless output. Moreover, we should select those homologous sequences that we expect to be informative. In the simple and taxonomically limited scenarios, we simply input all homologs found, or, if the database is very large, we select a subset that represents the diversity of the hits and species. In the lineage-specific duplication scenario, if we think a gene duplicated in, e.g., animals, we include all sequences from organisms that diverged just before (e.g., the choanoflagellate *S. rosetta*) and after (e.g., the sponge *Amphimedon queenslandica*) the hypothesized duplication. Conversely, under the ancient duplication scenario, we include two or more sequences from various evolutionary more distant genomes, while ensuring that each eukaryotic supergroup is well-represented.[64] After having constructed the tree, we compare it to the proposed scenario in order to assess whether it was drafted correctly. Moreover, we use the tree to explain in detail what happened during the

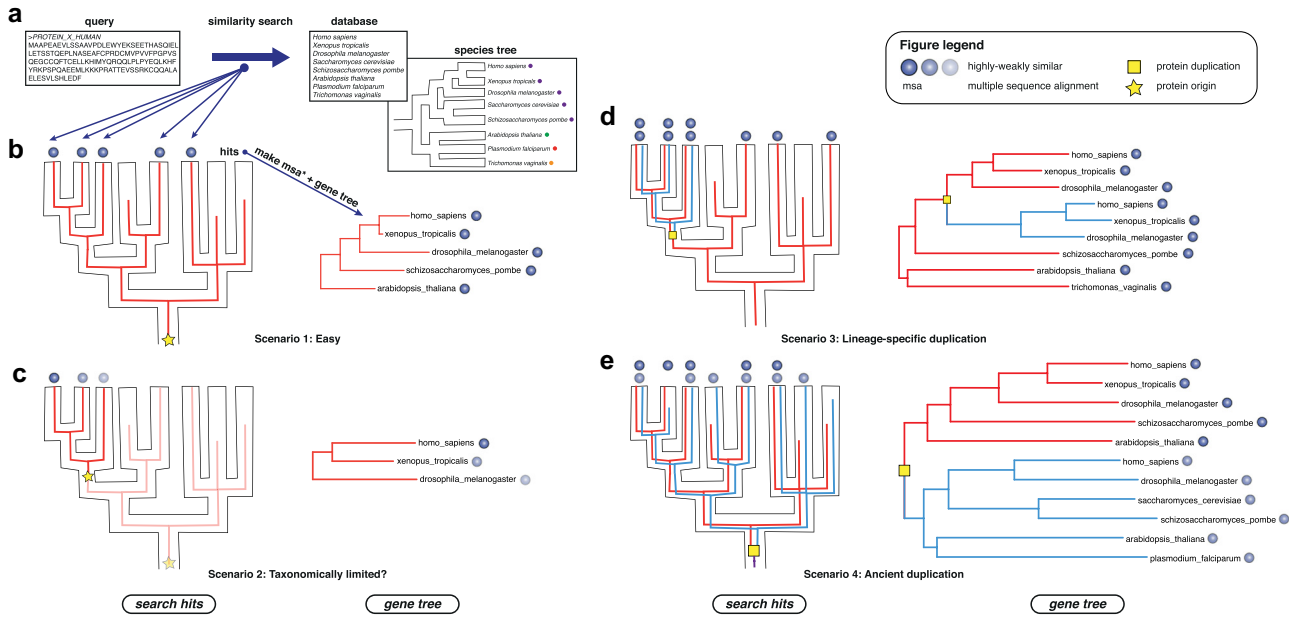


Figure 1. Outline of the initial analysis steps and their results in the case of scenario 1: easy. a) After having collected basic protein information, the initial analytic step is to perform sequence similarity searches with a query sequence (here illustrated as a human sequence) against a proteome database. This database may consist of a subset of species, ideally species of which we know the phylogenetic relationships, as indicated by the species tree. The colored bullets indicate the supergroups to which the species belong, as in Box 3. Here we illustrated the proteome database as a subset of eukaryotic species. The search yields sequence similarity hits that inform about potential scenarios. After hit protein sequences are collected, a MSA and gene tree can be inferred. b–e: Similarity search results and expected gene tree for four evolutionary scenarios. b) The easy scenario is hinted at by highly similar hits, mostly single hits in a variety of species. Likely, the protein of interest is old and relatively well conserved at the sequence level. Not all species in the database might have it due to gene loss. As a result, the gene tree does not include all species in the species tree. c) The taxonomically limited (?) scenario delivers few hits in the sequence similarity searches, only in species closely related to that of the query sequence. Although this might suggest a recent origin (upper left star), the protein might in fact be older (faded star below). This protein might evolve rapidly in different lineages, due to which sequence similarity is below significance and hence various orthologs are not seen in the similarity output. If no further, more advanced searches are executed, the gene tree includes only sequences from closely related species. d) The lineage-specific duplication scenario predicts highly similar, multiple hits in the query species and in species that are closely related. The gene tree aids in determining when this duplication occurred. In this case, this duplication took place in a common ancestor of vertebrates and fruit fly, after their divergence from fungi. This may have been the common ancestor of animals. e) The ancient duplication scenario predicts multiple hits across the species that are part of the database, in which one has a higher similarity to the query sequence than the other(s). In this scenario, the hits include sequences from multiple OGs (Box 2) and a gene tree is needed to discriminate which sequences belong to which OG. NB: The left panels contain gene trees within the species trees. These illustrate the protein's evolutionary history. The right panels show the actual gene trees. Here these gene trees do not have branch length values and support values, such as bootstrap percentages. Furthermore, while we here use the term "gene tree," because the evolution occurs at the genetic level, it also describes the evolution of the protein as a consequence of that. Generally, amino acid sequences are being used to infer such a gene tree, particularly when studying evolution on longer evolutionary timescales.

evolution of the protein (Box 1). When, in which ancestral lineages, did the protein duplicate, was lost or even horizontally transferred between species? This process of interpreting the gene tree by comparing it to the species tree is called tree reconciliation, on which detailed guides can be found in bioinformatic textbooks.[65,66] Tree reconciliation can be challenging, particularly if the gene tree contains errors (see Section 4.4). New computational methods are being developed regularly in order to facilitate tree reconciliation.[67] Altogether, after making the gene tree we are able to sketch a more refined evolutionary scenario for the protein of interest.

4. Challenging the Evolutionary Scenario

While executing evolutionary analyses, we often encounter challenges in either the data or in the analysis itself. Such challenges may force us to revise the evolutionary scenario for our protein of interest, and sometimes even to reformulate our research question. We here discuss these challenges in order to recognize them and apply strategies to overcome them, or to make informative disclaimers in a manuscript.

4.1. Multi-Domain Proteins May Require Multiple Independent Analyses

The protein of interest may consist of multiple domains, which is difficult if these domains have their own evolutionary history. The latter is quite common, given that domains seem to fuse more often than they split[68,69] and the evolution of new domain architectures has been relatively frequent in

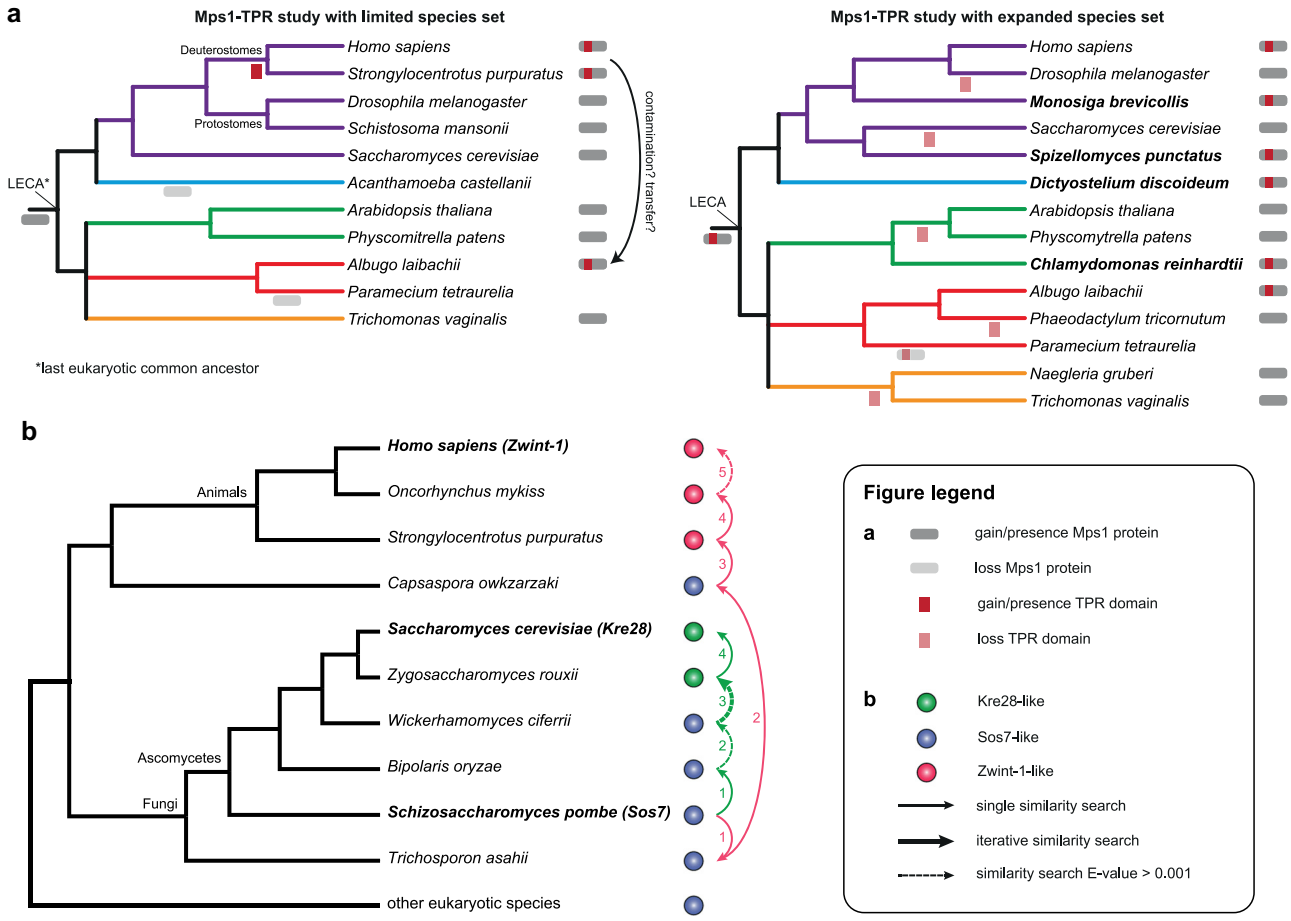


Figure 2. a) Sample size affects the interpretation of the evolutionary history of a protein domain. Left panel: Evolutionary analysis of the protein kinase Mps1 initially showed that this kinase was likely ancient, present in LECA, but it seemed to have fused to a TPR domain more recently. This TPR domain was inferred to have fused to the kinase in the common ancestor of deuterostomes.[40] The presence of this domain in the Mps1 protein of *A. laibachii* was hypothesized to result from contamination of the genome or horizontal gene transfer. Right panel: After genomes of other species were included in the analysis, various early branching species turned out to possess the TPR domain in their Mps1 proteins.[41] Hence, likely the Mps1 protein of LECA already had this domain. As a result, it must have been lost in various eukaryotic clades. The presented phylogenies are species trees with the branches colored according to the eukaryotic supergroups to which the species belong. Purple: Opisthokonta (animals and fungi), blue: Amoebozoa, green: Archaeplastida (land plants, algae), red: SAR (Stramenopila, Alveolata, and Rhizaria), orange: Discoba. The species that were important for inferring the ancient origin of Mps1's TPR domain are indicated in bold. b) Orthology established via bridging sequences. Zwint-1 (human), Kre28 (budding yeast), and Sos7 (fission yeast) are orthologous to one another and belong to a single OG. This conclusion was based on successive similarity searches that indicated homology between various Zwint-1-like, Kre28-like, and Sos7-like proteins of different fungal, animal, and animal-related species.[35] Starting with the Sos7 protein (fission yeast, *Schizosaccharomyces pombe*), the ordered searches indicate which "bridging" sequences were used to establish orthology to Kre28 and Zwint-1. Note that in one instance, iterative homology searching was needed, and in two others, hits with *E*-values higher than 0.001 were included, which can be considered gray zone hits. The species tree indicates which species' sequences were used to establish these homologies.

metazoan.[70] Different evolutionary histories may be hinted at by inconsistencies in the similarity search results, e.g., if hits from different taxa align to different regions of the query. If the domains indeed have different histories, we suggest three ways to investigate and represent the evolutionary history of the protein. First, if all domains are relevant, all deserve a thorough investigation and representation of their evolutionary history. Second, if one domain can be considered the primary domain and the other an accessory domain, we perform primarily a detailed analysis of the primary domain. Afterwards, we may infer when during evolution the accessory domain was acquired and/or lost. This is the way we represented the evolution of the kinetochore protein kinase Mps1, which in humans is joined to a tetratricopeptide repeat (TPR) domain (**Figure 2**a).[41] We also approached the evolution of human RasGRP proteins this way, whereby the Cdc25 homology domain, the main domain, appeared to have been fused to different other domains in different lineages.[71] Third, sometimes the domain combination itself is relevant for the biological function of interest. In this case, although the composite domains have their own possibly byzantine history, only after they fused "this protein" arose, and therefore we investigate the phylogenetic time point

of this fusion and the evolution thereafter. Although orthologous sequences in species that diverged before the fusion likely exist (having only one of the domains), these are probably not very informative. An example of this third case is R-spondin, a protein of the Wnt signaling pathway, that likely came into existence after two Fu domains joined a TSP1 domain in the ancestor of deuterostomes.[72]

4.2. When and How to Consider Gray Zone Hits

As discussed in the taxonomically limited scenario, quite often we expect a protein to be widely present, e.g., because it is involved in a core cellular process, but end up with a set of sequences from only a limited set of species. This occurs if the protein under investigation is largely unstructured, and therefore its amino acid sequence is poorly conserved. In order to challenge the taxonomically limited scenario, we scrutinize hits with a higher *E*-value than the cut-off, which we call "gray zone" hits. A gray zone hit may gain support from additional, less conventional similarity searches. For example, we may search iteratively during multiple rounds, thereby starting the search not with the initial (e.g., human) query sequence, but with another trusted ortholog that may serve as a "bridge." We experienced that searching with another sequence occasionally yields orthologs in many species that previously appeared to lack them. We select such bridging sequences by the taxonomic position of their species: we search with a trusted ortholog in a species that is closely related, e.g., in another ascomycete in case of a budding yeast. In addition to the taxonomic position, we take into account the evolutionary rate of a potential bridging species. If a species is known to have a rapidly evolving genome (Table 1), it is unlikely to serve as a useful bridge. Sometimes, newly found sequences trace even other, previously undetected sequences (homology is a transitive feature, Box 2). For example, in order to establish whether the kinetochore protein Sos7 of fission yeast is orthologous to the Kre28 protein of budding yeast, we used Sos7 homologs from ascomycete fungi to search in other yeasts' proteomes, and this indeed hit Kre28 (Figure 2b). Similarly, stepping from one orthologous sequence to another also helped us to identify orthology between Sos7 and human Zwint-1.

We regularly attempt to increase confidence in the gray zone hit with other information, e.g., from the actual alignment: Does the gray zone hit contain functionally characterized sequence motifs, or possibly a reduced version thereof? Moreover, the actual alignment may reveal that the protein is truncated, due to which the *E*-value is higher. Other information that would increase the likelihood of gray zone hit to be homologous includes whether the gray zone hit is a bidirectional best hit to the query, whether it has been reported to execute the same function, or whether it interacts with a protein that is orthologous to the query's interaction partner.[59,73] The gray zone hit may have a similar secondary or higher-level structure as the query sequence, which can be assessed with various tools. Using this multimodal approach we were, for instance, able to determine a much broader distribution of the RPGRIP1 proteins, crucial components of the ciliary transition zone,[74] than previously found.[75]

4.3. Compositionally Biased Proteins May Evolve Convergently

Nonglobular protein regions, such as coiled-coils, transmembrane domains, and disordered regions, are prone to evolve convergently, specifically if they have compositionally biased amino acid frequencies or simple, repeated sequence motifs.[76] When applying sequence similarity searches to these segments, we run the risk of obtaining hits with a low *E*-value that are nevertheless not homologous.[77] They acquired a similar amino acid sequence independently.[76,78,79] While standard BLASTP and phmmer mask such low-complexity regions, we think that one should be particularly careful in using iterative searches for proteins having such segments. If iterative searches do not converge and continue to return new hits, this may indicate that nonhomologous sequences are being included. Because inferring common descent for nonglobular regions is notoriously difficult,[76] we think that sets of orthologs that are only based on sequence similarity of coiled-coil and disordered regions should be treated with caution. However, if a candidate sequence shares a coiled-coil region with the protein of interest in addition to other homologous regions, the coiled-coil may serve as additional support to infer homology.

4.4. How Gene Trees May Be Incorrect

In each of the scenarios discussed, the evolutionary analysis is aided by a gene tree, and a gene tree is particularly essential in scenarios involving duplications and horizontal gene transfer. What sort of problems might a gene tree encounter? Quite regularly, the tree suffers from poor statistical supports, as indicated by low bootstrap values, approximate likelihood-based measures or posterior probabilities. This undermines the tree's reliability as representation of the proteins' evolutionary scenario. Another, often coinciding, reason for distrusting the tree could be that it is typically hard to reconcile with the species tree. Such reconciliation might require an inconceivable number of duplications and losses, or multiple horizontal gene transfers. Note that, even for proteins that are present in single copy in most of the species studied, the gene tree in fact seldom exactly follows the species tree.[80] Therefore we recommend to, generally speaking, not reconciling too strictly. Problematic trees may arise if the aligned sequences are not homologous along the full length, if sequences are short and hence contain little information, or if they are highly divergent, which may lead to so-called long-branch attraction (the tendency of rapidly evolving sequences to cluster together in the gene tree, although they are in reality not closely related).[81,82] The first problem can easily be solved by selecting the homologous positions in the alignment, for which also specific tools exist.[83] The other problems are more difficult to overcome, although it may help in adding more sequences, particularly ones that increase the phylogenetic diversity. However, increased sample size also increases the chance of obtaining phylogenetically incorrect branches with low bootstrap supports, for which solutions have been offered.[84] Sequences that align poorly and/or have very long branches could be removed from the alignment and be replaced with a more conserved ortholog from a closely related species.

4.5. Mixed Scenarios Face Multiple Challenges

Some proteins have evolutionary histories that do not fall into a single scenario category but combine features of different categories. For example, in investigating the evolution of the plant polycomb repressive complex 1 (PRC1) component EMF1, through sensitive sequence analysis, we first discovered multiple flowering plant-specific duplicates before we discovered that gymnosperms in fact also have, highly diverged, orthologs.[85] Here, lineage-specific duplications and "invisible" orthologs thus conspired to make the previous inferences in the literature flawed. Moreover, even proteins with common domains, such as certain kinases, sometimes diverged extensively, due to which the initial gene tree is very likely to lack orthologs. As a result, some lineages may incorrectly appear to have lost the protein. In this case, again, iterative similarity searching with a sequence profile might help in finding sequences that should be added to the tree. Such a search is most sensitive if it starts with an OG-specific profile, i.e., a profile that is based on the alignment of the OG that was identified from an earlier tree.

4.6. Is My Protein Truly Absent/Present in This Species?

Quite often, the evolutionary scenario entails unexpected absences of orthologs in certain species, or unexpected presences. Are these unexpected observations true or not? Absences may be false if the similarity searches fail to detect homologous sequences, or if the predicted proteome or assembled genome is incomplete. We assessed that protein-coding sequences quite frequently are not, or incompletely, predicted.[86] It might therefore pay off to search for homologous sequences on the (six-frame translated) genomic DNA. However, unpredicted genes may also be due to global codon reassignments and local recoding, to which this may not offer a solution. Less frequently, unexpected presences are observed, e.g., if a protein that seemed eukaryote-specific is found in a few bacteria. While such presences may be true, e.g., caused by eukaryote-to-bacterium gene transfer, they may also be false, e.g., if the bacterial genome is contaminated with eukaryotic sequences.[87] It is not always easy to discriminate between these because A) contamination can be difficult to prove, and B) although horizontal transfer is rare in eukaryotes, it does occur.[88] Moreover, an unexpected "presence" may result from falsely inferred homology, as discussed in Section 4.3.

4.7. When Did My Protein Originate?

Above we referred to "the origin of a protein" mostly as the position on the species tree in which the protein came into existence. Logically, errors in the proposed evolutionary scenario may also result in errors in the inferred origin. A protein may be estimated too young, due to undetected homologs in more distantly related species, which is mainly observed in the taxonomically limited scenario. In some instances, the age of a protein might simply be underestimated because crucial genomes that point to an older age are not (yet) available. The TPR domain of the protein kinase Mps1 had been inferred to have originated in the ancestor of deuterostomes, and its presence in the oomycete *Albugo laibachii* (*A. laibachii*) was considered contamination or horizontal gene transfer.[40] However, when more species were included, this TPR domain was found in various other clades across the eukaryotic tree of life (Figure 2a). Hence, it likely was already present in LECA but lost in multiple eukaryotic lineages in parallel.[41] In other cases, a protein may be estimated too old if a wide taxonomic distribution is actually due to horizontal gene transfer, which is, e.g., the case for genes that were transferred between plants and fungi.[89]

4.8. A Tailor-Made Database Yields More Comprehensible and More Interesting Results

Since public databases nowadays contain tons of sequence data, it is computationally and practically quite a challenge to interpret the results of a simple similarity search. Furthermore, this output does not reveal which species do not have a hit. For this reason, in the previous section, we recommended to select a subset of species as a search database (Box 3). As an alternative, we prefer to use a tailored, local proteome database with a fixed set of species of which we know the (approximate) species tree. Although this is considerably more work and requires more advanced skills in bioinformatics, we think it does pay off if one intends to study the evolution of proteins regularly. An in-house database also facilitates the detection of coevolution among different proteins (Table 1). The database may be revised or recompiled if a protein turns out to have a particularly interesting function or evolutionary history in a specific clade. Various studies demonstrated that zooming-in into the species tree and adding species at key phylogenetic positions yield highly interesting patterns, such as when the protein turns out to be present in species having a certain (cellular) biological feature. A good example is provided by the centromeric histone variant CenH3, a protein that turned out to be absent from species that have a specific type of centromere (a so-called holocentromere, which runs along the length of the chromosome), in certain lineages of insects.[5]

5. Conclusions

With this article, we hope to have given relevant insights into how we approach the evolutionary analysis of proteins, and we hope that these insights may serve as a guide. Essentially, we consider this analysis a process that entails many feedback loops that aim to revise a proposed scenario of evolution. This process is regularly unfinished: new data, such as a resolved 3D structure or newly sequenced genomes, may alter the scenario of the protein's evolution, as may technological advances, such as improved homology detection and tree building. The derived evolutionary scenario should therefore be considered the best estimate of the moment, not necessarily the definitive one. To explicate this uncertainty, we suggest that authors report doubtful cases: which set of orthologs possibly contains false positives or false negatives, and why? Which evolutionary events may require re-examination? Also, reporting the actual

orthologous sequences, not just their identifiers, will facilitate such re-examination.

The common efforts of many researchers generated detailed hypotheses on the evolution of a wide array of proteins. Likely, many others would be interested to quickly retrieve this information. Unfortunately, these data are now often hidden in research articles. If we and others would share manually defined orthologs on a wide and homogeneously formatted platform, these would be easier to access. This does not necessarily need to be a new platform: maybe our research community could aid to improve existing databases such as EggNOG, PANTHER, Quest for Orthologs, or TreeBASE by adding the knowledge of the proteins we studied.[26,53,90–92]

Acknowledgements

B.S. designed the manuscript. J.J.E.v.H. wrote the primary manuscript with significant contributions from E.T., T.J.P.v.D., G.J.P.L.K., and B.S. The authors kindly thank Carlos Sacristan, Simona Antonova, and Mathilde Galli for their critical and useful feedback on the manuscript. This work was supported by The Netherlands Organisation for Scientific Research (NWO-Vici 016.160.638 to B.S.). E.T. was supported by a postdoctoral fellowship from the Herchel Smith Fund, Cambridge, UK.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

comparative genomics, orthology, paralogy, protein evolution, timing gene duplication, timing gene origin

Received: January 8, 2019
Revised: February 17, 2019
Published online: April 26, 2019

[1] B. Li, S. Oestreich, T. de Lange, *Cell* **2000**, *101*, 471.
[2] A. A. W. M. Sanders, E. de Vrieze, A. M. Alazami, F. Alzahrani, E. B. Malarkey, N. Sorusch, L. Tebbe, S. Kuhns, T. J. P. van Dam, A. Alhashem, B. Tabarki, Q. Lu, N. J. Lambacher, J. E. Kennedy, R. V. Bowie, L. Hetterschijt, S. van Beersum, J. van Reeuwijk, K. Boldt, H. Kremer, R. A. Kesterson, D. Monies, M. Abouelhoda, R. Roepman, M. H. Huynen, M. Ueffing, R. B. Russell, U. Wolfrum, B. K. Yoder, E. van Wijk, F. S. Alkuraya, O. E. Blacque, *Genome Biol.* **2015**, *16*, 293.
[3] S. S. Schalm, J. Blenis, *Curr. Biol.* **2002**, *12*, 632.
[4] J. Falck, J. Coates, S. P. Jackson, *Nature* **2005**, *434*, 605.
[5] I. A. Drinnenberg, D. deYoung, S. Henikoff, H. S. Malik, *eLife* **2014**, *3*, e03676.
[6] T. A. Richards, D. M. Soanes, M. D. M. Jones, O. Vasieva, G. Leonard, K. Paszkiewicz, P. G. Foster, N. Hall, N. J. Talbot, *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 15258.
[7] D. Jebb, M. Hiller, *eLife* **2018**, *7*, e38906.
[8] D. P. Anderson, D. S. Whitney, V. Hanson-Smith, A. Woznica, W. Campodonico-Burnett, B. F. Volkman, N. King, J. W. Thornton, K. E. Prehoda, *eLife* **2016**, *5*, e10147.
[9] K. Mykytyn, R. F. Mullins, M. Andrews, A. P. Chiang, R. E. Swiderski, B. Yang, T. Braun, T. Casavant, E. M. Stone, V. C. Sheffield, *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 8664.
[10] A. G. Bick, S. E. Calvo, V. K. Mootha, *Science* **2012**, *336*, 886.
[11] J.-Y. Lee, L. N. Kinch, D. M. Borek, J. Wang, J. Wang, I. L. Urbatsch, X.-S. Xie, N. V. Grishin, J. C. Cohen, Z. Otwinowski, H. H. Hobbs, D. M. Rosenbaum, *Nature* **2016**, *533*, 561.
[12] S. J. E. Suijkerbuijk, T. J. P. van Dam, G. E. Karagöz, E. von Castelmur, N. C. Hubner, A. M. S. Duarte, M. Vleugel, A. Perrakis, S. G. D. Rüdiger, B. Snel, G. J. P. L. Kops, *Dev. Cell* **2012**, *22*, 1321.
[13] E. Tromer, D. Bade, B. Snel, G. J. P. L. Kops, *Open Biol.* **2016**, *6*, 160315.
[14] J. Atherton, K. Jiang, M. M. Stangier, Y. Luo, S. Hua, K. Houben, J. J. E. van Hooff, A.-P. Joseph, G. Scarabelli, B. J. Grant, A. J. Roberts, M. Topf, M. O. Steinmetz, M. Baldus, C. A. Moores, A. Akhmanova, *Nat. Struct. Mol. Biol.* **2017**, *24*, 931.
[15] B. Wickstead, K. Gull, T. A. Richards, *BMC Evol. Biol.* **2010**, *10*, 110.
[16] P. B. Talbert, K. Ahmad, G. Almouzni, J. Ausió, F. Berger, P. L. Bhalla, W. M. Bonner, W. Z. Cande, B. P. Chadwick, S. W. L. Chan, G. A. M. Cross, L. Cui, S. I. Dimitrov, D. Doenecke, J. M. Eirin-López, M. A. Gorovsky, S. B. Hake, B. A. Hamkalo, S. Holec, S. E. Jacobsen, K. Kamieniarz, S. Khochbin, A. G. Ladurner, D. Landsman, J. A. Latham, B. Loppin, H. S. Malik, W. F. Marzluff, J. R. Pehrson, J. Postberg, R. Schneider, M. B. Singh, M. M. Smith, E. Thompson, M.-E. Torres-Padilla, D. J. Tremethick, B. M. Turneret al., *Epigenet. Chromatin* **2012**, *5*, 7.
[17] L. M. Iyer, D. D. Leipe, E. V. Koonin, L. Aravind, *J. Struct. Biol.* **2004**, *146*, 11.
[18] B. J. Foth, M. C. Goedecke, D. Soldati, *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 3681.
[19] T. A. Richards, T. Cavalier-Smith, *Nature* **2005**, *436*, 1113.
[20] M. Elias, A. Brighouse, C. Gabernet-Castello, M. C. Field, J. B. Dacks, *J. Cell Sci.* **2012**, *125*, 2500.
[21] A. Spang, J. H. Saw, S. L. Jørgensen, K. Zaremba-Niedzwiedzka, J. Martijn, A. E. Lind, R. van Eijk, C. Schleper, L. Guy, T. J. G. Ettema, *Nature* **2015**, *521*, 173.
[22] A.-C. Lindas, E. A. Karlsson, M. T. Lindgren, T. J. G. Ettema, R. Bernander, *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 18942.
[23] N. Yutin, E. V. Koonin, *Biol. Direct* **2012**, *7*, 10.
[24] C. J. Leonard, L. Aravind, E. V. Koonin, *Genome Res.* **1998**, *8*, 1038.
[25] F. Tekaia, *Genomics Insights* **2016**, *9*, 17.
[26] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, M. Punta, *Nucleic Acids Res.* **2014**, *42*, D222.
[27] C.-M. Train, N. M. Glover, G. H. Gonnet, A. M. Altenhoff, C. Dessimoz, *Bioinformatics* **2017**, *33*, i75.
[28] B. J. Liebeskind, C. D. McWhite, E. M. Marcotte, *Genome Biol. Evol.* **2016**, *8*, 1812.
[29] H. Chiba, I. Uchiyama, *BMC Bioinf.* **2014**, *15*, 148.
[30] P. Bork, E. V. Koonin, *Nat. Genet.* **1998**, *18*, 313.
[31] N. Vrielynck, A. Chambon, D. Vezon, L. Pereira, L. Chelysheva, A. De Muyt, C. Mezard, C. Mayer, M. Grelon, *Science* **2016**, *351*, 939.
[32] T. Robert, A. Nore, C. Brun, C. Maffre, B. Crimi, V. Guichard, H.-M. Bourbon, B. de Massy, *Science* **2016**, *351*, 943.
[33] M. J. E. Koster, B. Snel, H. T. M. Timmers, *Cell* **2015**, *161*, 724.
[34] A. Peviani, J. Lastdrager, J. Hanson, B. Snel, *Sci. Rep.* **2016**, *6*, 30444.
[35] J. J. van Hooff, E. Tromer, L. M. van Wijk, B. Snel, G. J. Kops, *EMBO Rep.* **2017**, *18*, 1559.
[36] T. Gabaldón, E. V. Koonin, *Nat. Rev. Genet.* **2013**, *14*, 360.
[37] E. V. Koonin, *Annu. Rev. Genet.* **2005**, *39*, 309.
[38] M. Lynch, J. S. Conery, *Science* **2000**, *290*, 1151.
[39] NCBI Resource Coordinators, *Nucleic Acids Res.* **2015**, *44*, D7.

[40] S. Lee, P. Thebault, L. Freschi, S. Beaufils, T. L. Blundell, C. R. Landry, V. M. Bolanos-Garcia, S. Elowe, *J. Biol. Chem.* **2012**, *287*, 5988.
[41] W. Nijenhuis, E. von Castelmur, D. Littler, V. De Marco, E. Tromer, M. Vleugel, M. H. J. van Osch, B. Snel, A. Perrakis, G. J. P. L. Kops, *J. Cell Biol.* **2013**, *201*, 217.
[42] A. J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, E. Birney, *Genome Res.* **2009**, *19*, 327.
[43] J. A. Lake, E. Henderson, M. Oakes, M. W. Clark, *Proc. Natl. Acad. Sci. U. S. A.* **1984**, *81*, 3786.
[44] M. C. Rivera, J. A. Lake, *Science* **1992**, *257*, 74.
[45] K. Zaremba-Niedzwiedzka, E. F. Caceres, J. H. Saw, D. Bäckström, L. Juzokaite, E. Vancaester, K. W. Seitz, K. Anantharaman, P. Starnawski, K. U. Kjeldsen, M. B. Stott, T. Nunoura, J. F. Banfield, A. Schramm, B. J. Baker, A. Spang, T. J. G. Ettema, *Nature* **2017**, *541*, 353.
[46] T. A. Williams, P. G. Foster, C. J. Cox, T. M. Embley, *Nature* **2013**, *504*, 231.
[47] R. Derelle, G. Torruella, V. Klimeš, H. Brinkmann, E. Kim, Č. Vlček, B. F. Lang, M. Eliáš, *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, E693.
[48] D. R. Maddison, K.-S. Schulz, http://tolweb.org (accessed: May 2012) **2009**, *8*.
[49] R. D. Finn, J. Clements, W. Arndt, B. L. Miller, T. J. Wheeler, F. Schreiber, A. Bateman, S. R. Eddy, *Nucleic Acids Res.* **2015**, *43*, W30.
[50] W. Li, A. Cowley, M. Uludag, T. Gur, H. McWilliam, S. Squizzato, Y. M. Park, N. Buso, R. Lopez, *Nucleic Acids Res.* **2015**, *43*, W580.
[51] E. V. Kriventseva, D. Kuznetsov, F. Tegenfeldt, M. Manni, R. Dias, F. A. Simão, E. M. Zdobnov, *Nucleic Acids Res.* **2018**, *47*, D807.
[52] E. L. L. Sonnhammer, G. Östlund, *Nucleic Acids Res.* **2015**, *43*, D234.
[53] H. Mi, X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang, P. D. Thomas, *Nucleic Acids Res.* **2017**, *45*, D183.
[54] D. R. Zerbino, P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Girón, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K. To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D. N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D., *Nucleic Acids Res.* **2018**, *46*, D754et al.
[55] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, *Nucleic Acids Res.* **1997**, *25*, 3389.
[56] S. Altschul, The Statistics of Sequence Similarity Scores. https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html, **2011**.
[57] T. Sicheritz-Pontén, C. G. Kurland, S. G. E. Andersson, *Biochim. Biophys. Acta, Bioenerg.* **1998**, *1365*, 545.
[58] E. V. Koonin, Y. I. Wolf, L. Aravind, *Adv. Protein Chem.* **2000**, *54*, 245.
[59] R. Szklarczyk, B. F. Wanschers, T. D. Cuypers, J. J. Esseling, M. Riemersma, M. A. van den Brand, J. Gloerich, E. Lasonder, L. P. van den Heuvel, L. G. Nijtmans, M. A. Huynen, *Genome Biol.* **2012**, *13*, R12.
[60] M. W. Hahn, *J. Hered.* **2009**, *100*, 605.
[61] T. Burmester, T. Hankeln, *Acta Physiol.* **2014**, *211*, 501.
[62] P. Findeisen, S. Mühlhausen, S. Dempewolf, J. Hertzog, A. Zietlow, T. Carlomagno, M. Kollmar, *Genome Biol. Evol.* **2014**, *6*, 2274.
[63] G. Jékely, *BioEssays* **2003**, *25*, 1129.
[64] A. Simpson, C. H. Slamovits, J. M. Archibald, Protist diversity and eukaryote phylogeny, in *Handbook of the Protists*, (Eds: (Eds: A. Simpson, C. H. Slamovits, J. M. Archibald), Springer, Cham **2017**.
[65] O. Eulenstein, S. Huzurbazar, D. A. Liberles, Reconciling phylogenetic trees, in *Evolution after Gene Duplication*, (Eds: K. Dittmar, D. Liberles), **2011**.
[66] O. K. Kamneva, N. L. Ward, Chapter 9 - Reconciliation approaches to determining HGT, duplications, and losses in gene trees, in *Methods in Microbiology*, Vol. 41 (Eds: (Eds: M. Goodfellow, I. Sutcliffe, J. Chun), Academic Press, Cambridge, MA **2014**, pp. 183–199.
[67] L. Nakhleh, *Trends Ecol. Evol.* **2013**, *28*, 719.
[68] S. K. Kummerfeld, S. A. Teichmann, *Trends Genet.* **2005**, *21*, 25.
[69] B. Snel, P. Bork, M. Huynen, *Trends Genet.* **2000**, *16*, 9.
[70] D. Ekman, Å. K. Björklund, A. Elofsson, *J. Mol. Biol.* **2007**, *372*, 1337.
[71] T. J. P. van Dam, H. Rehmann, J. L. Bos, B. Snel, *Cell. Signalling* **2009**, *21*, 1579.
[72] W. B. De Lau, B. Snel, H. C. Clevers, *Genome Biol.* **2012**, *13*, 242.
[73] L. Fokkens, S. M. Botelho, J. Boekhorst, B. Snel, *BMC Bioinf.* **2010**, *11*, 86.
[74] N. J. Lambacher, A.-L. Bruel, T. J. P. van Dam, K. Szymańska, G. G. Slaats, S. Kuhns, G. J. McManus, J. E. Kennedy, K. Gaff, K. M. Wu, R. van der Lee, L. Burglen, D. Doummar, J.-B. Rivière, L. Faivre, T. Attié-Bitach, S. Saunier, A. Curd, M. Peckham, R. H. Giles, C. A. Johnson, M. A. Huynen, C. Thauvin-Robinet, O. E. Blacque, *Nature Cell Biol.* **2014**, *18*, 122.
[75] A. R. Barker, K. S. Renzaglia, K. Fry, H. R. Dawe, *BMC Genomics* **2014**, *15*, 531.
[76] J. Surkont, J. B. Pereira-Leal, *Genome Biol. Evol.* **2015**, *7*, 545.
[77] W.-C. Wong, S. Maurer-Stroh, B. Eisenhaber, F. Eisenhaber, *BMC Bioinf.* **2014**, *15*, 166.
[78] J. Mistry, R. D. Finn, S. R. Eddy, A. Bateman, M. Punta, *Nucleic Acids Res.* **2013**, *41*, e121.
[79] O. J. L. Rackham, M. Madera, C. T. Armstrong, T. L. Vincent, D. N. Woolfson, J. Gough, *J. Mol. Biol.* **2010**, *403*, 480.
[80] C. Ku, S. Nelson-Sathi, M. Roettger, S. Garg, E. Hazkani-Covo, W. F. Martin, *Proc. Natl. Acad. Sci. U. S. A.* **2015**.
[81] A. Som, *Briefings Bioinf.* **2015**, *16*, 536.
[82] H. Philippe, H. Brinkmann, D. V. Lavrov, D. T. J. Littlewood, M. Manuel, G. Wörheide, D. Baurain, *PLoS Biol.* **2011**, *9*, e1000602.
[83] S. Whelan, I. Irisarri, F. Burki, *Bioinformatics* **2018**, *34*, 3929.
[84] F. Lemoine, J. B. Domelevo Entfellner, E. Wilkinson, D. Correia, M. Dávila Felipe, T. De Oliveira, O. Gascuel, *Nature* **2018**, *556*, 452.
[85] L. Berke, B. Snel, *BMC Evol. Biol.* **2015**, *15*, 44.
[86] E. S. van Deutekom, T. J. P. van Dam, B. Snel, unpublished.
[87] G. Koutsovoulos, S. Kumar, D. R. Laetsch, L. Stevens, J. Daub, C. Conlon, H. Maroon, F. Thomas, A. A. Aboobaker, M. Blaxter, *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 5053.
[88] M. M. Leger, L. Eme, C. W. Stairs, A. J. Roger, *BioEssays* **2018**, *40*, 1700242 DOI: 10.1002/bies.201700115
[89] T. A. Richards, D. M. Soanes, P. G. Foster, G. Leonard, C. R. Thornton, N. J. Talbot, *Plant Cell* **2009**, *21*, 1897.
[90] J. Huerta-Cepas, D. Szklarczyk, K. Forslund, H. Cook, D. Heller, M. C. Walter, T. Rattei, D. R. Mende, S. Sunagawa, M. Kuhn, L. J. Jensen, C. von Mering, P. Bork, *Nucleic Acids Res.* **2016**, *44*, D286.
[91] W. H. Piel, L. Chan, M. J. Domunis, J. Ruan, R. A. Vos, V. Tannen, TreeBASE v. 2: A Database of Phylogenetic Knowledge, e-BioSphere, London **2009**.
[92] A. M. Altenhoff, B. Boeckmann, S. Capella-Gutierrez, D. A. Dalquen, T. DeLuca, K. Forslund, J. Huerta-Cepas, B. Linard, C. Pereira, L. P. Pryszcz, F. Schreiber, A. S. da Silva, D. Szklarczyk, C.-M. Train, P. Bork, O. Lecompte, C. von Mering, I. Xenarios, K. Sjölander, L. J. Jensen, M. J. Martin, M. Muffato, Quest for Orthologs consortium, T. Gabaldón, S. E. Lewis, P. D. Thomas, E. Sonnhammer, C. Dessimoz, *Nat. Methods* **2016**, *13*, 425.