Proteome-scale detection of differential conservation patterns at protein and sub-protein levels with BLUR

Audrey Defosset, Arnaud Kress, Yannis Nevers, Raymond Ripp, Julie Thompson, Olivier Poch, Odile Lecompte

Proteome-Scale Detection of Differential Conservation Patterns at Protein and Subprotein Levels with BLUR

Audrey Defosset [1], Arnaud Kress[1], Yannis Nevers[1,2,3,4], Raymond Ripp[1], Julie D. Thompson[1], Olivier Poch[1], and Odile Lecompte[1,*]

[1]Complex Systems and Translational Bioinformatics, ICube UMR 7357, Université de Strasbourg, France

[2]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

[3]Department of Computational Biology, University of Lausanne, Switzerland

[4]Center for Integrative Genomics, University of Lausanne, Switzerland

*Corresponding author: E-mail: odile.lecompte@unistra.fr.

Accepted: 18 November 2020

Abstract

In the multiomics era, comparative genomics studies based on gene repertoire comparison are increasingly used to investigate evolutionary histories of species, to study genotype–phenotype relations, species adaptation to various environments, or to predict gene function using phylogenetic profiling. However, comparisons of orthologs have highlighted the prevalence of sequence plasticity among species, showing the benefits of combining protein and subprotein levels of analysis to allow for a more comprehensive study of genotype/phenotype correlations. In this article, we introduce a new approach called BLUR (BLAST Unexpected Ranking), capable of detecting genotype divergence or specialization between two related clades at different levels: gain/loss of proteins but also of subprotein regions. These regions can correspond to known domains, uncharacterized regions, or even small motifs. Our method was created to allow two types of research strategies: 1) the comparison of two groups of species with no previous knowledge, with the aim of predicting phenotype differences or specializations between close species or 2) the study of specific phenotypes by comparing species that present the phenotype of interest with species that do not. We designed a website to facilitate the use of BLUR with a possibility of in-depth analysis of the results with various tools, such as functional enrichments, protein–protein interaction networks, and multiple sequence alignments. We applied our method to the study of two different biological pathways and to the comparison of several groups of close species, all with very promising results. BLUR is freely available at http://lbgi.fr/blur/.

Key words: comparative genomics, evolution, sequence analysis, genotype/phenotype relations.

Significance

Current tools are designed to compare gene repertoires between species, or to study the modularity of annotated protein domains between clades. Our work is designed to allow for the detection of differences between groups of species on both these levels, and more. The tool we designed, BLUR (BLAST Unexpected Ranking), can highlight divergences between clades at the whole protein level (presence/absence) as well as at the subprotein level, by detecting differences in protein sequences ranging from complete functional domains to small motifs. Our resource allows a more in depth study of the protein modularity that arised from evolution and will help with gaining a better understanding of genotype/phenotype relations.

Introduction

Technological advances in recent years have given rise to an ever-increasing amount of sequencing data, providing opportunities to capitalize on the available diversity of living organisms to study the evolution of various biological processes. Data from genome sequencing have been used to establish correlations between genotype and phenotype to improve gene function prediction. Full proteomes of distinct species can be compared with identify genes that are conserved, gained, or lost, and could be linked to phenotypical differences or species specificity. Comparison of genes that are present or absent in various species can not only help with understanding evolution and the adaptation of living organisms to different environments but it is also a useful comparative genomics approach for the inference of gene function. It is assumed that genes participating in the same mechanism will generally be conserved and lost together through evolution, and that functionally linked genes often present similar phylogenetic distributions (Pellegrini et al. 1999). It is thus possible to infer gene function and associate genes with various processes by matching a phenotype distribution to that of a set of genes. This method has been successfully applied to various processes and organelles, such as cilia (Li et al. 2004; Dey et al. 2015; Nevers et al. 2017), mitochondria (Cheng and Perocchi 2015), thermophily (Jim 2003), and the DOXP/MEP metabolic pathway (Cunningham et al. 2000).

Although phylogenetic profiling is a very insightful approach to explore evolutionary histories of species at the gene/protein level, it does not account for the modular nature of protein evolution. Many studies have quantified and characterized protein domain evolution, showing that domain gains and losses are quite common, and that sequence architectures are often rearranged between taxa, participating in lineage-specific adaptations (Zmasek and Godzik 2011; Moore and Bornberg-Bauer 2012; Lees et al. 2016; Dohmen et al. 2020). Such sequence divergences have been observed even between orthologs of closely related species, such as members of the genus *Drosophila* (Forslund et al. 2011; Moore et al. 2013). It has also been shown that sequence divergence on the scale of a region or a small motif can have nonnegligible impact, such as in homeotic genes in arthropods. Variations in sequences in several Hox orthologs have indeed been linked to developmental differences between various arthropod species (Löhr et al. 2001; Ronshaugen et al. 2002; Shiga et al. 2002). It is expected that such interspecific sequences divergences can also be observed when dealing with proteins participating in multiple processes, such as moonlighting proteins, which can exhibit two or more biological functions (Jeffery 1999). So far, several hundreds of proteins have been found to be involved in more than one process, and many more may exist that remain to be discovered (Mani et al. 2015).

Differences in sequences at various levels (motif, block, or domain) between orthologs can be challenging for traditional orthology inference methods, making it difficult to predict the correct relations between divergent sequences. In terms of comparison of gene repertoires, this means that the regions' variations, losses, or gains that may be observed in certain species will not only make it hard to predict the true orthologous relations but also to properly annotate their function through co-occurrence methods. Consequently, although it is important to consider gain and loss of complete genes, it is also crucial to take into account the domain composition and sequence divergences between orthologs to gain better insight into the complex relations between phenotype and genotype, and potentially predict specializations and phenotype divergences between closely related species.

Some attempts have been made to extend the classical gene-level phylogenetic profiling approach, either to fixed-length protein segments (Kim and Subramaniam 2005) or to conserved domains (Pagel et al. 2004; Persson et al. 2019) found in databases such as PFAM (El-Gebali et al. 2019) or SMART (Letunic and Bork 2018), in order to infer domain interactions and help identify physical and functional relationships between proteins. The PhyloPro2.0 phylogenetic profile database allows the visualization of PFAM domain conservation through heatmaps generated for 164 eukaryotes and can display up to 1,000 genes at a time (Cromar et al. 2016). The PhyloGene server allows users to retrieve coevolving genes by computing normalized phylogenetic profile according to sequence conservation and calculating *Z*-score between these profiles to assess coevolution (Sadreyev et al. 2015). Recently, Han et al. (2020) designed a new method, RASfam, based on subgene regions, called modules, and species phylogeny to infer evolutionary scenarios and construct homologous gene families. Some resources have also been designed that facilitate the identification of variable domains in protein families, such as PROBE, that allows users to find conserved blocks in a multiple sequence alignment (Kress et al. 2018), or TreeDom, a web tool designed to graphically represent domain architecture evolution in multidomain proteins (Haider et al. 2016). Other software tools, such as DoMosaics (Moore et al. 2014) or DomArch (Vera-Parra et al.,2016), have been developed to work in conjunction with available domain annotation services, and enable the comparison, analysis, and visualization of the evolution of domain architectures.

Generally, these tools are limited to the study of individual genes or gene families, and are not adapted to the study of complete proteomes. The programs also mostly focus on well-characterized functional domains such as PFAM, which prevents the analysis of uncharacterized domains or of regions without domain annotations, and do not allow the detection of subtle sequence divergence, which has been shown to alter

domain function entirely, even when the change affects only one amino acid (Anderson et al. 2016). The obvious need for a high-throughput method that would allow for the search of lineage-specific conservation patterns, at both the gene and subgene levels, in a complete proteome led us to develop a novel approach based on BLAST homology searches (Camacho et al. 2009), that is capable of detecting genotype divergence or specialization between two related lineages in a wide selection of organisms.

Here, we present the BLAST Unexpected Ranking (BLUR) method, a rapid, proteome-scale approach to analyze the protein conservation of two related taxa in order to detect atypical patterns. BLUR is designed to facilitate the study and understanding of genotype/phenotype relations by providing information both on the gain/loss of complete proteins and on the specific divergences of subprotein regions, ranging from small motifs to complete domains. It can be used both as an exploratory tool to compare two groups of interest with no previous knowledge, or to study specific phenotypes and identify proteins linked to them. To facilitate the exploitation of results, a website was developed that includes a variety of resources for in-depth analyses, including functional annotation, interaction networks or multiple sequence alignment visualization. Finally, we demonstrate the usefulness of our method, by applying it to different use cases, notably the detection of cilia-related proteins in Eukaryotes, and sulfur oxidation-related proteins in Bacteria, as well as by using it to compare various groups of species in different life domains.

Materials and Methods

Definition of Differential Conservation

We define differential conservation as the unexpected divergence that can be observed between taxonomic groups in an otherwise well-conserved protein family, which can correspond to a diverging or missing region of variable size in the sequences of specific species. This can be due to varying evolutionary pressures between clades, resulting in a higher rate of sequence evolution leading to variations along the protein sequence or in the complete or partial gain/loss of one or several proteins. Complete protein gain/loss can be detected either by searching for homologous sequences through BLAST searches, or by predicting orthologous relations with dedicated programs such as OrthoInspector 3.0 (Nevers et al. 2019). In the case of partial protein gain/loss or sequence divergence, relative conservation between groups in a protein family can be inconsistent with what is expected based on the species tree. The proposed approach is based on the analysis of the respective conservation of two groups of closely related species compared with a more distant query species used as a reference. For instance, we can estimate the relative conservation of two groups of Teleost fish (e.g., Otomorpha and Euteleosteomorpha) to *Homo sapiens*. The two groups of Teleost fish are expected to have a similar conservation when compared with human. If one group of Teleost fish is significantly closer to human than the other in a given protein family, it may reflect a case of differential conservation.

For the two chosen sister groups of species, a comparison is done to establish a baseline behavior of conservation in the whole proteome, which can then be used to highlight cases where the conservation is atypical. Relative conservation and taxonomic proximities compared with a query species can be assessed by using BLAST homology searches as a proxy. By using a more distant reference species, we ensure that for most protein families, the two selected taxa of interest should be indistinguishable from one another in a BLAST result (i.e., in the same range of ranks in a BLAST output), whereas in proteins presenting an atypical conservation pattern, there should be a clear separation between the two groups (fig. 1).

Relative Conservation at the Protein Family Level

BLAST homology searches are computed, for a complete query proteome (used as reference species) and each BLAST result is then processed individually, with the first hit of each species from both selected groups extracted, under the hypothesis that the sequences are homologous to the query sequence. Alternatively, BLAST can be used in conjunction with orthology program OrthoInspector 3.0 (Nevers et al. 2019), and hits corresponding to predicted orthologs in species of both groups can be considered.

For each homolog or ortholog detected in the two groups, BLUR retrieves various statistics of the BLAST hits (e.g., E-value, rank of hit in the BLAST, start and end positions of the pairwise alignment, etc.) and compares the average conservation behavior of both groups for each protein family (fig. 2). To avoid any bias caused by accelerated evolutionary rate in an individual species or by badly predicted sequences (missing or mispredicted exons, missed genes/exons boundaries, etc.), hits where ranks are detected as group outliers by Tukey's fences statistical method using a 1.5 interquartile range are not taken into consideration for the calculations (Tukey 1977). Comparisons are only executed for proteins where, for each of the groups, hits were found for at least 50% (33% for orthologous sequences) of the species that are available in the BLUR database (see below). These cutoffs were introduced to avoid biases caused by the detection of sequences in only a few species, which could correspond to paralogs. The cutoff for orthology searches is lower due to the stringency of the inference method compared with the homology search.

For each protein family, the relative conservation of the two groups of species is evaluated according to three parameters:

- The ratio between the mean (in log space) of the E-values of both groups

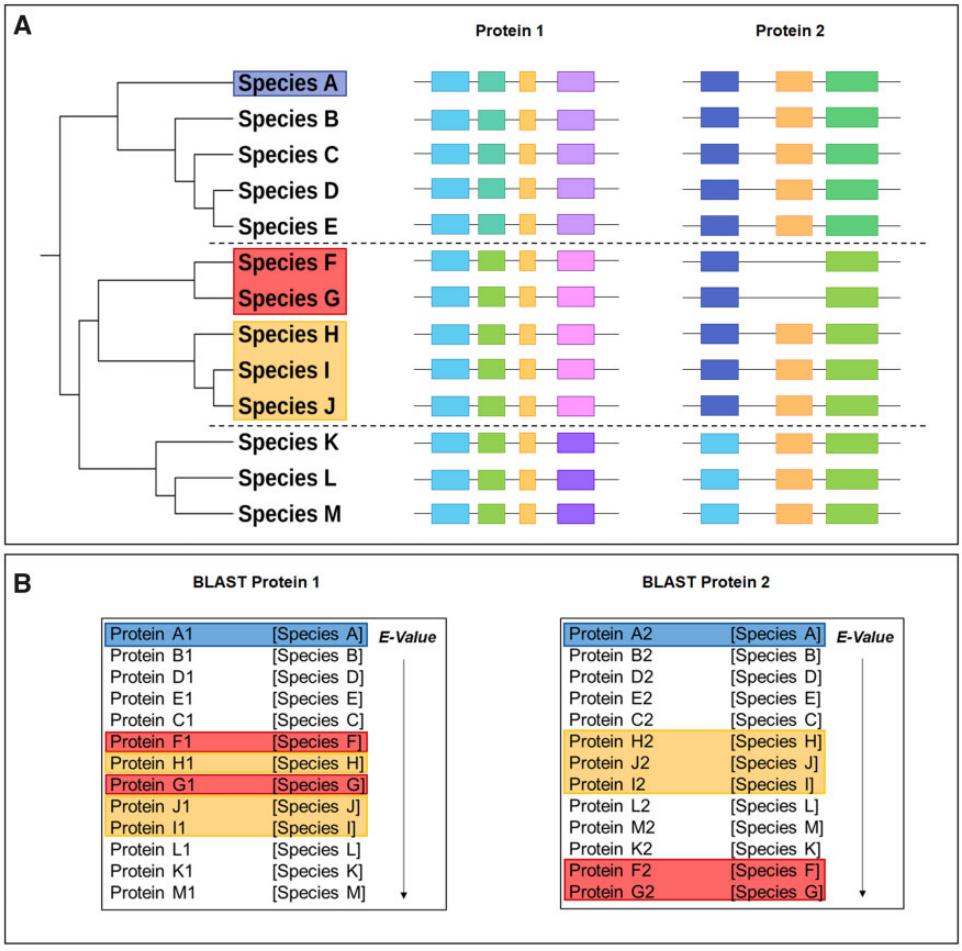

Fig. 1.—Schematic representation of the proposed approach. (*A*) The relative conservations for two proteins (1 and 2) in 13 different species. Colored blocks represent conserved sequence regions (blocks). A variation of hue between two blocks of the same color indicates a small divergence in sequence. Protein 1 shows expected taxonomic variations. For protein 2, the orange block is missing in species F and G. (*B*) The BLAST results for proteins 1 and 2 using Species A as the query. In the Protein 1 BLAST, the species F, G and H, I, J are ranked together, since their respective sequences are similar. In the Protein 2 BLAST, Species H, I, and J are ranked similarly to Protein 1, whereas species F and G are ranked further down due to the missing orange block.

- The difference between the mean distances to the query of each group, where the distance is defined as:

$$1 - \frac{(\text{end}_q - \text{start}_q) - \text{mismatches} - \text{gap}_o\text{pens}}{\text{length}_q}.$$

  With end_q, start_q, mismatches, gap_opens, and length_q being the position on the query where the aligned hit ends, the position on the query where the aligned hit starts, the number of mismatches in the alignment, the number of gaps opened, and the length of the query sequence, respectively.
- The percentage of hits of one group ranked higher in the BLAST than the other groups' best-ranked hit.

Detection of Outliers at the Proteome Level

The distributions of these parameters for the complete proteome are then analyzed using Tukey's fences method with a 1.5 interquartile range. Protein sequences with outlying values compared with the standard conservational behavior in the whole proteome are classified into two categories: "High priority," if all three criteria are detected as outliers, and "Mid priority," if only two out of the three criteria are detected as outliers. Proteins with no hits in one or both groups of species are classified in a third category. These three categories can then be analyzed in depth using various tools (see below).

BLUR Databases

BLAST searches have been precalculated with default parameters for 27 different query species (15 Eukaryotes, 8 Bacteria, and 4 Archaea) in protein databases of the corresponding life domain (e.g., eukaryote queries on a database containing only eukaryotic proteins, etc.), using BLAST+ 2.5.0 (Camacho et al. 2009) with an E-value

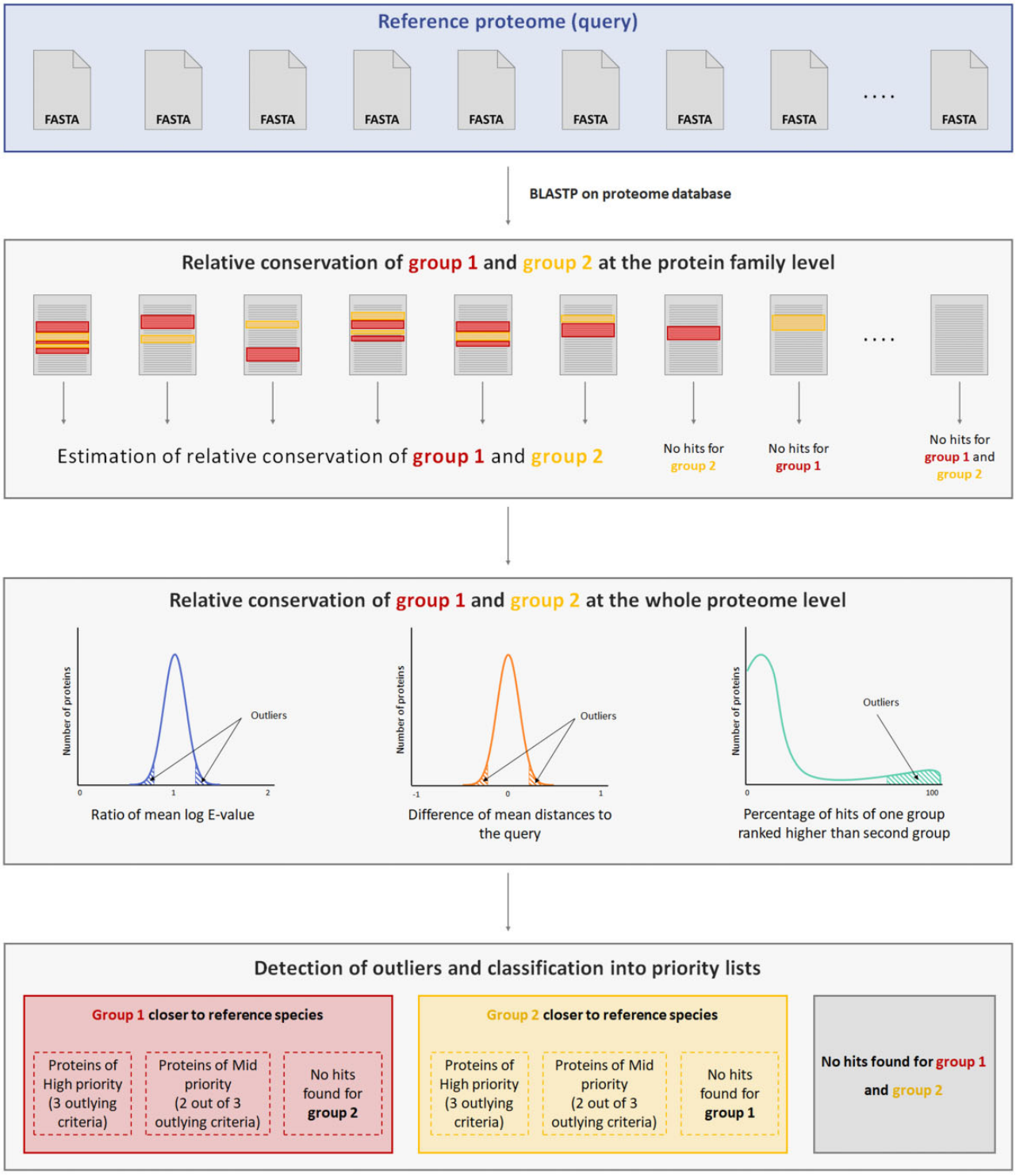

Fig. 2..—Schematic representation of the BLUR protocol. A reference proteome is compared with a proteome database with BlastP, and the results are stored in a database (not shown here). For each user-selected groups 1 and 2, BLUR establishes the relative conservation of both groups for each protein using three criteria: ratio of mean E-value in log space, difference of mean distance to the query, and ranking of one group compared with the other. The relative conservation is then analyzed on the whole proteome level, and outliers are detected using Tukey's fences method, and classified into priority lists.

Table 1
Query Species Available in BLUR for Each of the Three Life Domains, with the Number of Proteins in the Proteome Used

Domain	Query Species (Taxonomy ID)	Number of Proteins	Life Group (number of species)
Eukaryota	*Homo sapiens* (9606)	21,044	Opisthokonta (557)/Metazoa (169)
	Mus musculus (10090)	22,298	
	Xenopus tropicalis (8364)	24,125	Metazoa (169)
	Drosophila melanogaster (7227)	13,780	
	Caenorhabditis elegans (6239)	19,990	
	Saccharomyces cerevisiae (559292)	6,049	Fungi (384)
	Schizosaccharomyces pombe (284812)	5,142	
	Cryptococcus neoformans (214684)	6,601	
	Arabidopsis thaliana (3702)	27,619	Viridiplantae (73)
	Chlamydomonas reinhardtii (3055)	14,266	
	Cyanidioschyzon merolae (280699)	4,995	Eukaryota (734)
	Plasmodium falciparum (36329)	5,340	
	Dictyostelium discoideum (44689)	12,731	
	Leishmania major (5664)	8,031	
	Ectocarpus siliculosus (2880)	15,903	
Bacteria	*Thermotoga maritima* (243274)	1,852	Bacteria (3,846)
	Bacillus subtilis (224308)	4,260	
	Streptomyces coelicolor (100226)	8,038	
	Treponema pallidum (243276)	1,027	
	Chlamydia trachomatis (272561)	895	
	Escherichia coli (83333)	4,347	
	Bacteroides thetaiotaomicron (226186)	4,782	
	Aquifex aeolicus (224324)	1,553	
Archaea	*Nanoarchaeum equitans* (228908)	536	Archaea (179)
	Pyrococcus abyssi (272844)	1,788	
	Sulfolobus solfataricus (273057)	2,938	
	Candidatus Thorarchaeota archaeon SMTZ1-45 (1706444)	3,208	

NOTE.—The last column indicates in which life group the query species can be used, as well as the number of species in the group.

threshold of 1.0e-3 and a maximum of 5,000 hits (table 1). Reference species were selected to offer a broad coverage of the tree of life and allow users to study any specific groups of organisms. The Eukaryota, Bacteria, and Archaea databases comprise 734, 3,863, and 179 complete proteomes respectively, from the Uniprot reference proteomes (Bateman et al. 2017) (Downloaded in November 2016) and the RefSeq database (O'Leary et al. 2016) (Downloaded in October 2017). The proteomes included in the database are the same as those found in the OrthoInspector 3.0 database and were selected based on the following criteria of quality: protein number, low proportion of small proteins (<100 amino acids), proteins that do not start with a methionine. The last two criteria are used to estimate the number of fragmentary proteins in the proteomes to filter out low-quality ones. Proteomes of Archaea and Bacteria with >20% of small proteins and/or 10% of false-start proteins and/or >10% proteins annotated as fragments were excluded. For Eukaryotes, the same threshold was used for small proteins content and proteomes with >55% of false start proteins were excluded. The BLUR relational database contains information (e.g., associated gene name, description, sequence length), for all the proteins available in the various proteomes used as queries for the BLAST searches (table 1). It also stores conservation features pertaining to the first homologous or orthologous hit of each species (e.g., percent identity to the query, length of the BLAST pairwise alignment, E-value, taxonomic id of the associated species, etc.) for all BLAST searches. When several high-scoring pairs exist in a BLAST output, the Expect value of the best hit is kept, but the number of gaps, mismatches, and the alignment length are recalculated according to the overlapping ratio of the different existing HSPs in order to calculate a distance to the query as accurately as possible. Orthologous relations were predicted with OrthoInspector 3.0 and used to select relevant hits when populating the database with the results of the BLAST searches. In this case, ortholog relations were retrieved from the OrthoInspector resource for each query sequence and only BLAST hits corresponding to orthologs are selected to fill the database and ranked according to BLAST outputs. The NCBI taxonomy (Federhen 2012) was used both in the BLAST searches, and in the database to enable an easy manipulation of the data and retrieval of target hits.

Web Implementation

To make BLUR user-friendly, a web interface was developed using the Symfony PHP web application framework (https://symfony.com/), with the Twig template engine (https://twig.symfony.com/). The website offers the opportunity to perform both global and individual analyses of the results, as well as the possibility to export the results in a CSV file. For the various lists of results, protein interaction networks can be generated using data from the STRING database when available (Szklarczyk et al. 2019), containing only direct interactions between proteins of the lists with a score greater than 0.7, and Gene Ontology (GO) enrichments can be computed using the Panther API (Mi et al. 2019). Individual analyses provide information about each protein detected by BLUR, with GO annotations, protein domain annotations provided by the InterPro webservice (Mitchell et al. 2019) and links to external resources such as UniProt and OrthoInspector. We also provide a multiple sequence alignment precomputed using DBClustal (Thompson et al. 2000) containing up to 2,000 homologous sequences and a visual representation of the BLAST result. The generated networks, GO enrichments, and the precomputed multiple sequence alignments can be exported from the website, as SIF, text, and TFA files respectively.

Results

To address the need for a method capable of detecting both complete protein gain/loss and block-level divergences in a group of species, we developed a new approach based on BLAST homology search results designed to highlight atypical conservation patterns between orthologs or homologs. To facilitate both the use of BLUR and the analysis of the results, we developed a web interface that includes a variety of tools.

BLUR Webserver

The home page of the website (http://lbgi.fr/blur/) shows the three steps necessary to run a BLUR analysis (fig. 3). The first step is the selection of the life domain in which the species of interest belong, and the query species (reference) to use for the BLAST searches using a drop down menu. In order to represent a large taxonomic diversity, 27 species spanning the three life domains are available as queries. The reference species should be chosen to be distant enough from both groups of interest so that in most cases they appear undistinguishable in a BLAST search. In other words, the two groups must share a more common ancestor than the one they share with the reference species. The second and third steps are the selection of the two groups of species to be compared. For each group, the user can choose several species, a single clade, or several clades, using a search bar containing an autocomplete feature. Only species belonging to the selected life group can be chosen. To help in the selection of groups, BLUR can automatically determine a set of possible second groups containing at least three species, according to the taxonomy of the user-defined first group. In this case, BLUR will propose taxa sharing a common ancestor with the first group and containing at least 3 species present in the database. If more than one clade is selected, BLUR first retrieves their common ancestor, and looks for 1) other children taxa of the common ancestor containing at least 3 species and 2) sister clade to the common ancestor with at least 3 species present in the database. Lastly, the user has the possibility of choosing whether to use only orthologs computed with OrthoInspector, or extend the search to homologs found in the BLAST search.

The results obtained from the BLUR software are presented on a Results page in three sections. The first section contains a list of proteins where the second group is closer to the query species than the first group. The second section contains a list of proteins where the first group is closer to the query species than the second group. The third section contains a list of proteins where no hits were found in the BLAST for either groups. The first two sections are divided into three subcategories: absence of homolog/ortholog, High priority, and Mid priority proteins. The two latter correspond to differentially conserved proteins fulfilling respectively three or two BLUR criteria of differential conservation.

For each of the three blocks, and each subcategory within these blocks, interaction networks and GO term enrichments can be generated. Selecting an individual protein in any of the lists will open a protein page containing diverse information. Firstly, a header provides general data on the protein such as the associated gene name, the protein description, the length of the protein, links to external resources, GO terms, and InterPro domains associated with the protein. Secondly, the user can access BLUR-specific data: a representation of the BLAST output with the hits of both groups highlighted for easier analysis and a multiple sequence alignment. This alignment, displayed with the MSAViewer library (Yachdav et al. 2016), contains a subset of sequences of both groups of species, the query species as well as sequences of a few organisms related to the query. It is also possible to display a more complete multiple sequence alignment, containing up to 2,000 homologous sequences, and in this case, species of interest will be highlighted.

The BLUR approach has been tested on different groups of species, demonstrating the advantages of combining subprotein level and protein level information, in order to highlight lineage specialization and obtain a comprehensive view of genotype/phenotype correlations. Two examples of studies performed on the BLUR website are presented below: prediction of cilia-related proteins in Eukaryotes, and prediction of proteins involved in sulfur oxidation in Bacteria.

Use Case: Cilia-Related Proteins in Fungi

Cilia are small microtubule-based organelles present in the Last Eukaryotic Common Ancestor that exhibit an unusual

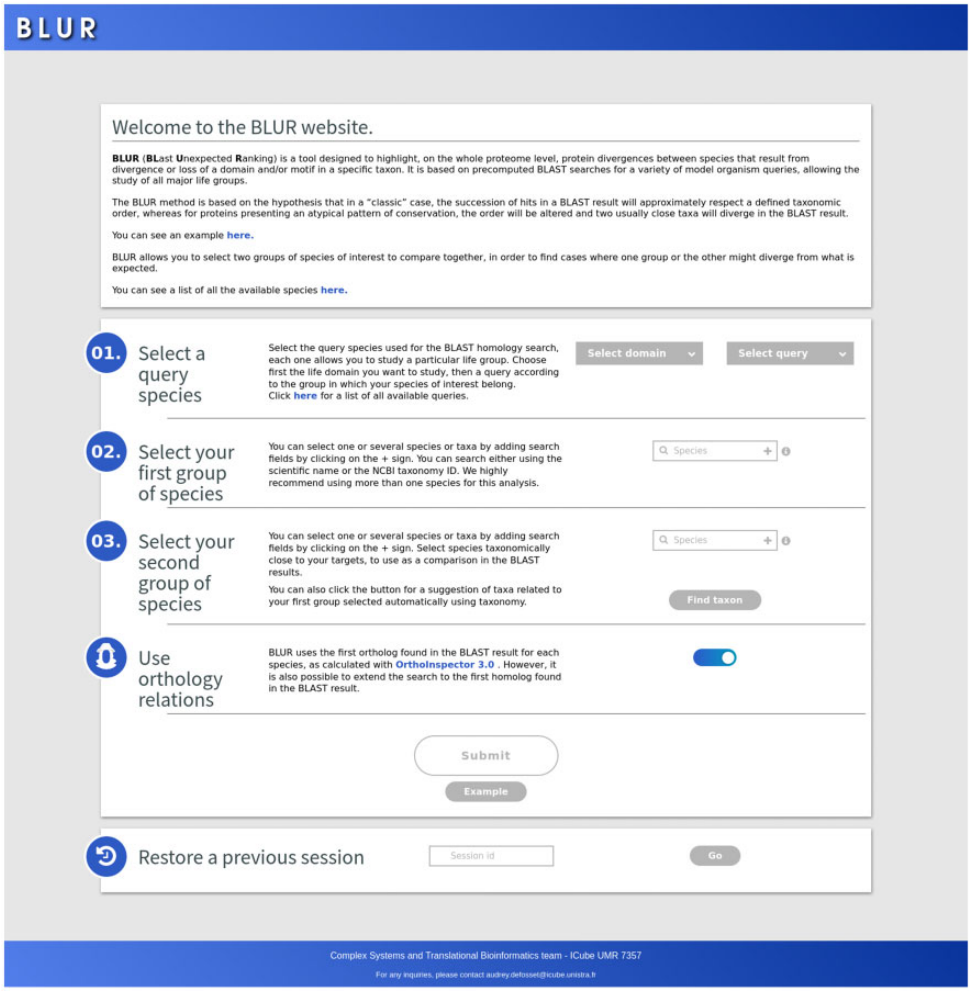

Fig. 3..—Home page of the BLUR website with the different steps necessary to run BLUR. Step 1 allows the user to select one of the three life domains (Eukaryota, Bacteria, Archaea), then the query species used for the BLAST search, as well as the life group to study. Step 2 allows the user to select the first group of interest, which can either be a clade, several species, or several clades, but must be in the life group selected in Step 1. Finally, Step 3 consists in the selection of the second group to be compared, which can either be chosen by the user, or automatically using taxonomy. The last step is the selection of the type of relations to use for the BLAST computation: orthology (default) or homology. The user can also restore a previous session using a session ID provided on the result page.

evolutionary history with various independent losses in the eukaryotic lineage, which makes them a good candidate for comparative genomics studies. Most Fungi are devoid of cilia, with a few known exceptions, namely Chytridiomycota, Blastocladiales, and *Rozella* (Adl et al. 2012). We used our method to identify cilia-related proteins with the assumption that in ciliated Fungi, proteins linked to cilia should be more similar to their metazoan homologs than to their homologs found in nonciliated fungal species.

We chose Opisthokonta as the life group of interest, with *H. sapiens* as the query proteome. We used Chytridiomycota, Blastocladiales, and *Rozella* taxa as the first group (with a total of six species), and Dikarya (350 nonciliated species) as the second group, using ortholog sequences.

For the category corresponding to our hypothesis, where ciliated Fungi proteins are closer to Human than Dikarya, 1,081 proteins were absent in Dikarya, 18 were classified as High priority, and 81 as Mid priority. A manual analysis of the multiple sequence alignments showed the presence of divergent regions in most proteins, with 12 false positives found in the Mid priority list, due to either an insufficient number of sequences, or the presence of low-quality sequences. As an example, the multiple alignment of RFX1 is provided as supplementary data, Supplementary Material online, showing the presence of only 3 badly predicted sequences of ciliated Fungi. A GO enrichment analysis of the 1,180 proteins showed that they were significantly enriched in terms related to cilia, such as "cilium" (*P* value: 2.23E-74) or "intraciliary transport" (*P* value: 2.05E-22). To further assess the quality of our results, we compared the 1,180 proteins to a negative set of 971 proteins from pathways unlikely to be related to cilia constructed in a previous study (Nevers et al. 2017). Only 22

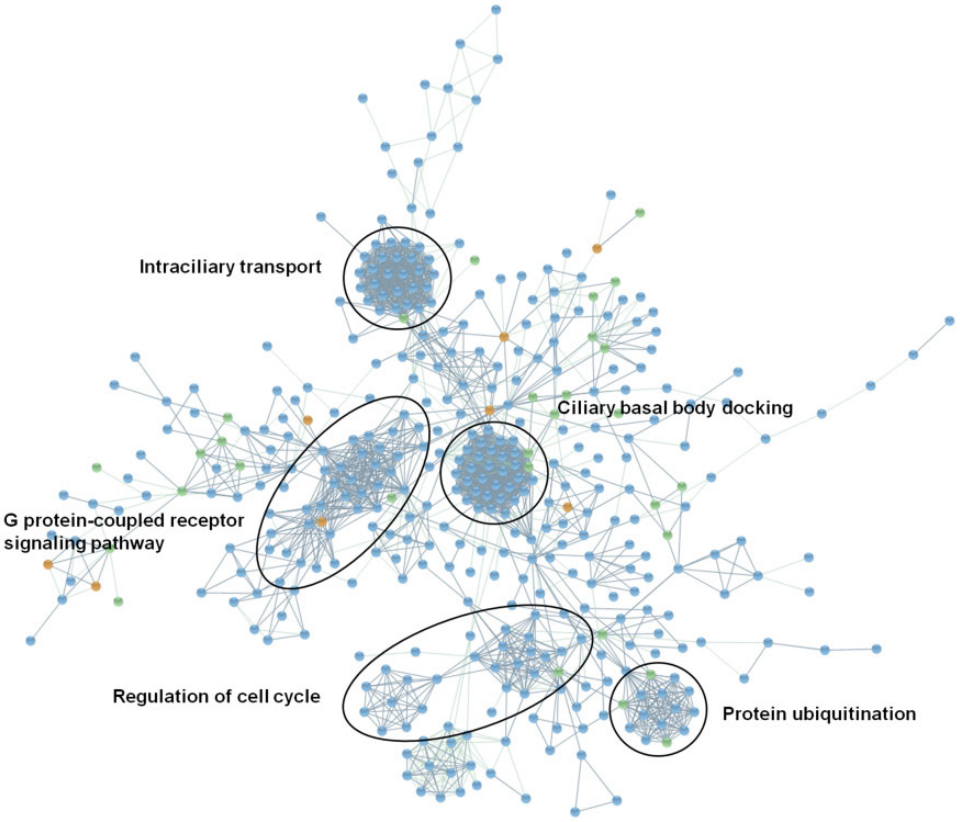

Fig. 4..—Main interaction network of proteins absent in Dikarya (blue nodes), and proteins predicted to have differential conservation with High priority (orange nodes) or Mid priority (green nodes). The network contains highly linked clusters of proteins that are both absent and divergent in Dikarya, and that are enriched in GO terms corresponding to ciliary components, thus validating the proposed method.

proteins of this negative set were included in the 1,180 proteins, with 2 in the High priority list, and 2 in the Mid priority.

Of the 1,180 proteins detected, 526 presented a high confidence interaction with at least one other. The interaction networks generated showed one main network of 400 proteins consisting of several highly linked clusters, including ones enriched in intraciliary transport, centriole elongation and basal body docking, and cell proliferation regulation (fig. 4). Among the 400 proteins present in the main network, 362 are absent in Dikarya, including 76 related to cilia previously detected using a phylogenetic profiling method (Nevers et al. 2017). The other 38 proteins present in the network come from both the High priority list (orange nodes in fig. 4) and the Mid priority list (green nodes in fig. 4). Thus, these proteins are present in Dikarya, but exhibit a probable differential conservation. About ten of them are already annotated as related to cilia, whereas the other 28 represent potential new cilia-related candidates. Many clusters include both proteins that are totally absent in nonciliated Fungi and proteins that are differentially conserved at the subgene level, illustrating the interplay of these levels of differential conservations and the relevance of our approach.

Among the 99 proteins in the High and Mid priority lists (including the 38 proteins found in the interaction network), 17 had annotations linked to cilia, centrosome, centriole, or microtubule, of which at least 14 presented a clear differential conservation confirmed by visual inspection of the multiple alignment. A particularly striking example is ARMC4, a ciliary protein involved in left/right symmetry and axonemal outer dynein arm assembly, with homologs found in most eukaryotic clades, including Metazoa and Fungi. A multiple sequence alignment of the ARMC4 family showed a clear distinction between the sequences of ciliated versus nonciliated Fungi, with a higher similarity between vertebrate sequences and ciliated Fungi sequences (fig. 5). In particular, Vertebrates and ciliated Fungi proteins present a long N-terminal region that could constitute a yet undiscovered functional domain, whereas nonciliated Fungi proteins have a much shorter sequence.

Finally, we compared the results presented above to the results found when doing the same search but using homology relations. Although using homology, we obtained a list of 1,122 proteins, among which 868 are absent in nonciliated Fungi, 78 are of High priority, and 176 of Mid priority. A comparison with the negative gene set previously used showed an overlap of 27 genes, 2 of which were in the High priority list, and 7 in the Mid priority list. Using homology relations thus appears to be more permissive with more false

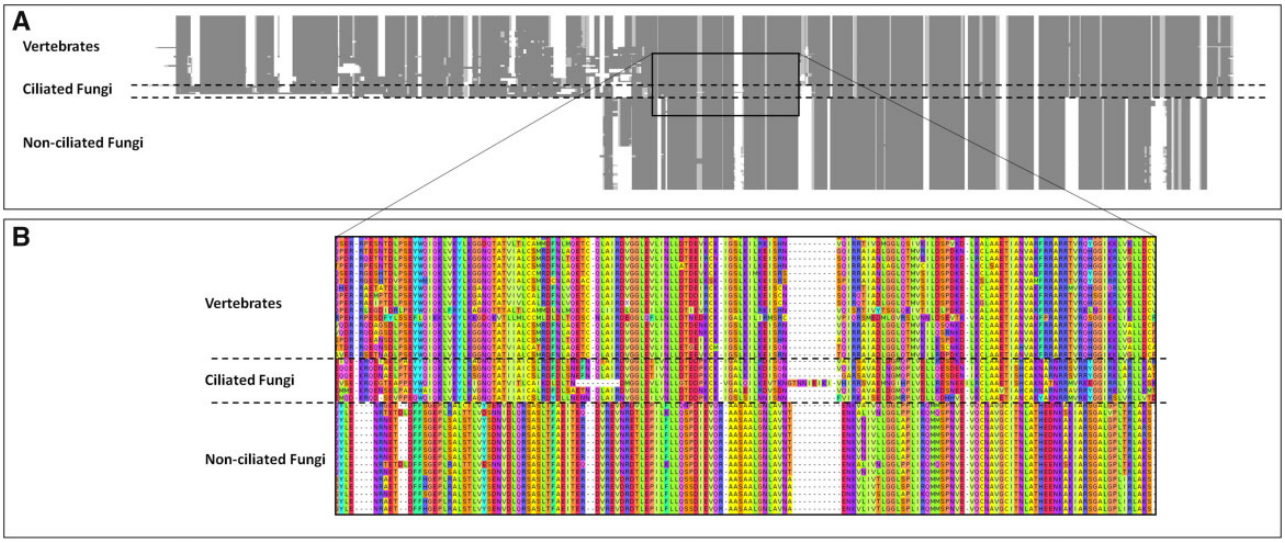

FIG. 5..—Multiple sequence alignment of ARMC4. (*A*) Overview of the multiple sequence alignment of ARMC4. Vertebrates (ciliated species) and ciliated Fungi sequences are similar with a long N-terminal domain that is absent in nonciliated Fungi. (*B*) Zoom on a portion of the alignment where differential conservation can be observed. Ciliated Fungi are very similar to Vertebrates, whereas other, nonciliated Fungi are more divergent.

positives. In return, it increases the number of genes linked to cilia as attested by a still better functional enrichment ("cilium," *P* value: 1.3E-78).

Use Case: Sulfur Oxidation in Bacteria

In certain ecosystems, hydrogen sulfide is more abundant than oxygen, allowing certain microorganisms to use sulfur as a means to produce energy. Sulfur oxidation is performed almost exclusively by Archaea and Bacteria, with a few eukaryotic exceptions. Here, we used BLUR to predict proteins related to sulfur oxidation in Bacteria, using the known sulfur-oxidizing Bacteria *Aquifex aeolicus* as a query proteome. We selected two close groups of Gammaproteobacteria for comparison, with one group able to oxidize sulfur (Chromatiales) and the other not (Enterobacterales). Our hypothesis is that most proteins from Chromatiales are highly similar to their orthologs in Enterobacterales and more divergent compared with *Aquifex* orthologs. In contrast, proteins involved in sulfur oxidation should be highly similar between Chromatiales and *Aquifex*, and very different from the orthologs (if any) found in Enterobacterales.

Using BLUR, we detected 223 proteins in the category where Chromatiales are closer to *Aquifex* than Enterobacterales, with 186 absent in Enterobacterales, 16 classified as High priority, and 21 as Mid priority. As for the previous example, a manual analysis of the multiple sequence alignments showed divergence in most cases, with 6 false positives in the Mid priority list. A GO enrichment analysis of these 223 proteins was not useful due to the lack of GO annotations for the majority of *Aquifex* proteins. However, the interaction networks showed the presence of several clusters (fig. 6). To investigate further the functions associated with these clusters, we used ortholog annotations provided by OrthoInspector (Nevers et al. 2019). We identified the Sox protein cluster (fig. 6), essential for sulfur oxidation that includes proteins absent from the Enterobacterales group (SoxAX, SoxF, SoxW, SoxX, SoxY, SoxZ) and also the High priority SoxB protein, well conserved in Chromatiales but highly divergent in Enterobacterales. The dimethyl sulfoxide (DMSO) reductase associated with the Sox cluster was also detected with DmsA, DmsB1, and DmsC protein subunits classified as High priority, Mid priority, and absent in Enterobacteriales respectively.

We also identified a large iron–sulfur protein cluster (fig. 6), containing the proteins from the hdr gene cluster (*dsrE2A*, *dsrE3B*, *dsrE3C*, *hdrA*, *hdrB1*, *hdrB2*, *hdrC1*, *hdrC2*), known to be involved in sulfur oxidation (Quatrini et al. 2009; Boughanemi et al. 2016), which were found to be absent in Enterobacterales. Other proteins with no known interactions were found to have a clear distinction between Chromatiales and Enterobacterales sequences, such as Peroxiredoxin, which was verified using a multiple sequence alignment. We do not have a benchmark to assess the specificity of this analysis, especially since the oxidation pathways are extremely variable, even within the same genus (Berben et al. 2019) and other cellular processes may vary between Chromatiales and Enterobacterales. However, we were able to detect a loss or a differential conservation in Enterobacterales for all the 5 genes of the Sox system reported as the core pathway in the oxidation of sulfur, as well as for the 5 genes of the Hdr systems tightly coupled to the SOX system in a majority of sulfur-oxidizing organisms (Watanabe et al. 2019). Additional well-known genes linked to core sulfur oxidation pathways are also detected, further attesting of the sensitivity of our approach.

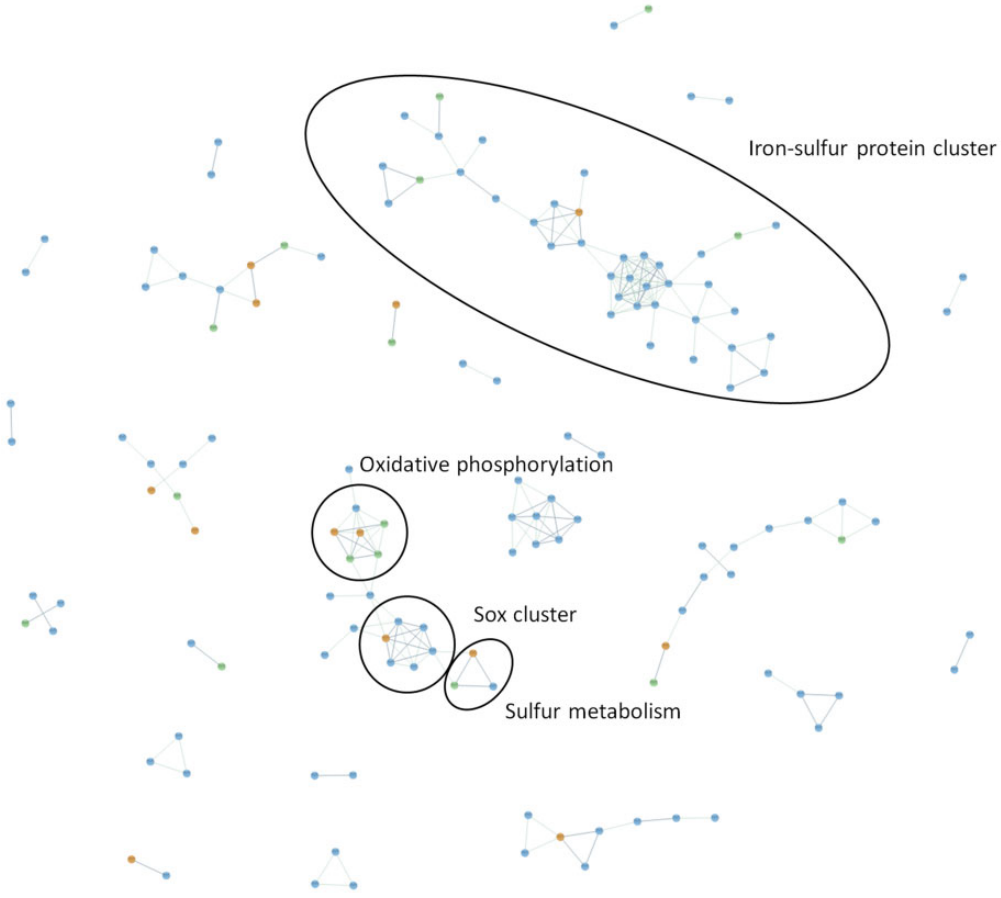

Fig. 6..—Interaction networks of proteins absent in Enterobacterales (blue nodes), of high priority (orange nodes) and of mid priority (green nodes). Several clusters contained over ten proteins with high confidence links between them, including a cluster containing the main Sox proteins, and a cluster corresponding to the iron–sulfur proteins found in the hdr cluster.

Examples of Proteome Comparisons without Prior Knowledge

We have shown with two use cases that BLUR can be used to study phenotypes of interest by comparing species that present a specific character and species devoid of that character. More generally, BLUR can be used to compare two groups of species without focusing on any specific process. Table 2 shows the results obtained when performing various searches on the BLUR webserver, using different query species in different life domains.

In all examples, BLUR detected proteins that were absent, and proteins that showed divergences of both high and mid priority, with significant functional enrichments in all lists. Most of the networks generated showed highly linked clusters of proteins that are both absent and divergent, with GO enrichment in specific biological functions. These functional links between families showing loss/gain of a complete gene and differential conservation at the subgene level highlights the added value of our approach compared with an analysis based on the sole presence/absence of genes.

Discussion

BLUR represents an online resource capable of rapidly detecting differential conservation from BLAST search results at the whole proteome level, in any of the 4,776 species available in the precalculated database. Our original approach addresses the problems generated by variable evolutionary rates between taxa, by using a reference species to perform relative comparisons and establishing an average conservation behavior over a whole proteome. It is, in this way, similar to relative-rate tests used to compare evolutionary rates between species to assess the existence of molecular clocks by comparing two ingroups and an outgoup (Kumar 2005). These comparisons can be performed among orthologs or homologs; while using orthologs allow for a more restricted search and limit the false positives that could be attributed to the detection of close paralogs, it can also create false negatives due to the problems of orthologs inference caused by highly diverging sequences. These sequences could be detected by using homologs, although the presence of hidden paralogs could introduce a bias in the results.

Table 2
Examples of Application of BLUR Using Various Query Species and Groups of Interest

Query species	Comparison	Protein lists	GO enrichment	Network	Network enrichment
Homo sapiens	Basidiomycota over Ascomycota	469 absent in Ascomycota, 32 High priority, 112 Mid priority	RNA processing (*P* value: 2.12E-10) Protein modification process (*P* value: 3.17E-9) RNA splicing (*P* value: 3.04E-8)	Main network of 208 proteins: 140 absent, 14 High priority, 54 Mid priority	Several clusters: mRNA splicing ; ribosome biogenesis; regulation of signal transduction
Mus musculus	Lophotrochozoa over Ecdysozoa	775 Absent in Ecdysozoa, 23 High priority, 105 Mid priority	Nervous system process (*P* value: 1.34E-12) Sterol metabolic process (*P* value: 5.62E-7) Cilium assembly (*P* value: 1.37E-6)	224 Proteins with a least one interaction: 177 Absent, 10 High, 37 Mid priority)	Several small networks: steroid biosynthetic process; regulation of apoptotic process; cilium assembly; cell cycle
Chlamydomonas reinhardtii	Liliopsida over Eudicotyledons	107 Absent in Eudicotyledons, 18 High priority, 81 Mid priority	Photosynthesis (*P* value: 2.25E-10) Oxidation-reduction process (*P* value: 1.41E-9)	44 Proteins with at least one interaction: 15 absent, 7 High priority, 22 Mid priority	Photosynthesis
Escherichia coli	Betaproteobacteria over Alphaproteobacteria	252 Absent in Alphaproteobacteria, 5 High priority, 28 Mid priority	Pilus organization (*P* value: 5.31E-16) Submerged biofilm formation (*P* value: 2.69E-6)	Main network of 91 proteins: 77 absent, 2 High priority, 12 Mid priority	Several clusters: cell motility; pilus organization; asexual reproduction
Bacillus subtilis	Selenomonadales over Veillonellales	635 Absent in Veillonellales, 23 High priority, 34 Mid priority	Locomotion (*P* value: 7.65E-15) Chemotaxis (*P* value: 1.63E-7)	Main network of 401 proteins: 364 absent, 18 High priority, 19 Mid priority	Several clusters: spore germination; locomotion; antibiotic metabolic process

Although our approach is not as precise as one based on multiple sequence alignments would be, as it is a proxy for relative conservation, it has the large advantage of being able to process complete proteomes in a small amount of time. To assess this relative conservation, we chose three criteria derived from BLAST similarity search results. We selected E-value rather than bit-score, as tests showed that although mean bit-score ratio and mean E-value ratio were similar, E-values were generally more homogenous for species groups in BLAST results, thus allowing outlying values to be detected more easily. Distance to the query and E-value are partially dependent features, but the distance criteria takes into account alignment length, giving us the opportunity to detect potentially missing regions more easily. Finally, ranks are used to confirm that variations that are detected for distances and E-values are indeed due to diverging pattern on a clade level and not to a subset of sequences. The homogeneity of the distribution of values for the three criteria used in BLUR (supplementary fig. 1, Supplementary Material online) is well adapted to outlier detection using Tukey's fences and tests done using different interquartile range values showed 1.5 as the value with the least amount of false-positive and false-negative results. We provide an accessible and easy to navigate website, with a substantial amount of complementary information that allows for more in-depth analysis. We have shown that our method is not limited to any specific biological process or life domain, by identifying cilia-related proteins in Eukaryotes, as well as proteins related to sulfur oxidation in Bacteria. Both examples demonstrate the usefulness of an approach combining complete protein loss/gain and subprotein variation by presenting results containing clusters of strongly interacting proteins that were both completely lost and only partially divergent in some regions. Further tests were done comparing two groups of fish, Otomorpha and Euteleosteomorpha (data not shown) showed that our method is capable of detecting subprotein divergences of varying sizes, from large regions down to single amino acids (fig. 7).

It is difficult to estimate the sensitivity and specificity of our approach as there are currently no suitable benchmarks for differential conservation detection. Studies have been conducted to assess evolutionary phenotype specializations between species at the protein domain level (Nasir et al. 2014; Sun et al. 2017), however their focus is mostly on the comparison between the three domains of life, which makes them unsuited for a comparison with BLUR.Manual inspection of multiple alignments of proteins detected by our approach showed that in both use-cases, most of the proteins from the High priority list exhibited a more- or less-pronounced differential conservation, with false positives in the Mid priority lists. This manual analysis showed that the precision and the quality of the results are mostly dependent on the number

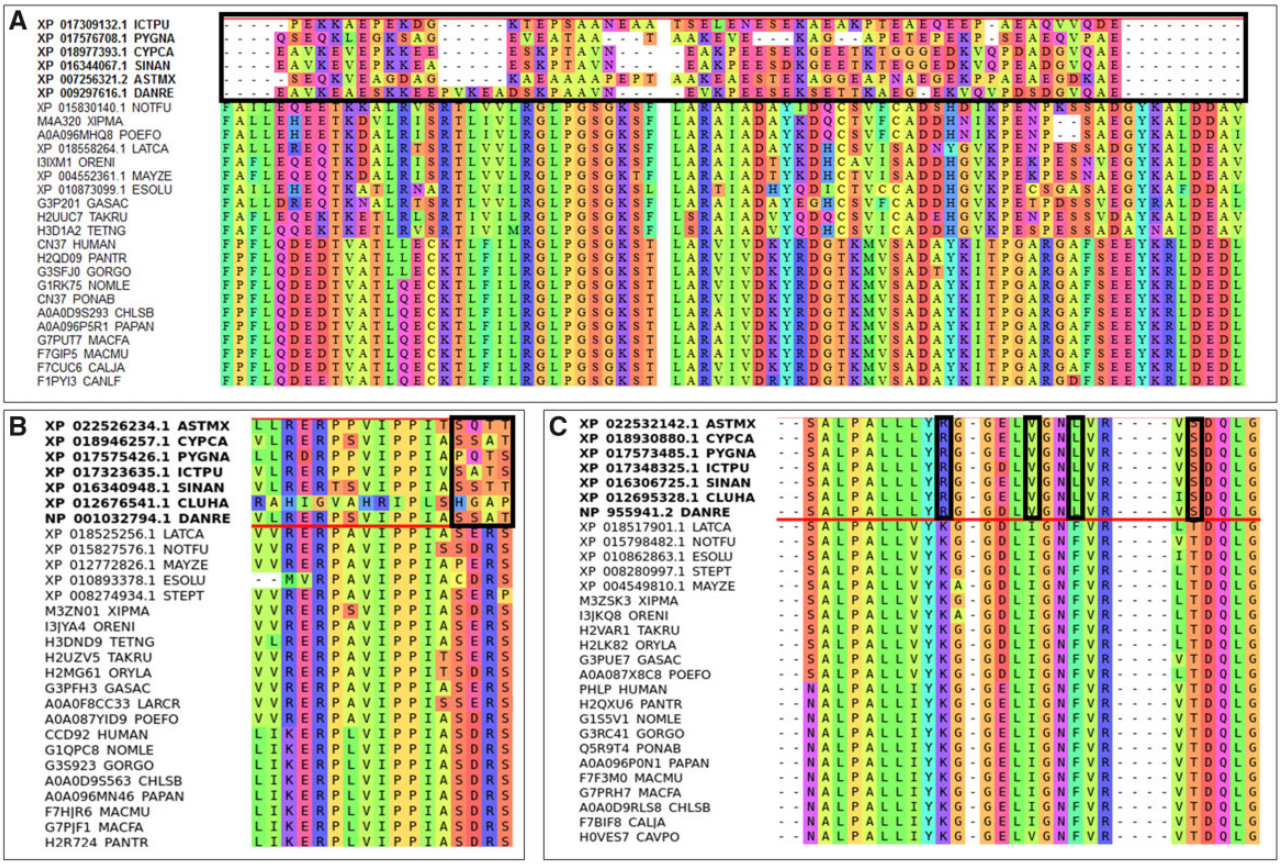

Fig. 7.—Examples of differential conservation detected by BLUR. Comparison was done between two groups of Actinopterygii, Otomorpha (above the red line), and Euteleosteomorpha (below the red line). Homo sapiens was used as a query species, the multiple sequence alignments contain sequences of mammals. (*A*) Multiple sequence alignment of CNP. Differential conservation of a large region can be seen in protein sequences of Otomorpha (*B*) Multiple sequence alignment of CCDC92. Differential conservation of a small motif can be seen in protein sequences of Otomorpha. (*C*) Multiple alignment of PDCL. Differential conservation of single amino acids can be observed in protein sequences of Otomorpha.

of species in each group, and more importantly on the quality of the sequences available. In some cases, one group did not contain enough reliable sequences to properly assess the conservation between the two groups. The quality of the BLUR results are clearly dependent on the parameters chosen (number of species in each group, distance between the query and the groups, complexity of the phenotypic differences between groups), and are entirely correlated with the quality of the sequences available in the database, leading to a small proportion of false positives. We have assessed the impact of query choice on the results of BLUR (data not shown), and it appeared clear that choosing similar query species (e.g., *H. sapiens* and *Mus musculus*) produces similar results. Choosing a query that is too distant to the groups of interest will result in missing candidate genes, as the sequences will have naturally diverged too much over time and relative conservation of the groups of interest will be harder to assess. Similarly, if the query is too close to one group, the sequences will not have diverged enough to detect abnormal conservation patterns. As a general rule of thumb, when comparing two groups with no previous knowledge, we recommend choosing the query species that is the closest to the two groups' common ancestor. When studying a specific phenotype, we recommend selecting a query species and a group that share the phenotype of interest, and select a sister group for comparison that does not possess the phenotype.

In conclusion, we have shown that our method is effective in the detection of proteins related to a given phenotype and to generate relevant new candidates that can be analyzed easily and rapidly with the various tools available on the website. It also opens the way to more specific studies on domain rearrangements and evolution by highlighting potential candidate families for such analyses. Future developments will include the release of the underlying software to allow analysis of user-specific proteomes, as well as the addition of new reference proteomes to extend the comparison possibilities, as well as an extension of the sequence databases with more species to analyze and compare.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank the Bio-statistics, Informatics and Complex System platform (BICS) and BISTRO bioinformatics platforms for informatics support and the European Grid Infrastructure for cloud computing facilities. This work was supported by the IdEx Unistra in the framework of the "Investments for the future" program of the French government and Institute funds from the Centre National de la Recherche Scientifique and the Universitéde Strasbourg.

Data Availability

The data underlying this article are available in the article and in its Supplementary Material online.

Literature Cited

Adl SM, et al. 2012. The revised classification of eukaryotes. J Eukaryot Microbiol. 59(5):429–514.
Anderson DP, et al. 2016. Evolution of an ancient protein function involved in organized multicellularity in animals. eLife 5:e10147.
Bateman A, et al. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45:D158–D169.
Berben T, Overmars L, Sorokin DY, Muyzer G. 2019. Diversity and distribution of sulfur oxidation-related genes in Thioalkalivibrio, a genus of Chemolithoautotrophic and Haloalkaliphilic sulfur-oxidizing bacteria. Front Microbiol. 10:160.
Boughanemi S, et al.. 2016. Microbial oxidative sulfur metabolism: biochemical evidence of the membrane-bound heterodisulfide reductase-like complex of the bacterium Aquifex aeolicus. FEMS Microbiol Lett. 363(15):fnw156.
Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10(1):421.
Cheng Y, Perocchi F. 2015. Prediction of mitochondrial protein function by comparative physiology and phylogenetic profiling. In: Weissig V, Edeas M, editors. Mitochondrial medicine. Vol. 1264. New York: Springer. p. 321–329.
Cromar GL, et al. 2016. PhyloPro2.0: a database for the dynamic exploration of phylogenetically conserved proteins and their domain architectures across the Eukarya. Database 2016:baw013.
Cunningham FX, Lafond TP, Gantt E. 2000. Evidence of a role for LytB in the nonmevalonate pathway of isoprenoid biosynthesis. J Bacteriol. 182(20):5841–5848.
Dey G, Jaimovich A, Collins SR, Seki A, Meyer T. 2015. Systematic discovery of human gene function and principles of modular organization through phylogenetic profiling. Cell Rep. 10(6):993–1006.
Dohmen E, Klasberg S, Bornberg-Bauer E, Perrey S, Kemena C. 2020. The modular nature of protein evolution: domain rearrangement rates across eukaryotic life. BMC Evol Biol. 20(1):30.
El-Gebali S, et al. 2019. The Pfam protein families database in 2019. Nucleic Acids Res. 47(D1):D427–D432.
Federhen S. 2012. The NCBI taxonomy database. Nucleic Acids Res. 40(D1):D136–D143.
Forslund K, Pekkari I, Sonnhammer EL. 2011. Domain architecture conservation in orthologs. BMC Bioinformatics 12(1):326.
Haider C, Kavic M, Sonnhammer ELL. 2016. TreeDom: a graphical web tool for analysing domain architecture evolution. Bioinformatics 32(15):2384–2385.
Han X, Guo J, Pang E, Song H, Lin K. 2020. Ab initio construction and evolutionary analysis of protein-coding gene families with partially homologous relationships: closely related drosophila genomes as a case study. Genome Biol Evol. 12(3):185–202.
Jeffery CJ. 1999. Moonlighting proteins. Trends Biochem Sci. 24(1):8–11.
Jim K. 2003. A cross-genomic approach for systematic mapping of phenotypic traits to genes. Genome Res. 14(1):109–115.
Kim Y, Subramaniam S. 2005. Locally defined protein phylogenetic profiles reveal previously missed protein interactions and functional relationships. Proteins 62(4):1115–1124.
Kress A, Lecompte O, Poch O, Thompson JD. 2018. PROBE: analysis and visualization of protein block-level evolution. Bioinformatics 34(19):3390–3392.
Kumar S. 2005. Molecular clocks: four decades of evolution. Nat Rev Genet. 6(8):654–662.
Lees JG, Dawson NL, Sillitoe I, Orengo CA. 2016. Functional innovation from changes in protein domains and their combinations. Curr Opin Struct Biol. 38:44–52.
Letunic I, Bork P. 2018. 20 years of the SMART protein domain annotation resource. Nucleic Acids Res. 46(D1):D493–D496.
Li JB, et al. 2004. Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene. Cell 117(4):541–552.
Löhr U, Yussa M, Pick L. 2001. *Drosophila fushi tarazu*: a gene on the border of homeotic function. Curr Biol. 11(18):1403–1412.
Mani M, et al. 2015. MoonProt: a database for proteins that are known to moonlight. Nucleic Acids Res. 43(D1):D277–D282.
Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. 2019. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 47(D1):D419–D426.
Mitchell AL, et al. 2019. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res. 47(D1):D351–D360.
Moore AD, Bornberg-Bauer E. 2012. The dynamics and evolutionary potential of domain loss and emergence. Mol Biol Evol. 29(2):787–796.
Moore AD, Grath S, Schüler A, Huylmans AK, Bornberg-Bauer E. 2013. Quantification and functional analysis of modular protein evolution in a dense phylogenetic tree. Biochim Biophys Acta Proteins Proteomics. 1834(5):898–907.
Moore AD, Held A, Terrapon N, Weiner J, Bornberg-Bauer E. 2014. DoMosaics: software for domain arrangement visualization and domain-centric analysis of proteins. Bioinformatics 30(2):282–283.
Nasir A, Kim KM, Caetano-Anollés G. 2014. Global patterns of protein domain gain and loss in superkingdoms. PLoS Comput Biol. 10(1):e1003452.
Nevers Y, et al. 2017. Insights into ciliary genes and evolution from multilevel phylogenetic profiling. Mol Biol Evol. 34(8):2016–2034.
Nevers Y, et al. 2019. OrthoInspector 3.0: open portal for comparative genomics. Nucleic Acids Res. 47(D1):D411–D418.
O'Leary NA, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44:D733–D745.
Pagel P, Wong P, Frishman D. 2004. A domain interaction map based on phylogenetic profiling. J Mol Biol. 344(5):1331–1346.
Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A. 96(8):4285–4288.

Persson E, Kaduk M, Forslund SK, Sonnhammer ELL. 2019. Domainoid: domain-oriented orthology inference. BMC Bioinformatics 20(1):523.
Quatrini R, et al. 2009. Extending the models for iron and sulfur oxidation in the extreme acidophile *Acidithiobacillus ferrooxidans*. BMC Genomics 10(1):394.
Ronshaugen M, McGinnis N, McGinnis W. 2002. Hox protein mutation and macroevolution of the insect body plan. Nature 415(6874):914–917.
Sadreyev IR, Ji F, Cohen E, Ruvkun G, Tabach Y. 2015. PhyloGene server for identification and visualization of co-evolving proteins using normalized phylogenetic profiles. Nucleic Acids Res. 43(W1):W154–W159.
Shiga Y, Yasumoto R, Yamagata H, Hayashi S. 2002. Evolving role of Antennapedia protein in arthropod limb patterning. Development 129(15):3555–3561.
Sun C-T, Chiang AWT, Hwang M-J. 2017. A proteome view of structural, functional, and taxonomic characteristics of major protein domain clusters. Sci Rep. 7(1):14210.
Szklarczyk D, et al. 2019. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 47(D1): D607–D613.
Thompson JD, Plewniak F, Thierry J-C, Poch O. 2000. DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. Nucleic Acids Res. 28(15):2919–2926.
Tukey JW. 1977. Exploratory data analysis. Reading (MA): Addison-Wesley Publishing Company Reading.
Vera-Parra N, Gutiérrez-Ramirez M, López-Sarmiento D. 2016. Automatic construction and graph-making of functional domain architectures. Adv Nat Appl Sci. 10:99–105.
Watanabe T, et al. 2019. Genomes of neutrophilic sulfur-oxidizing chemolithoautotrophs representing 9 proteobacterial species from 8 genera. Front Microbiol. 10:316.
Yachdav G, et al. 2016. MSAViewer: interactive JavaScript visualization of multiple sequence alignments. Bioinformatics 32(22):3501–3503.
Zmasek CM, Godzik A. 2011. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. Genome Biol. 12(1):R4.

Associate editor: Alba Mar