



SuperPartitions: Detection and classification of orthologs

Fredj Tekaia ^{a,*}, Edouard Yeramian ^b

^a Institut Pasteur, Unité de Génétique Moléculaire des Levures (URA 2171 CNRS and UFR927 Univ. P.M. Curie), 25, Rue du Dr Roux, 75724 Paris Cedex 15, France

^b Institut Pasteur, Unité de Bioinformatique Structurale, (URA 2185 CNRS), 25, Rue du Dr Roux, 75724 Paris Cedex 15, France

ARTICLE INFO

Article history:

Accepted 11 October 2011

Available online 20 October 2011

Received by Takashi Gojobori

Keywords:

Evolution
Paralogs
Clustering
Genomics
Conservation profiles

ABSTRACT

The proper detection of orthologs is crucial for evolutionary studies of genes and species. Despite large efforts to solve this problem the methodological situation appears unsettled to a large extent and the “quest for orthologs” is still an ongoing task in large-scale genome comparisons.

Here, we introduce a simple operational framework for the detection of orthologs and their classification. The operational framework relies on well-established principles, optimizing their implementation for the considered purposes, and chaining components in coherent procedures: 1) We take advantage of the efficiency and simplicity of the Reciprocal Best Hit (RBH) detections, remedying (by design) the drawback concerning the limitations in terms of 1:1 detections. The procedure is based on the partitioning of Reciprocal Best Hits, with the further merging of partitions including members of the same paralogous classes (“SuperPartition of Orthologs” (SPOs)). 2) We then resort to the conservation profiles of the obtained clusters, allowing simple detection of SPOs containing duplicated members. Based on accepted evolutionary principles, such members can be further tagged as in-paralogs (co-orthologs) or out-paralogs.

The method is illustrated and validated by extensive genomic analyses. The performances of the overall approach are characterized in global terms for three sets of species (*Chlamydiae*, *Mycobacteria*, *Aspergilli*), showing that at least 75% of the sets of orthologs contain at most one protein from a given species. The sets including more than one protein from a given species are shown to contain in-paralogs in proportions varying from 28% to 58%. The characterizations also show that the large majority of SPOs are associated with ancestral motifs, and accordingly not prone to chaining effects that might be triggered by multi-domain proteins. Further the SPO formulation is compared to other similarity based ortholog detection methods. Beyond core common results, significant differences are observed between various methods, which can be accounted for to a large extent on conceptual grounds, relative to the different merging schemes involved. Such comparisons highlight a major advantage of the SPO approach concerning the proper clustering of associated paralogs, which appear to be often dispatched spuriously into distinct orthologous classes. Finally the perspectives for future applications and elaborations of SPO-based compositional analyses are discussed.

© 2011 Elsevier B.V. All rights reserved.

Abbreviations: RSH, Reciprocal Significant Hit (in intra-species comparisons); RBH, Reciprocal Best Hit (in inter-species comparisons); PRBH, Partition of Reciprocal Best Hit proteins; Pn.m, Partition of n RSH or RBH proteins (m being an arbitrary indexing order); mcl, “Markov Cluster Algorithm”: cluster algorithm for graphs, allowing the clustering of similar proteins; Cp.q, mcl cluster containing p RSH linked proteins (q arbitrary indexing order); SPOr.s, SuperPartition of Orthologs (SPO) containing r proteins (s arbitrary indexing order); PE, *M. tuberculosis* protein family characterized by Proline–Glutamic acid motifs; PPE, *M. tuberculosis* protein family characterized by Proline–Proline–Glutamic acid motifs; MP, MultiParanoid method; OM, OrthoMCL method; COMr.s, OrthoMCL cluster containing r proteins (s arbitrary indexing order); CMPr.s, MultiParanoid cluster containing r proteins (s arbitrary indexing order); eggNOG, “evolutionary genealogy of genes: non-supervised Orthologous Groups”: Orthologs detection and clustering method; LMX1A/B, LIM homeobox transcription factor 1 alpha/beta; RPF1, Ribosome production factor 1; IMP4, U3 small nucleolar ribonucleoprotein.

* Corresponding author.

E-mail address: tekaia@pasteur.fr (F. Tekaia).

1. Introduction

Comparative genome analyses represent a powerful approach for the detection of similarities between different species, notably with the exponentially increasing amount of data generated by genome projects. With this respect one of the primary tasks of evolutionary genomics is the determination of gene families from sets of taxa. The reconstruction of the evolutionary histories of genes and species relies critically on the accurate identification of orthologs. Orthologs i.e. genes resulting from speciation events (Fitch, 1970, 2000) are to be differentiated from paralogs i.e. genes resulting from duplication events before or after speciation. Paralogy is defined with reference to a speciation event (usually the last one that gave rise to the actual species): paralogs resulting from duplications that occurred after that event are called in-paralogs; those resulting from duplications that occurred before this event are called out-paralogs (Sonnhammer and Koonin, 2002).

The proper identification of orthologous genes is a major challenge because accumulation of gene-loss, duplication, horizontal gene transfer, protein domain gain and loss events tend to blur the recognition of true orthologs among the set of remaining homologs (Koonin, 2005; Kristensen et al., 2010). Most methods for large-scale detection of orthologs are based on homology as inferred from sequence similarity. It is generally admitted that large-scale orthology assignment is not a simple task. A large number of methods have been elaborated to solve this problem, with different advantages and limitations (Altenhoff and Dessimoz, 2009; Kristensen et al., 2010; Kuzniar et al., 2008). However, in the most general case, all such methods suffer from a common limitation, relevant to the difficulties in constructing orthologous classes in the presence of paralogs. More specifically, for distantly related species, assessing orthology can become very difficult typically because of low similarity between protein sequences and the likely increase of gene-birth and death (Lynch and Conery, 2003).

In this general background, we develop here a methodological framework for the detection of orthologs, which relies on the recognized advantages of existing schemes, remedying in the same time to some of their known shortcomings. We assess the effectiveness of the developed scheme, and its possible advantages, through conceptual and practical comparisons with other similarity based methods.

There are essentially two approaches to determine orthologous gene classes, based either on phylogeny (tree-based methods) or sequence similarity (mainly with Blast searches). Phylogenetic approaches to orthology rely on the reconciliation of the phylogeny of orthologous genes with their corresponding species phylogeny (Altenhoff and Dessimoz, 2009; Dehal and Boore, 2006; Gabaldón, 2008; Goodstadt and Ponting, 2006; Zmasek and Eddy, 2004). Blast based approaches to orthology rely on genome comparisons and classification of highly similar genes (Chen et al., 2007; Enright et al., 2003; Kriventseva et al., 2008; Li et al., 2003; Linard et al., 2011; Remm et al., 2001; Tatusov et al., 2003). Phylogenetic methods deemed to be most reliable (Gabaldón, 2008), are difficult to automate and the complexity of their use grows with the number of taxa (Poptsova and Gogarten, 2007), particularly in large families where orthology is most difficult to assess. Such approaches are also subject to controversy on conceptual grounds, as species phylogeny is not always straightforwardly established (Forest, 2009; Koonin, 2005; Tekaia and Yeramian, 2005), with gene and species phylogenies not necessarily coinciding.

On the other hand Blast approaches, relying on Reciprocal Best Hits (RBH), are recognized to perform well in terms of comparative accuracy. It was shown (Moreno-Hagelsieb and Latimer, 2008) that Blast searches with “soft filtering” and “Smith–Waterman alignment” options produce the highest number of orthologs with the minimal error rates, as compared to other methods. Furthermore recent benchmark studies (Altenhoff and Dessimoz, 2009; Salichos and Rokas, 2011) concluded that the RBH method outperforms other approaches involving more complex algorithms. However, the RBH approach was criticized for under appreciation of orthology in the presence of paralogs. Most importantly, this simple approach was criticized to suffer from conceptual drawbacks: 1) RBH analyses are restricted to the class of 1:1 orthologs, failing thus in the detection of many-to-(one and/or many) orthologs and 2) the classification of orthologs detected with RBH methods is suspected to be prone to chaining effects in the motif and domain compositions. Chaining effects can be associated with gene fusion and domain shuffling. There is then the possibility that various domains in multi-domain proteins have evolved through different speciation and duplication events (Jensen et al., 2008), with the corresponding artifactual increase in the number of detected similar sequences.

In addition, RBH detections are suspected (Catchen et al., 2009; Salichos and Rokas, 2011) to become over-inclusive when gene losses are involved in some of the species. In such cases a gene could be

erroneously considered as best hit only because its genuine counterpart is lost.

Taking into account the various facets of the methodological situation earlier we introduce here a working framework, to identify potential orthologs and classify them. In this context we take advantage of the efficiency and simplicity of the RBH-based detection of orthologs, remedying in the same time – by design – the drawback concerning the limitations in terms of 1:1 detections. We resort to a partitioning of Reciprocal Best Hits with the subsequent merging of partitions containing members of the same paralogous classes into a higher-order cluster, we call “SuperPartitions of Orthologs” (SPOs). A major advantage of this clustering method is to avoid the discrepancy of associated paralogs being dispatched into different clusters of orthologs (a problem frequently pointed out; see for example Fig. 1 in Kuzniar et al., 2008).

The effectiveness of the conceptual design is further demonstrated in practice with detailed analyses, involving notably comparisons with other methods.

More precisely, for the validation of the SPO approach we proceed following two complementary directions: assessing internal consistency with global characterizations and assessing detailed performances through comparisons with other methods.

For the global characterizations we consider three sets of related species spanning significantly different levels of genome and organism complexity: *Chlamydiae* (13 genomes), *Mycobacteria* (13 genomes) and *Aspergilli* (8 genomes). With the merging scheme developed, it appears that, in predominant cases, all the proteins in a given SPO share at least one ancestral motif. This result hints towards the ancestral organization of such proteins, prior to speciation events. It thus appears that the SPO-based approach avoids to a large extent the problem of chaining effects, which might be associated with multi-domain proteins or with gene-fusions.

For methods comparisons we consider MultiParanoid (Alexeyenko et al., 2006) and OrthoMCL (Li et al., 2003) and three publicly available datasets related to *H. sapiens*, *D. melanogaster*, *C. elegans* and *C. intestinalis*, with associated analyses. The most prominent result from these comparisons concerns the different outcomes, of the merging schemes underlying the various methods. In most cases, it appears that the SPO method brings together a more or less large number of sub-clusters that can be associated with distinct orthologous classes in the other approaches. Different criteria (such as motif analyses, functional annotations or obvious relatedness of proteins) suggest then that the higher-order clustering with SPOs appears to avoid spurious dispatching of paralogs into distinct classes.

Further, for fine-grained evolutionary analyses, we construct conservation profiles associated with SPOs. With such profiles, it is straightforward to discriminate between clusters containing at most one gene per species from clusters containing at least one duplicated gene. It becomes then possible to sort out in-paralogs and out-paralogs in such clusters. For the sets of SPOs with more than one protein per species, we rely on an operational definition put forward by (Koonin, 2005) to differentiate in-paralogs from possible out-paralogs. Following this operational definition, genes recently duplicated are less constrained functionally than their corresponding orthologs, sharing similar original functions. Accordingly such duplicated genes can more easily undergo divergence, with new functional adaptation. Although sometimes questioned (Studer and Robinson-Rechavi, 2009), this assumption is generally accepted and we used it to detect in-paralogs (genes duplicated after speciation) by comparing, within a given SPO, the similarities between proteins from the same species with their respective similarities with proteins from other species. Because of their recent duplication, in-paralogous protein-pairs are expected to be more similar to each other than to any other protein from other species.

Finally we discuss the perspectives for the SPO-based framework developed here, highlighting the specific potentialities of the approach for various types of evolutionary studies.

2. Materials and methods

The methodological steps in this work are schematically diagrammed in Fig. 1, under the form of a flowchart. For the detection of orthologs, large-scale intra-species and inter-species comparisons were performed for a set of predicted proteomes. Based on such comparisons, the subset of Reciprocal Best Hits was determined and further classified into clusters that are called SuperPartition of Orthologs. In what follows we detail each step in the flowchart in Fig. 1, with the construction of clusters of orthologs and the corresponding conservation profiles.

2.1. Species-specific comparisons

Protein sequence comparisons were performed following previously described methodology for large-scale proteome comparisons (Tekai et al., 2000). Blastp was used with the substitution matrix *blosum62*, the “soft filtering” and “Smith–Waterman alignment” options ($-F$ “*m S*” $-s$ *T*). With the soft filtering option, low information segments are masked only during the search phase. This option produces also better alignment scorings. With the Smith–Waterman option, alignments are performed based on the dynamic programming scheme, leading to mathematically optimal local alignments (Altschul et al., 1997; Moreno-Hagelsieb and Latimer, 2008; Schäffer et al., 2001). Genome species-specific comparisons were performed for each pair of species (i.e. four whole proteome runs per species pair: 2 runs both ways and 2 self-self runs). Blastp significant similarity between sequences was set to $e\text{-value} \leq 1.0 \times 10^{-9}$ for eukaryotic species and $e\text{-value} \leq 1.0 \times 10^{-5}$ for bacterial species (Tekai and

Dujon, 1999; Tekai and Yeramian, 2005). For intra-species comparisons, in order to avoid weak links (one way similarity only), only proteins with Reciprocal Significant Hits (RSH) were taken into account for classification. For each pair of distinct species, only significant Reciprocal Best Hits (RBH) were taken into account for classification.

2.2. Intra-species comparisons: paralogs

Classification of non-unique proteins in intra-species comparisons was performed following three steps (Fig. 1):

- 1) Reciprocal Significant Hits (RSHs), were determined in each species and partitioned into subsets of proteins. Partitions are disjoint subsets, each subset contains proteins significantly similar to at least one other protein in the subset and displays no similarity with proteins not contained in the subset. The subsets are minimal structures, i.e. they cannot be further divided. Each non-unique protein is assigned to a partition denoted $P_{n,m}$, with n the number of proteins in the partition and m an arbitrary order for indexing.
- 2) The set of aforementioned RSHs was subdivided into classes using an alternative method, the *mcl* program (Enright et al., 2002, 2003; <http://micans.org/mcl/>), using “ $-\log(\text{blastp}(e\text{-value}))$ ” and an inflation index $I=3.0$. Each non-unique protein was assigned to a cluster denoted $C_{p,q}$, with p the number of proteins it contains and q an arbitrary order for indexing.
- 3) Finally, the basic merging scheme relies on the *mcl* clustering, leading to the $C_{p,q}$ classes. In order to keep track of the information relative to $P_{n,m}$ partitioning as well, the *mcl* ($C_{p,q}$) cluster

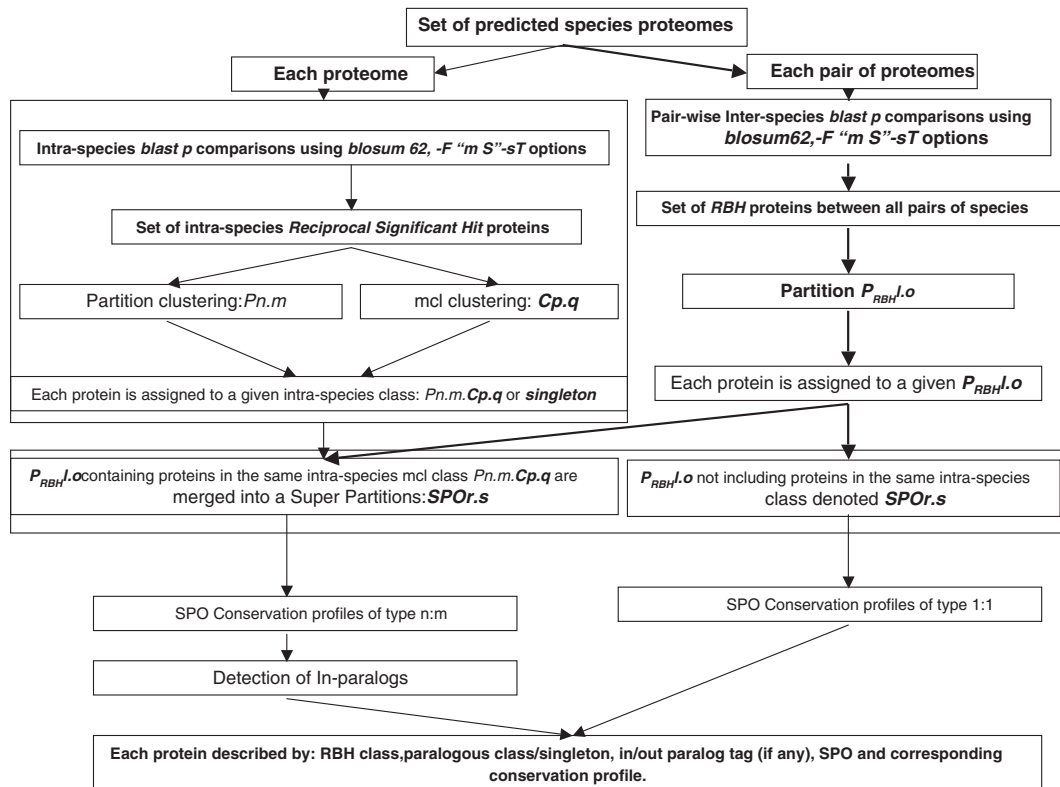


Fig. 1. Flow chart for the construction of SuperPartitions of Orthologs (SPOs). The figure shows the different steps in the construction of SuperPartitions of Orthologs, with the detection of possible in-paralogs (see details in Materials and methods). Given a set of species, with their predicted proteomes, the procedure performs intra-species and inter-species *blastp* comparisons (with the indicated options), detecting the intra-species unique (or singleton) proteins, the sets of Reciprocal Significant Hits (RSH) and the sets of inter-species Reciprocal Best Hits (RBH). Partition and *mcl* (Enright et al., 2002) clustering of RSH sets and partition of RBH sets are performed next (see Materials and methods for details). Partitions of RBHs (P_{RBH}) containing members of the same intra-species cluster are merged into a SuperPartition of Orthologs denoted $SPO_{r,s}$ (with r the number of members and s an arbitrary order for indexing). Partitions of RBHs containing at most one protein in the considered species are also denoted $SPO_{r,s}$. The SPO conservation profiles are then derived (see Materials and methods). These profiles are classified as type 1:1 if they contain at most one protein from a given species and as type $n:m$ if they contain at least one duplicated protein from a given species. SPO conservation profiles of type $n:m$ are used to detect in-paralogs and/or out-paralogs (see Materials and Methods).

is prefixed with its associated partition Pn.m. Each protein in a Cp.q cluster was then further labeled as Pn.m.Cp.q. Proteins with no significant matches were assigned to “singletons” classes (Tekaia and Latge, 2005).

The mcl cluster is the one that will be used in the construction of SPOs in next steps, using the annotation scheme indicated earlier. The indication Pn.m provides additional information as for the size of the corresponding partition, in which the mcl cluster might be contained.

The various concepts and notations mentioned are illustrated with a detailed example from *M. tuberculosis* (see additional material Figure S1).

2.3. Inter-species comparisons: RBHs

RBHs were determined following three steps (Fig. 1):

- 1) Pairs of proteins showing Reciprocal Best Hits (RBHs) were detected through the pair-wise comparisons of proteomes.
- 2) RBHs obtained from all pair-wise proteome comparisons were partitioned into subsets of proteins denoted $P_{RBH,l,o}$ with l the subset size (number of proteins) and o an arbitrary order for indexing.
- 3) Each protein was assigned to the appropriate partition $P_{RBH,l,o}$.

2.4. Construction of SPOs and corresponding Conservation Profiles

RBHs partitions ($P_{RBH,l,o}$) were further processed, with the merging of partitions including proteins belonging to the same intra-species class Cp.q (labeled Pn.m.Cp.q), into a SuperPartition (see examples in Fig. 2) denoted $SPO_{r,s}$ with r the number of proteins in the SPO and s an arbitrary order for indexing. The $SPO_{r,s}$ denotation was also adopted for RBH partitions not containing multiple members of the same intra-species class.

In addition, each SPO was characterized by the associated conservation profile along the considered species and their known phylogeny. SPOs including at most one protein from a given species were designated as type 1:1 and those including more than one protein from at least one species were designated as type n:m (Fig. 3).

2.5. Detection of in-paralogs

SPOs of type n:m were analyzed to detect in-paralogs and out-paralogs following the criteria put forward in (Koonin, 2005). In such SPOs, pairs of proteins belonging to the same species were compared to all proteins from other species in the same SPO. Such protein pairs were classified as in-paralogs, whenever their intra-species similarities were more significant than the similarity with proteins in the same SPO from other species (based on associated blastp e-values). Otherwise the protein pair was classified as out-paralogs.

2.6. MEME/MAST motifs in SPOs

Motifs were searched using meme/mast programs (Bailey and Elkan, 1994). For a set of species, motif searches were performed for all SPOs, taking into account their specific amino-acid compositions, with possible significant differences relative to the default compositions (Tekaia and Yeramian, 2006; Tekaia et al., 2002). The distributions of motifs were analyzed for proteins in a given SPO, detecting motifs shared by all proteins in the SPO, or by subsets of them.

2.7. Overall protein characterizations

As a result of the various steps earlier, each protein, involved in at least one RBH pair, was characterized by the following descriptors: a) RBH partition ($P_{RBH,l,o}$); b) Intra-species mcl cluster Cp.q (with the attached labeling Pn.m.Cp.q); c) Superpartition SPO ($SPO_{r,s}$) and d)

Conservation profile. In addition, a given protein could be identified as in-paralog or out-paralog.

2.8. Comparisons with other methods

Comparisons were performed with MultiParanoid (designated as MP; Alexeyenko et al., 2006) and OrthoMCL (designated as OM; Li et al., 2003), based on publicly available data and pre-computed clustering results. The corresponding datasets were downloaded (30/08/2011) from <http://multiparanoid.sbc.su.se/download/> for MultiParanoid and from http://info.gersteinlab.org/Ortholog_Resources for OrthoMCL (corresponding to the results obtained by Fang et al., 2010).

2.8.1. MultiParanoid (MP) dataset

The dataset (ensCIOIN.fa-ensHOMSA.fa-modCAEEL.fa-modDROME.fa.MP.sql) from the MultiParanoid server corresponds to comparisons between four species: *H. sapiens* (ensHOMSA.fa: 22983 proteins), *C. elegans* (modCAEEL.fa: 20084), *D. melanogaster* (modDROME.fa: 13854) and *Ciona intestinalis* (ensCIOIN.fas: 14278), representing a total of 71199 proteins, that have been clustered into 7453 clusters.

The original MultiParanoid cluster labeling was adapted so as to follow the $SPO_{r,s}$ labeling: $CMPr_s$ with r the number of proteins in the cluster and s an arbitrary order for indexing.

2.8.2. OrthoMCL (OM) dataset

The dataset (OrthoMCL_raw_human_fly_worm_groups) downloaded from the Ortholog Resources server, corresponds to comparisons between three species (Fang et al., 2010): *H. sapiens*, *C. elegans* and *D. melanogaster*. For these three species, the OrthoMCL analyses were based on the same data as those used in MultiParanoid, representing a total of 56,921 proteins, and that have been clustered into 6467 clusters.

As for MultiParanoid, the original OrthoMCL cluster labeling was adapted: $COMr_s$ with r the number of proteins in the cluster and s an arbitrary order for indexing.

2.8.3. Common dataset

For the overall comparisons, we considered the set of proteins that have been detected by at least one of the three methods, resulting in 30693 proteins from Human, fly and *C. elegans* (common species in the aforementioned datasets) with the associated number of clusters (7453 for MP, 6467 for OM and the 3766 determined SPOs). The associated clusters in MultiParanoid were resized after removing *C. intestinalis* proteins if any.

3. SuperPartitions: elaborations and results

We consider the construction of clusters of orthologs for a given set of species. The methodological elaborations are highlighted with regard to the conceptual difficulties encountered. Assessing internal consistency, the constructions are detailed with results obtained for three sets of species.

3.1. SuperPartitions of Orthologs

Fig. 2 illustrates schematically the difficulties encountered with the handling of RBH sets in the presence of paralogs. In such situations RBH partitioning can lead to incomplete classification of orthologous proteins, with connections between pairs of proteins broken into different RBH sets (spurious dispatching of paralogs into distinct classes). The construction of SuperPartitions in what follows is precisely to remedy this situation.

3.1.1. Design and construction of SuperPartitions

Given a set of species (see flow chart in Fig. 1), our procedure starts with intra-species and inter-species specific comparisons to identify: 1) unique (or singleton) and non-unique proteins in each

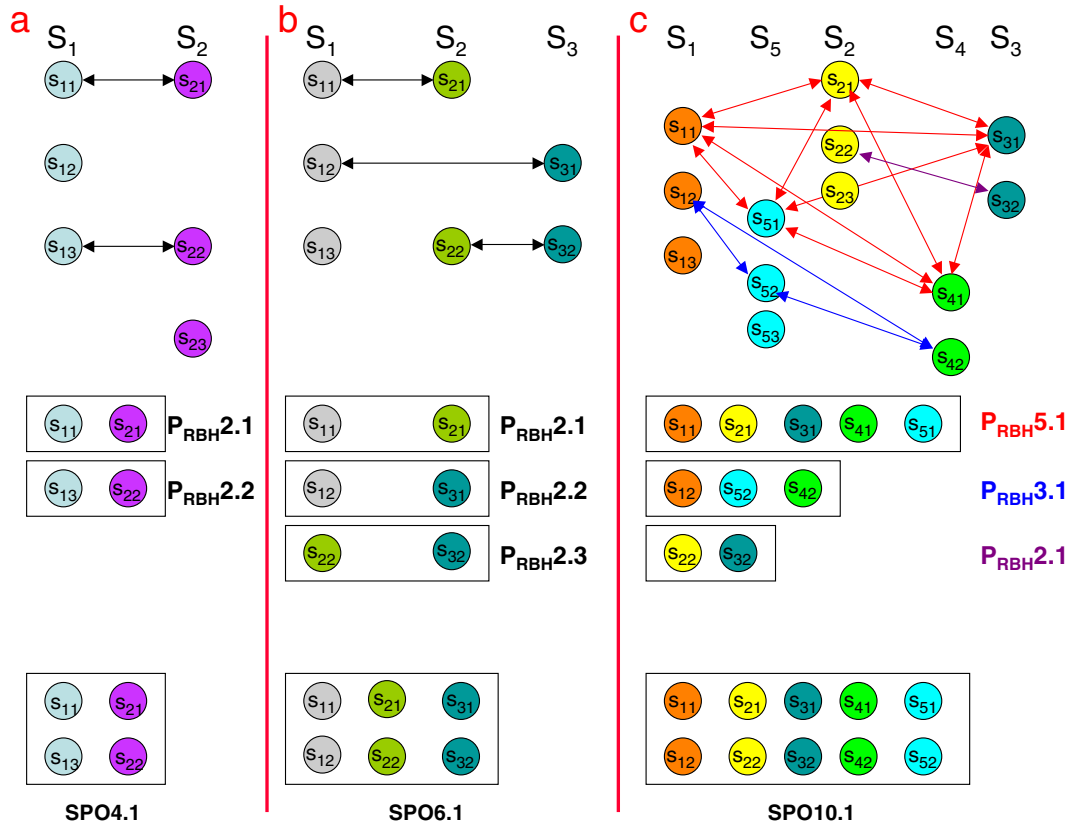


Fig. 2. RBH partitions in presence of paralogs and construction of SuperPartitions (SPOs). The construction of SuperPartitions of Orthologs (SPOs) is illustrated with toy examples ((a) to (c) in increasing complexity with respectively 2, 3 and 5 species). RBH relationships are indicated by arrows between proteins from various species (denoted S₁ to S₅). In each species, proteins belonging to a given class of paralogs are represented with the same colour (mcl cluster **Cp,q** with p members, also annotated as Pn.m.Cp,q with the labeling Pn.m associated with RBH partitioning; see text). (a) RBH detections lead to 2 partitions P_{RBH}2.1 and P_{RBH}2.2. The two partitions containing paralogous proteins (S₁₁ and S₁₃ from species S₁; S₂₁ and S₂₂ from species S₂), are merged into the SuperPartition SPO4.1. The proteins S₁₂ (paralogous to S₁₁ and S₁₃) and S₂₃ (paralogous to S₂₁ and S₂₂) are not included in SPO4.1 as they are not involved in RBH links. (b) For the 3 species (S₁, S₂, S₃), the SuperPartition SPO6.1 results from the merging of RBH partitions P_{RBH}2.1, P_{RBH}2.2 and P_{RBH}2.3 (S₁₁ and S₁₂ are paralogs in species S₁; S₂₁ and S₂₂ are paralogs in species S₂ and S₃₁ and S₃₂ are paralogs in species S₃). S₁₃ is not included in SPO6.1, as it is not involved in any RBH link. (c) For the 5 species (S₁ to S₅), the SuperPartition SPO10.1 results from the merging of RBH partitions P_{RBH}5.1, P_{RBH}3.1 and P_{RBH}2.1. The SPO contains only proteins involved in RBH pairs.

a:

SPO8.1	ASFL-ASOR-ASTE-ASNG-NEFI-ASFU-ASCL-ASNI
SPO8.1	1 1 1 1 1 1 1 1

b:

SPO4.2ASOR-.....ASNG-NEFI-ASFU-.....
SPO4.2	0 1 0 1 1 1 0 0

c:

SPO15.3	ASFL-ASOR-ASTE-.....NEFI-ASFU-ASCL-ASNI
SPO15.3	1 1 1 0 4 3 3 2

Fig. 3. SPO Conservation Profiles. SPO Conservation Profiles are illustrated for a set of 8 *Aspergilli* species (see Table 1 for species coding). An SPO conservation profile is represented as a vector of dimension n (n the number of species, 8 in this example), with at position i (associated with species i) information relative to the presence or absence of orthologs in the corresponding species. In the various examples (a to c), the representations of the conservation profiles are either in terms of simple presence/absence information (first line in each example; dot symbols for absence) or in terms of more detailed information concerning the number of proteins in each species (second line in each example; number p, with possible value 0, corresponding to the number of proteins in the species associated with the SPO). With the detailed representation, the number p at each position corresponds to the number of proteins with RBH hits. (a) Conservation profile associated with SPO8.1 (type 1:1): one protein in each of the 8 species. (b) Conservation profile associated with SPO4.2 (type 1:1): no representatives for species ASFL, ASTE, ASCL and ASNI; and 1 representative for ASOR, ASNG, NEFI and ASFU (see Table 1 for species coding). (c) Conservation profile associated with SPO15.3 (type n:m): no representative for ASNG, 4 for NEFI, 3 for ASFU, 3 for ASCL and 2 for ASNI.

species along with the intra-species classes to which they belong (classification is based on partition and mcl clustering methods of proteins showing Reciprocal Significant Hits), and 2) pairs of Reciprocal Best Hits of proteins from distinct pairs of species (see Fig. 1 and Materials and methods for details).

All such pair-wise RBH connections are then partitioned into subsets designated by P_{RBH}n,m, with n the number of proteins in the subset and m an arbitrary order for indexing. As illustrated schematically in Fig. 2, each member of such a subset is connected to at least one other member of the subset with RBH links, leading to disjoint P_{RBH}n,m partitions. Such partitioning is represented schematically in Fig. 2, with increasingly more complex situations (a, b and c, with respectively 2, 3 and 5 species), illustrating the possible discrepancies in the classification of RBH partitions. For example the RBH partitioning in Fig. 2c, leads to the two distinct partitions denoted as P_{RBH}3.1 and P_{RBH}2.1, which can be erroneously considered as independent partitions of orthologs.

In a second step, minimizing as much as possible the number of broken RBH sets, RBH partitions which contain members of common paralogous classes (defined by the intra-species mcl clusters Cp,q labeled as Pn.m.Cp,q) are merged into a single set, we call SuperPartitions of Orthologs (SPOs). In such construction the merging concerns only paralogs that are members of RBH sets. For further specification, we adopt in what follows the notation SPOr,s for SuperPartitions, with r the number of proteins in the merged set and s an arbitrary order for indexing. Such constructions are illustrated in detail for the various examples in Fig. 2.

In the merging scheme earlier, based on the mcl classes **Cp,q**, each protein is further labeled as Pn.m.Cp,q, to keep track of the additional accessory information relative to the Pn.m partitions as well (see example on Figure S1 in additional material). More precisely, such dual labeling of proteins in a given mcl class is well tailored for handling and classifying protein sets with different levels of mutual relatedness, following their respective similarity scores. The Pn.m.Cp,q construction can then be useful to highlight cases where different mcl clusters of paralogs are found in the same Pn.m partition, and hence display possible weak relations to other proteins in that partition (For details and illustrations see additional material Figure S1).

The construction of SPOs relies solely on RBH predictions between distinct pairs of species. It is then important to stress in this context the rationale in orthology predictions from RBHs, and the possible pitfalls and difficulties associated with such predictions. The construction relies on a fundamental assumption in orthology predictions from RBHs, following which two orthologous genes resulting from a speciation event are expected to be more similar to each other than to any paralogous genes in the other species (Koonin, 2005). Indeed, duplicated genes are assumed to evolve relatively rapidly, with possible adaptive properties to the species environment. For RBH ortholog predictions, potential pitfalls were discussed in various reviews (see for example Fig. 1 in (Poptsova and Gogarten, 2007) and Fig. 2 in (Gabaldón, 2008)). One such reported pitfall concerns the case of two paralogs in a given species, with RBH predictions based on comparisons with another species detecting only one ortholog. Such a pitfall could be resolved in the SPO scheme by resorting to many pair-wise species comparisons. With multi-species comparisons, it is expected that the “missed” ortholog will establish at least one RBH link, and will be accordingly recovered in the proper SPO. Such a situation is illustrated schematically in Fig. 2b, which highlights the importance for RBH analyses to involve at least 3 species. In this example, the final SPO (SPO6.1, Fig. 2b), resulting from the comparisons between 3 species (S_1 to S_3), includes proteins from the two first species S_1 and S_2 (S_{12} and S_{22}), which would have been missed without the inclusion of the third species (S_3) in the analyses.

Another important potential difficulty with RBH-based analyses concerns the handling of orthologs in the presence of paralogs. To address this difficulty, we consider next the construction of SPO conservation profiles.

3.1.2. SPO conservation profiles and detection of in-paralogs

As illustrated schematically in Fig. 2, based on the configurations of the various RBH links, an SPO may contain paralogs (proteins that are members of intra-species clusters). Of course such situations occur only for SPOs containing more than one protein from the same species. The analysis of such cases can be non-trivial, notably because of the difficulty in differentiating between “in-paralogs” (duplicated in the species after speciation) and “out-paralogs” (duplicated before speciation). The construction of SPOs offers then a convenient possibility to separate the relatively difficult situations, involving families of paralogs, from the simple situations involving only singletons, associated with straightforward interpretations. With this respect it will be sufficient to separate the set of SPOs that contain at most one protein from a given species, from the SPOs containing duplicated proteins in at least one species. It is convenient to analyse the two types of situations concerning SPOs, based on conservation profiles. For a given SPO, such a profile is a vector of dimension n (number of species), specifying at position i (following species) the number p of protein coding genes from the species (with possibly the value 0) in the SPO (see examples in Fig. 3). Of course in such representation, p corresponds to the number of different proteins found by RBH predictions in a given species (which means that p does not correspond necessarily to the number of paralogs for the given protein; as illustrated in Fig. 2). A simplified version of the profiles can be in terms of presence/absence, with no specification of the precise number of

proteins (Figs. 3a,b). With the SPO conservation profiles we can straightforwardly distinguish between SPOs of type 1:1 (containing at most one protein for each represented species) and SPOs of type $p:q$ (containing at least two proteins from one of the represented species). The interest in such a separation is to isolate the set of SPOs of type $p:q$, which are more difficult to interpret in terms of orthology. Indeed for such SPOs we need to distinguish between proteins resulting from duplication events within given species from those resulting from duplications which took place before speciation events. To handle SPOs in this category, it is possible then to resort to more specific treatments, such as “manual” annotations or taking into account various types of information, such as relevant to phylogeny analyses or to conservation of syntenic blocks. Here, we detect in-paralogs (i.e. recently duplicated genes), relying on the simple (and generally accepted) assumption following which in-paralog protein pairs are expected to be more similar to each other than to any other protein from other species (Koonin, 2005). Corresponding to the complementary case, the detection of out-paralogs is then straightforward in such a scheme. It is also notable that such analyses take into account only the last speciation event: in-paralogy and out-paralogy are defined relatively to this event.

3.2. SPO analyses: global characterizations

We consider first global features emerging from SPO analyses, based on three sets of species whose genomes are available (Table 1): two sets of bacterial species (13 *Chlamydiae*, 13 *Mycobacteria*) and one set of eukaryotic species (8 *Aspergilli*). We report the distribution of the number of SPOs obtained and their description in terms of shared motifs.

3.2.1. Distributions of SPOs

Table 2 shows the distribution of the number of SPOs in the three sets of species. An important feature concerns the percentage of SPOs constituted of 1:1 orthologs, as compared to $n:m$ orthologs. For the three sets of species more than 75% of the SPOs are of type 1:1. Accordingly it can be assumed that with SPO constructions, analyses to distinguish orthologs from possible in-paralogs would be required in less than 25% of the cases. For such cases, it appears that a variable proportions of SPOs of type $n:m$ (28% to 58%) contain in-paralogs (Table 2).

Fig. 4 shows the distribution of SPO occurrences, following their sizes (number of proteins in given classes). It can be observed that significant SPO occurrences lie within the range of the total number of corresponding species. Thus most significant occurrences lie between sizes 2 and 13 for the *Chlamydiae* and *Mycobacteria* sets, and between sizes 2 and 8 for the *Aspergilli* set. The distributions reveal the presence of a number of large-sized SPOs, particularly in *Mycobacteria*. The large sizes for such SPOs might be due to the chaining effects, particularly in the case of PPE and PE-PGRS families (Cole et al., 1998; Tekaia et al., 1999; Gey van Pittius et al., 2006). Of course using such SPOs for the determination of orthologs would be of little relevance, because of the large intra-species families of proteins they contain, displaying short motifs in common (PE, PE-PGRS for example).

3.2.2. SPO motif compositions and evolution of proteins

In the analyses discussed earlier proteins are simply characterized by the overall SPO classifications. Such characterizations can be further refined, with the details of the SPO motif compositions (see Materials and methods). With such details, it becomes possible to differentiate between proteins in a given SPO, following evolutionary characteristics as revealed by motifs. We illustrate such rationale with a specific example, paving the way for further extensive similar analyses. Fig. 5 shows the motif composition of SPO26.27 (*Mycobacteria*), associated with protein sets mtc28 ($P_{RBH}13.134$) and lpqT ($P_{RBH}13.250$). The distribution reveals core-anchoring motifs (2, 4 and 1) shared by all proteins. lpqT proteins display a compact and highly conserved organization in

Table 1

List of the three sets of considered species.

The table shows the list of the three sets of considered species: a) 13 *Chlamydiae*, b) 13 *Mycobacteria* (corresponding proteomes have been downloaded from the ncbi ftp server: ftp.nlm.ncbi.nih.gov/genomes/Bacteria/species_genomes).

c) 8 *Aspergilli* (corresponding proteomes have been downloaded from the Broad Institute server: http://www.broad.mit.edu/annotation/genome/aspergillus_group/Multi-Home.html). Columns correspond respectively to the species code, size of the corresponding proteome and the organism identification.

Code	Size	Organism
<i>a)</i>		
CHMU	911	<i>Chlamydia muridarum</i>
CHTH	919	<i>Chlamydia trachomatis A HAR-13</i>
CHTR	895	<i>Chlamydia trachomatis</i>
CHTL	874	<i>Chlamydia trachomatis (LGV)</i>
CHTB	874	<i>Chlamydia trachomatis 434 Bu</i>
CHAB	932	<i>Chlamydomphila abortus S26 3</i>
CHFE	1013	<i>Chlamydomphila felis Fe C-56</i>
CHCA	1005	<i>Chlamydomphila caviae</i>
CHPN	1112	<i>Chlamydomphila pneumoniae AR39</i>
CHPC	1052	<i>Chlamydomphila pneumoniae CWL0299</i>
CHPT	1113	<i>Chlamydomphila pneumoniae TW 183</i>
CHPJ	1069	<i>Chlamydomphila pneumoniae J138</i>
PAUW	2031	<i>Parachlamydia sp UWE25</i>
<i>b)</i>		
MTTU	3996	<i>Mycobacterium tuberculosis H37R</i>
MYBO	3920	<i>Mycobacterium bovis</i>
MYTC	4203	<i>Mycobacterium tuberculosis CDC 1551</i>
MYUL	5105	<i>Mycobacterium ulcerans</i>
MYMA	5483	<i>Mycobacterium marinum</i>
MYLE	1614	<i>Mycobacterium leprae</i>
MYAV	5120	<i>Mycobacterium avium</i>
MYAP	4350	<i>Mycobacterium avium paratuberculosis</i>
MYJL	5739	<i>Mycobacterium JLS</i>
MYVA	5979	<i>Mycobacterium vanbaalenii PYR-1</i>
MYGI	5579	<i>Mycobacterium gilvum PYR GCK</i>
MYSM	6716	<i>Mycobacterium smegmatis MC2 155</i>
MYAB	4941	<i>Mycobacterium abscessus ATCC 19977T</i>
<i>c)</i>		
Code	Size	Organism
ASFL	12,587	<i>Aspergillus flavus</i>
ASOR	12,063	<i>Aspergillus oryzae</i>
ASTE	10,406	<i>Aspergillus terreus</i>
ASNG	8592	<i>Aspergillus niger</i>
NEFI	10,407	<i>Neosartorya fischeri</i>
ASFU	9630	<i>Aspergillus fumigatus</i>
ASCL	9124	<i>Aspergillus clavatus</i>
ASNI	10,701	<i>Aspergillus nidulans</i>

terms of motifs, in contrast to mtc28 proteins with significant variability at both ends. Following the most parsimonious evolutionary hypothesis it can be accordingly assumed that the ancestral gene for this SPO was duplicated before speciation, implying in turn that the duplicated proteins are out-paralogs. This analysis is coherent with results of in-paralogs detection (as described earlier), applied to this specific SPO.

3.2.3. Weights of ancestral motifs shared by SPO members

One complication that might undermine SPO construction is the effect of multi-domain proteins and gene fusions that may artificially connect different proteins in the same SPO. To characterize in more detail evolutionary information captured by SPOs, we consider the motif compositions of corresponding protein members, notably to determine the weight of ancestry in shared motifs. The motif compositions were determined using meme/mast programs (see [Materials and methods](#)), and their distributions were analyzed for each SPO. [Fig. 6](#) shows the distributions of the occurrences of SPOs following the number of motifs (0, 1, 2,... up to 15) shared by all proteins in the corresponding SPOs. The distributions for the three sets of species appear to be very similar, with only a small number of SPOs without any shared motif between

Table 2

Distribution of SPO conservation profiles and detection of in-paralogs.

The table shows for the three sets of species (13 *Chlamydiae*, 13 *Mycobacteria*, 8 *Aspergilli*): a) total number of detected partitions of RBHs, b) total number of SuperPartitions of Orthologs (SPOs), c) corresponding total number of distinct conservation profiles, d) and e) proportions of 1:1 and n:m conservation profiles (respectively: at most one protein per represented species and more than one protein in at least one represented species), f) proportion of SPOs of type n:m containing in-paralogs (see text), g) proportions of SPOs containing representatives from all species in the corresponding sets and h) proportions of SPOs containing exactly one protein from each species.

Sets of species	13 <i>Chlamydiae</i>	13 <i>Mycobacteria</i>	8 <i>Aspergilli</i>
a) Partitions of RBHs	1202	7560	11,887
b) SPOs	948	3708	6192
c) Distinct Conservation Profiles	73	414	213
d) 1:1 Conservation Profiles	823 (86.8%)	2780 (75.0%)	4837 (78.1%)
e) n:m Conservation Profiles	125 (13.2%)	928 (25.0%)	1355 (21.9%)
f) SPOs of type n:m containing in-paralogs	35 (28.0%)	535 (57.7%)	505 (37.3%)
g) SPOs containing proteins from each species	522 (55.1%)	869 (23.4%)	2474 (40.0%)
h) SPOs containing exactly one protein from each species	439 (46.3%)	535 (14.4%)	1597 (25.8%)

all protein members (n0: 4.1% in *Chlamydiae*, 6.4% in *Mycobacteria* and 7.6% in *Aspergilli*). For the three sets, the large majority of SPOs (more than 92%) displays at least one ancestral motif shared by all members.

In evolutionary terms this result confirms that members of an SPO are in most cases of common ancestral origin. On methodological grounds, following this result, it appears that the construction of SPOs should not be affected in any significant way by chaining effects involving multi-domain proteins or gene fusion.

4. Comparisons of methods

For validation purposes, beyond the analyses earlier concerning internal consistency, we consider comparisons between the SPO approach here, and other methods. We restricted the comparisons to the MultiParanoid (MP) and OrthoMCL (OM) methods (also completed by comparisons with eggNOG (Jensen et al., 2008 and <http://eggno.embl.de/>) for specific examples), for which datasets and associated analyses are available (concerning *H. sapiens*, *D. melanogaster*, *C. elegans* and *C. intestinalis* species).

We consider first global overview, characterizing quantitatively overlaps and discrepancies, between methods taken two by two, and all three methods together. Then in order to try to gain insight in this landscape, we consider several detailed examples to highlight relevant features.

Detailed comparison results for detected orthologs, along with corresponding classifications, are provided as additional files (see [Supplementary data tables in additional material](#)).

4.1. Global overview

Summarizing the global landscape, the comparison results are displayed in [Fig. 7](#), under the form of Venn diagrams: [Figs. 7a](#) and [b](#) concern the two-by-two comparisons (respectively between MP and SPO, and between OM and SPO), whereas the full comparison results between the three methods are displayed in [Fig. 7c](#). More specifically, the comparison between MP and SPO concerns a total of 71199 proteins, associated with the 4 species for which MP results are available (see [Materials and methods](#)). Among the 35,897 proteins detected as orthologs by at least one of the two methods, 24,070 proteins are

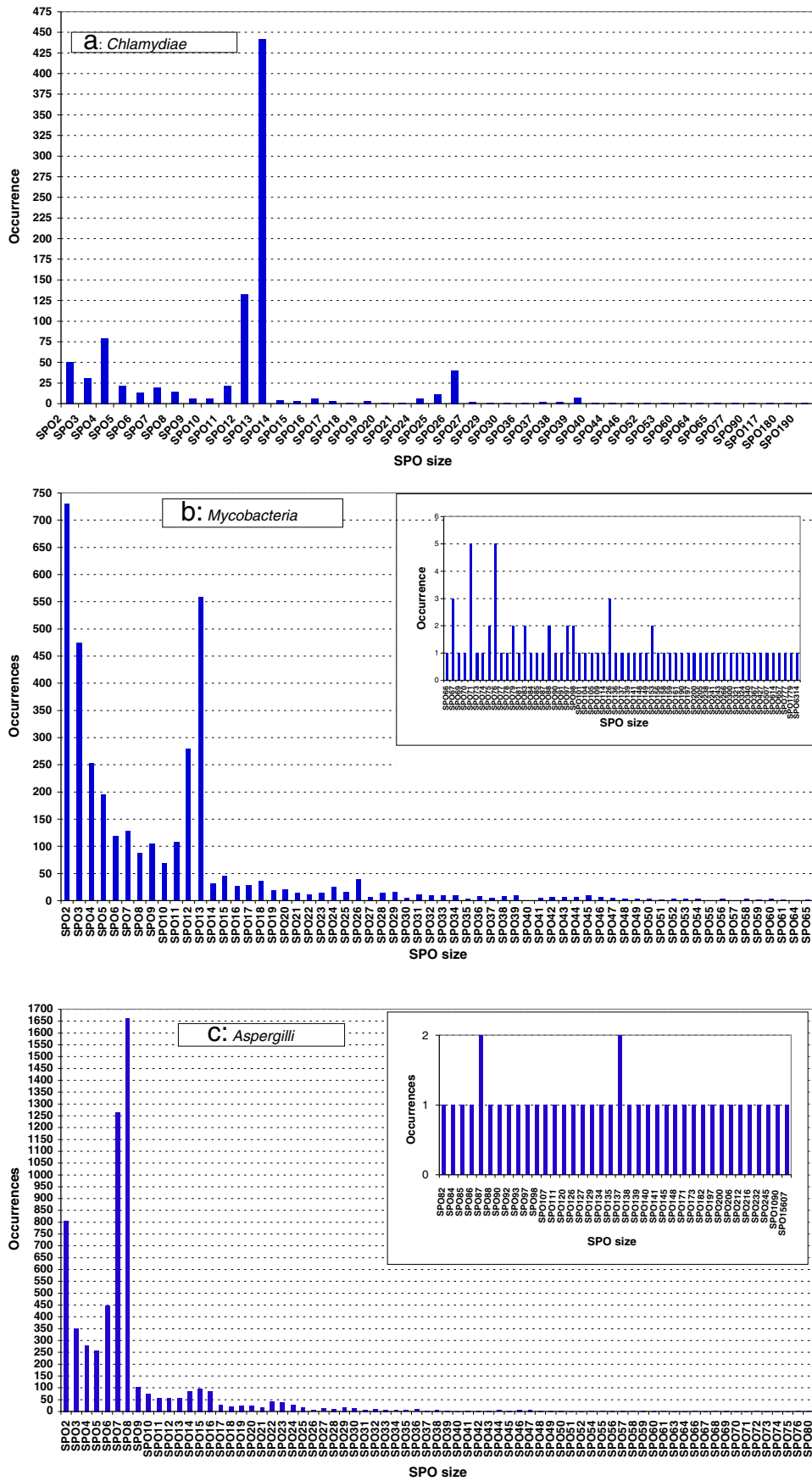


Fig. 4. Distributions of occurrences of SPOs. Occurrences of SPOs are represented following their size (number of proteins) for 3 sets of species : a) 13 *Chlamydiae*, b) 13 *Mycobacteria* and c) 8 *Aspergilli* species (see Table 1 for species coding). SPOs are reported following the x-axis as SPO_n (with n the relevant size). (a) *Chlamydiae* (948 SPOs): sizes between 2 and 190, (b) *Mycobacteria* (3708 SPOs): sizes between 2 and 65. Insert histogram highlights sizes larger than 65. (c) *Aspergilli* (6192 SPOs): sizes between 2 and 80. Insert histogram highlights sizes larger than 80.

Name	P_RBH	Par	Motifs SPO26.27 : mtc28, LpqT proteins
MYTU_mtc28	P13.134	P2.46.C2.82	6 7 8 14 2 4 1 3 8 13
MYBO_Mb0041c	P13.134	P2.205.C2.234	6 7 8 14 2 4 1 3 8 13
MYTC_MT0046	P13.134	P2.208.C2.239	6 7 8 14 2 4 1 3 8 13
MYUL_MUL0055	P13.134	P2.39.C2.106	6 7 8 8 14 2 4 1 3 15 8 8
MYMA_MMAR0056	P13.134	P2.77.C2.173	6 7 8 8 14 2 4 1 3 15 8 8
MYLE_ML0031c	P13.134	P2.83.C2.97	6 7 2 4 1 3
MYAV_MAV0054	P13.134	P2.211.C2.334	6 8 14 8 2 2 4 1 1 3 8 8 8 10
MYAP_MAP0047c	P13.134	P2.202.C2.269	6 8 14 8 2 4 1 3 8 8 8 10
MYJL_Mjls5766	P13.134	P2.3.C2.130	6 8 8 2 4 1 3 8
MYVA_Mvan6055	P13.134	P5.29.C2.98	6 7 14 2 4 1 3 8
MYGI_Mflv0852	P13.134	P2.244.C2.360	6 8 14 2 4 1 3 8
MYSM_MSMEG6919	P13.134	P2.4.C2.131	6 8 13 14 2 4 1 8
MYAB_MAB4924	P13.134	P2.2.C2.81	8 8 2 4 1 3 8 8 8
MYTU_lpqT	P13.250	P2.46.C2.82	9 5 2 4 1 11 3
MYBO_Mb1044c	P13.250	P2.205.C2.234	9 5 2 4 1 9
MYTC_MT1044	P13.250	P2.208.C2.239	9 5 2 4 1 11 3
MYUL_MUL4640	P13.250	P2.39.C2.106	5 2 4 1 11 3
MYMA_MMAR4470	P13.250	P2.77.C2.173	5 2 4 1 11 3
MYLE_ML0246c	P13.250	P2.83.C2.97	5 2 4 1 11 3
MYAV_MAV1154	P13.250	P2.211.C2.334	12 5 2 4 1 11 3
MYAP_MAP0981c	P13.250	P2.202.C2.269	12 5 2 4 1 11 3
MYJL_Mjls4630	P13.250	P2.3.C2.130	5 2 4 1 11 3
MYVA_Mvan4785	P13.250	P5.29.C2.98	5 2 4 1 11 3
MYGI_Mflv1941	P13.250	P2.244.C2.360	5 2 4 1 11 3
MYSM_MSMEG5429	P13.250	P2.4.C2.131	5 2 4 1 3
MYAB_MAB1145c	P13.250	P2.2.C2.81	8 2 4 1 3

Fig. 5. Detection of motifs in SPOs and their distribution. Motifs in SPOs are illustrated with the example of SPO26.27, from the 13 *Mycobacteria* considered. This SPO contains proteins corresponding to mtc28 and LpqT. Columns headings: a) Name: species code (see coding conventions in Table 1), followed by the protein identification. The proteins are shown following their species phylogeny. b) P_RBH: partition of RBHs. c) Par: Paralogous class (see text for the coding scheme of *Pn.m.Cp.q* classes). d) Motifs: distributions of motifs as obtained with the meme/mast programs. The distributions highlight motifs shared by all proteins (ancestral motifs: 2,4,1) and motifs shared by subsets of proteins.

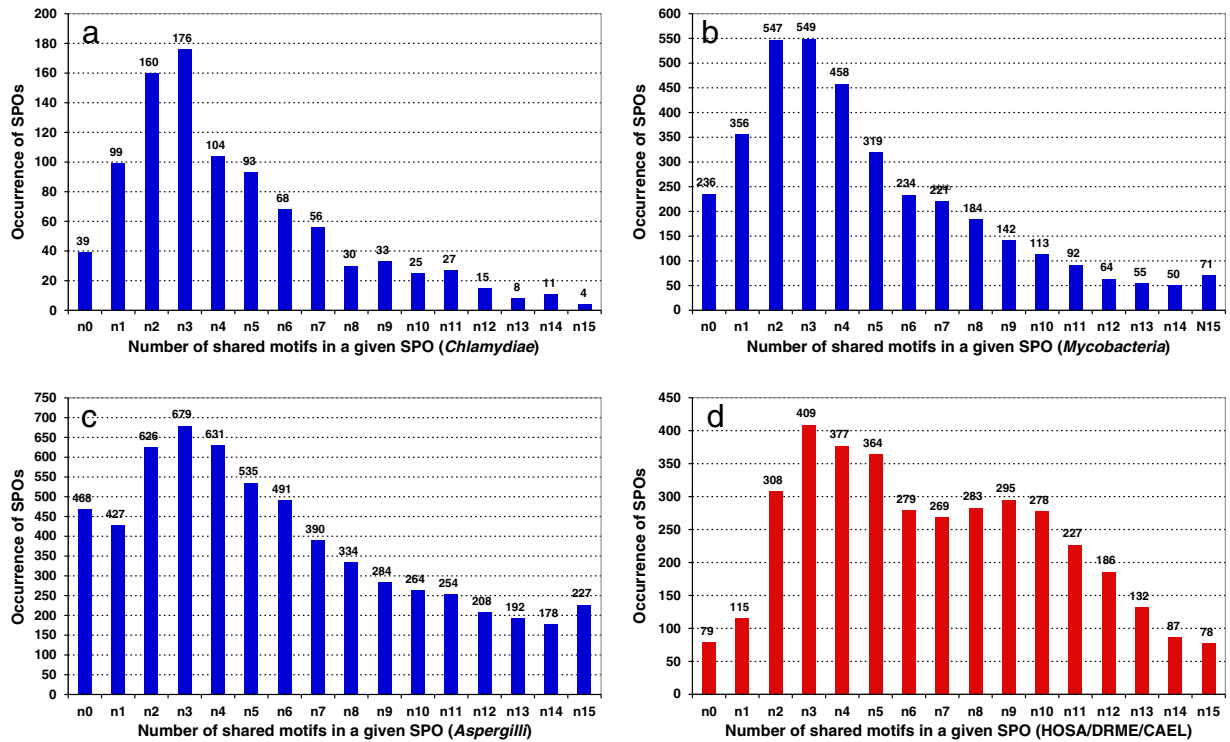


Fig. 6. Distribution of ancestral motifs in SPOs. (a), (b), (c) and (d): SPOs are constructed for the four sets of species (see text): (a) 13 *Chlamydiae* (948 SPOs), (b) 13 *Mycobacteria* (3691 SPOs), (c) 8 *Aspergilli* (6188 SPOs) and (d) *H. sapiens* (denoted HOSA), *D. melanogaster* (denoted DRME) and *C. elegans* (denoted CAEL) (3766 SPOs) used for the comparison of OrthoMCL, MultiParanoid and SPO. Motifs compositions are determined for proteins contained in each SPO, using meme/mast programs (see illustrative example in Fig. 5). These histograms show, for each set, the distribution of the number of SPOs following the number *n* of motifs shared by all proteins ($n_0 = 0, n_1 = 1, \dots, n_{15} = 15$) in a given SPO.

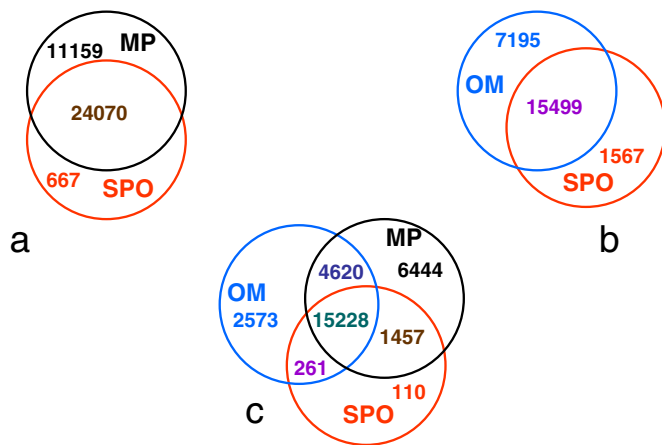


Fig. 7. Venn Diagrams for methods comparisons. The Venn diagrams display the numbers of orthologs detected by the MultiParanoid (MP), the OrthoMCL (OM) and SPO methods, corresponding to common or method-specific detections. (a): SPO and MultiParanoid (MP): comparisons concern predicted orthologs in *H. sapiens*, *D. melanogaster*, *C. elegans* and *C. intestinalis* (representing 35,896 orthologs from a total of 71,199 proteins, see text). (b): SPO and OrthoMCL (OM): comparisons concern predicted orthologs in *H. sapiens*, *D. melanogaster* and *C. elegans* (representing 24,260 orthologs from a total of 56,921 proteins, see text). (c) SPO, OrthoMCL (OM) and MultiParanoid (MP): comparisons concern detected orthologs related to *H. sapiens*, *D. melanogaster* and *C. elegans* (30,613 orthologs).

jointly detected by the two methods, 11,159 proteins detected solely by MP and 667 solely by SPO.

The 667 proteins involved in orthologous classes in SPO but not in MP were as follows: 199 *H. sapiens*, 156 *D. melanogaster*, 134 *C. elegans* and 178 *C. intestinalis* proteins. Similarly the 11,159 proteins involved in orthologous classes in MP but not in SPO were as follows: 5467 human, 1659 fly, 2212 *C. elegans* and 1821 *C. intestinalis* proteins.

Clearly MP led to many more putative orthologs than the SPO method. Also, the number of orthologous clusters in MP (7454 clusters) was larger than that in SPO (4311 clusters). As a matter of fact it appears that only 6240 proteins are identically classified by both methods in 1915 identical clusters (i.e. clusters including the same proteins).

Similarly, Fig. 7b displays the Venn diagram for the comparison between OrthoMCL (OM) and SPO. In this case (see [Materials and methods](#)) only three species (*H. sapiens*, *C. elegans* and *D. melanogaster*) are involved, with the corresponding 56,921 proteins being the same as those corresponding to the same species in MultiParanoid. OrthoMCL detected 6467 orthologous clusters and SPO 3766 clusters. For OM, the corresponding 22,694 proteins (Fig. 7b) are as follows: 9505 human proteins, 7258 fly proteins and 5931 *C. elegans* proteins. Fig. 7b shows that 15,499 proteins were detected by both methods, 7195 proteins were detected by OrthoMCL but not by SPO and 1567 by SPO and not by OM. Thus, as for MP it appears that OrthoMCL detects a significantly larger number of orthologs than SPO. Among the 15,499 detected orthologs by both methods, 4787 proteins are identically classified by both methods in 1772 clusters.

Among the 7195 proteins (3773 human, 2208 fly and 1214 from *C. elegans*) detected by OrthoMCL and not by SPO a subset of 2733 proteins have no significant hit in the two other species, following our Blast comparisons. Such a feature could be accounted for by the potential bias in the OrthoMCL procedure in the handling of the so-called “recent in-paralogs” relative to orthologs (Li et al., 2003). The 4462 proteins in the second subset are not connected by best reciprocal similarity links (i.e. display significant similarity without best reciprocity). This feature could explain why the corresponding proteins were not detected by SPO.

Finally a Venn diagram (Fig. 7c) shows the distribution of the detected orthologs according to the three methods (30,693 proteins; see [Materials and methods](#)). About 50% (15,228) are shared by the

three methods and a varying number of proteins are shared by two methods. MP shows the largest number of specifically detected orthologs (6444), OM (2573), whereas SPO shows the smallest number (110). Examinations of the associated clusters show that parts of them constitute 2623 MP, 118 OM and 6 SPO specific clusters.

The detailed results for the comparisons between methods are provided as additional files (see additional data tables in additional material).

Additional examples (comparing the clustering following different methods) and additional statistics are detailed in Figure S3 in additional material).

Finally, for the global characterization of the SPO merging, we consider in Fig. 6d, the distribution of motifs as determined by the meme/mast programs (see [Materials and methods](#)). This distribution shows that, in 98% of cases, members of a given SPO share at least one motif. This feature can be indicative of the ancestral organization of the proteins included in such SPOs, giving additional support the merging scheme of SPO construction.

4.1.1. Examples and insights

Beyond the overall comparative landscape earlier, we try to go into details to gain insights for the observed differences between the various methods. For the detailed analyses, we resort to an additional classification method for the proteins considered (eggNOG: <http://eggno.gembl.de/>), notably for the functional annotations provided with the associated clusters.

With the overall comparisons it appeared that SPOs provide in many cases much more compact representations than the other methods, in the sense that an SPO can encapsulate significant numbers of orthologous classes from the other methods, the reverse being not observed. Accordingly, to sample as many different cases as possible, we consider as examples SPOs of various sizes, from small to large (chosen at random from the respective size ranges), displaying variable compositions in terms of OM and MP classes (denoted respectively COM and CMP).

The example in Fig. 8a concerns SPO6.109, which is split into two classes both with OM and MP classifications. The corresponding classifications (COM3.505 and COM3.674, CMP3.975 and CMP3.894 respectively) appear to coincide simply with the partitioning scheme (P3.1885 and P3.3008, respectively). In addition eggNOG also leads to a splitting into two classes for the SPO: meNOG5806 and meNOG04739, respectively in correspondence with COM3.505/CMP3.975/P3.1885 and COM3.674/CMP3.894/P3.3008).

Beyond these observations, examination in composition of motifs shows clearly a common ancestral core for the various proteins considered in SPO6.109, along with the splitting into two classes in correspondence with the two partitions of RBHs (left-side variable parts in Fig. 8b). In functional terms (as provided by the eggNOG classification), the class corresponding to the partition P3.1885 is associated with the annotation “U3 small nucleolar ribonucleoprotein” and the class corresponding to the partition P3.3008 is associated with the annotation “Ribosome production factor 1”. It is then relevant to notice that the Ribosome production factor 1 (RPF1) is a Brix domain-containing protein, as for the IMP4 proteins associated with U3 small nucleolar ribonucleoprotein, and both are members of the Imp4 superfamily, with all members possessing a sigma(70)-like motif.

A larger size example concerns the cluster SPO23.5 (Fig. 9). As for the example earlier, we observe a correspondence between MP and OM and P_RBH partitioning for certain sub-clusters in this SPO (such as for example for COM4.70, CMP4.493, and P4.319, with 4 members in all cases). This sub-cluster (relative to the SPO scheme) is also recovered by eggNOG (meNOG07142). Another example for such coherent sub-clustering concerns the 4-membered class COM4.824/CMP4.649/P4.232/meNOG05926. However it is significant that such coherence is not observed in all instances of sub-clustering. For example CMP17.9 concerns a class of 17 members (with only 13 of them clustered in this SPO), with no correspondence with the OM, eggNOG and partitioning

a

Prot ID	OM Cluster	OM ID	MP Cluster	MP ID	SPO	P_RBH	Par Cluster	eggNOG
FBpp0084144	COM3.505	OG2_72253	CMP3.975	4594	SPO6.109	P3.1885	P2.242.C2.410	meNOG5806
ENSP00000259239	COM3.505	OG2_72253	CMP3.975	4594	SPO6.109	P3.1885	P2.272.C2.970	meNOG5806
WBGene00014083	COM3.505	OG2_72253	CMP3.975	4594	SPO6.109	P3.1885	P2.492.C2.791	meNOG5806
FBpp0079946	COM3.674	OG2_72561	CMP3.894	4363	SPO6.109	P3.3008	P2.242.C2.410	meNOG04739
ENSP00000359688	COM3.674	OG2_72561	CMP3.894	4363	SPO6.109	P3.3008	P2.272.C2.970	meNOG04739
WBGene00009711	COM3.674	OG2_72561	CMP3.894	4363	SPO6.109	P3.3008	P2.492.C2.791	meNOG04739

b

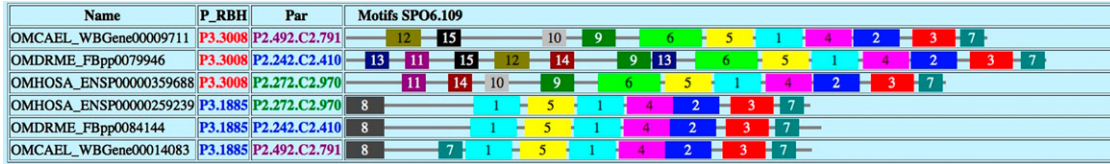


Fig. 8. Detailed examples for methods comparisons. The figure illustrates details of analyses for the 6 proteins in SPO6.109, following methods. (a) OM, MP and SPO clustering: the columns correspond to 1) Prot_ID: protein identifications, 2) OM_Cluster: OM cluster identifications (in Comp.q notation p concerns the number of proteins in the cluster and q an arbitrary index), 3) OM_ID: OM original cluster identifications (see Materials and Methods), 4) MP_Cluster: MP cluster identification (in CMPp.q notation p concerns the number of proteins in the cluster and q an arbitrary index), 5) MP_ID: MP original cluster identification (see Materials and Methods); 6) SPO: SPO cluster identification, 7) P_RBH: RBH partition identification and 8) Par_Cluster: mcl cluster identification (mcl clustering corresponds to the Cp.q class, with p concerning the number of proteins in the cluster and q an arbitrary index; each Cp.q is further labeled with additional information relative to Pn.m clustering, see text). (b) Motifs: motifs compositions for the proteins in SPO6.109, with motifs (5,1,4,2,3,7) shared by all members.

schemes. More precisely, the 13 members of this class in the SPO are associated with 5 different classes for the OM scheme, in turn either coherent or not with the respective eggNOG classes. At this level, the highest correspondence is observed between the associated OM classes and the simple P_RBH partitioning scheme (with however possible discrepancies, such as the splitting of the partition P4.415 into the two OM classes, COM2.1716 and COM3.2077).

In terms of ancestry, analyses of the distribution of shared motifs between all proteins in SPO23.5 show the existence of a core anchoring motifs (4,2,3) shared by all proteins (Fig. 10).

What rationale for the observations earlier, concerning the various merging schemes? Concerning the SPO23.5, for example, is the splitting of the partition P4.415 into the sub-clusters COM2.1716 COM3.2077 justified? In fact it appears that such splitting dispatches into two distinct clusters two human gene variants LMX1A and LMX1B. More generally speaking, what could be the rationale in the higher-order merging associated with this whole SPO as compared to the numerous sub-clusters associated with MP, OM, eggNOG methods (and of course P_RBH

partitioning)? The functional annotation provided by eggNOG reveals that all the members in the SPO are reported under one of the following three categories: “Lim homeobox protein”, “Rhombotin 1 protein” or “Insulin gene protein”. The Rhombotin 1 protein is further associated with a “Lim domain only protein 1” annotation. Also the Insulin gene protein is further associated with the annotation “Lim domain family protein”. These global functional annotations thus provide potential rationale at the basis of the clustering in this SPO.

5. Conclusions and discussion

The quest for orthologs is at the heart of genomics, in terms of methodological formulations as well as practical routine analyses. As a matter of fact, on practical grounds the identification of orthologs represents a cornerstone in many evolutionary genomic analyses. In parallel, methodological developments for accurate methods and more reliable performances still represent an active ongoing field, as testified by many recent reviews on the subject. With comparative

Prot_ID gene name	OM_Cluster	OM_ID	MP_Cluster	MP_ID	SPO	P_RBH	Par_Cluster	eggNOG_ID
FBpp0085394 CG8376 ap	COM3.2409	OG2_78806	CMP17.9	1429	SPO23.5	P2.285	P1576.1.C9.10	meNOG06518
ENSP00000362717 LHX2	COM3.2409	OG2_78806	CMP17.9	1429	SPO23.5	P2.285	P6875.1.C6.42	meNOG06518
FBpp0074370 CG6500 Bx	COM3.1753	OG2_76282	CMP3.239	6716	SPO23.5	P2.799	P1576.1.C9.10	meNOG09002
ENSP00000338207 TTG1	COM3.1753	OG2_76282	CMP3.239	6716	SPO23.5	P2.799	P6875.1.C6.42	meNOG09002
FBpp0071216 CG1135 Lim1	COM3.2496	OG2_79336	CMP17.9	1429	SPO23.5	P3.1956	P1576.1.C9.10	meNOG07169
ENSP00000254457 LHX1	COM3.2496	OG2_79336	CMP17.9	1429	SPO23.5	P3.1956	P6875.1.C2.91	meNOG07169
WBGene000003000 lin-11	COM3.2496	OG2_79336	CMP17.9	1429	SPO23.5	P3.1956	P96.1.C7.44	meNOG07169
FBpp0080712 CG10699 Lim3	COM4.702	OG2_75554	CMP4.493	2383	SPO23.5	P4.319	P1576.1.C9.10	meNOG07142
ENSP00000263726 LHX4	COM4.702	OG2_75554	CMP4.493	2383	SPO23.5	P4.319	P6875.1.C2.90	meNOG07142
ENSP00000360811 LHX3	COM4.702	OG2_75554	CMP4.493	2383	SPO23.5	P4.319	P6875.1.C2.90	meNOG07142
WBGene00000438 ceh-14	COM4.702	OG2_75554	CMP4.493	2383	SPO23.5	P4.319	P96.1.C7.44	meNOG07142
FBpp0080661 CG10619 tup	COM4.824	OG2_76275	CMP4.649	322	SPO23.5	P4.323	P1576.1.C9.10	meNOG05926
ENSP00000230658 ISL1	COM4.824	OG2_76275	CMP4.649	322	SPO23.5	P4.323	P6875.1.C2.120	meNOG05926
ENSP00000290759 ISL2	COM4.824	OG2_76275	CMP4.649	322	SPO23.5	P4.323	P6875.1.C2.120	meNOG05926
WBGene00002989 lim-7	COM4.824	OG2_76275	CMP4.649	322	SPO23.5	P4.323	P96.1.C7.44	meNOG05926
FBpp0075685 CG32105	COM2.1716	OG2_86734	CMP17.9	1429	SPO23.5	P4.415	P1576.1.C9.10	meNOG10273
ENSP00000340226 LMX1A	COM2.1716	OG2_86734	CMP17.9	1429	SPO23.5	P4.415	P6875.1.C2.108	meNOG10273
ENSP00000362573 LMX1B	COM3.2077	OG2_77489	CMP17.9	1429	SPO23.5	P4.415	P6875.1.C2.108	meNOG10273
WBGene00002988 lim-6	COM3.2077	OG2_77489	CMP17.9	1429	SPO23.5	P4.415	P96.1.C7.44	meNOG10273
FBpp0073014 Awh	COM3.1822	OG2_76526	CMP17.9	1429	SPO23.5	P4.476	P1576.1.C9.10	meNOG08070
ENSP00000294638 Lhx8	0	0	CMP17.9	1429	SPO23.5	P4.476	P6875.1.C2.99	meNOG08512
ENSP00000362861 LHX6	COM3.1822	OG2_76526	CMP17.9	1429	SPO23.5	P4.476	P6875.1.C2.99	meNOG08070
WBGene00002987 lim-4	COM3.1822	OG2_76526	CMP17.9	1429	SPO23.5	P4.476	P96.1.C7.44	meNOG08070

Fig. 9. Members of SPO23.5. The figure shows the members of SPO23.5 with associated OM, MP as well as eggNOG clusters. Columns headers are as in Fig. 8.



Fig. 10. Distribution of motifs in SPO23.5. The figure shows the distribution of the motifs in SPO23.5 as determined by meme/mast programs. A core motif (4,2,3) is shared by all proteins in this cluster, indicating their ancestral organisation.

analyses of the different methods, such reviews clearly highlight the various limitations and drawbacks.

In this general background, we formulate here a simple operational framework to solve efficiently the problem of the detection of orthologs and their classification. On conceptual backgrounds, our procedure overlaps to a large extent with previous methods based on “Reciprocal Best Hits” results, as a difference to phylogeny-based approaches. However, as compared to other methods, the SPO formulation here is designed with a specific handling of the RBH-related information for the building of orthologous clusters. Beyond this basic difference related to merging schemes, the SPO method is also characterised by the elaboration of a specific coherent environment, involving supplementary post-processing steps (such as related to motifs and conservation profiles), to further analyze and make full use of the information in the derived clusters. As a matter of fact, the post-processing components could be straightforwardly implemented in various other methods as well:

- 1) Construction of conservation profiles for the SPOs (or clusters derived from other methods) allows the detection of in-paralogs and out-paralogs (in SPOs containing more than one protein from a species) and further fine-grained evolutionary analyses.
- 2) Detailed characterizations of SPOs (or clusters derived with other methods) in terms of motifs compositions, allow to assess the presence of putative ancestral signals and the possible influence of chaining effects. With the SPO approach, as illustrated with detailed studies for bacterial and eukaryal species, it appears that chaining effects that might arise in multi-domain proteins are avoided to a large extent.

Concerning the basic merging scheme, the SPO method takes advantage of the recognized efficiency of RBH detections, and the associated ease of computation, remedying in the same time, by design, to the reported limitations concerning the detection of many-to-(one and/or many) orthologs. More precisely, at the heart of orthology detection methods, merging schemes need to handle appropriately the results of intra and inter-species comparisons for the building of orthologous clusters. With this respect, as a difference to other methods, in the SPO formulation the original RBH information (relative to the inter-species comparisons) are preserved under the form of P_{RBH} clusters, without being further introduced into the clustering process also involving the putative paralogs obtained from intra-species comparisons. In the SPO formulation the intra-species information (in terms of Reciprocal Significant Hits, RSHs) are processed separately, independently from the inter-species comparisons, with

mcl clustering leading to Cp.q classes. In our scheme, the final SPO orthologous classes are then obtained through higher-order merging, bringing together in the same class RBH partitions which contain members of common paralogous Cp.q classes (the overall process is schematically represented in Fig. 2, with proteins belonging to the same intra-species cluster represented in a given colour). The rationale in this design is to minimize as much as possible the number of broken RBH sets, thus avoiding the corresponding spurious dispatch of paralogous proteins into different orthologous classes. It is then important to notice that in various other RBH-based methods, as a difference to the SPO formulation, the merging scheme relies on the direct clustering of putative paralogs and orthologs, processed simultaneously with this respect. In such context it is also important to stress that the mcl algorithm concerns different operations in OrthoMCL and SPO methods, clustering respectively putative orthologs and paralogs in the first case and only putative paralogs in the second case.

What is the effect in practice of the design concepts in the effective building of orthologous classes for model systems? The effectiveness of the SPO approach was assessed through a series of studies, in terms of internal consistency with global characterisations for three sets of proteomes. More significantly, with respect to the different merging schemes, the SPO method was compared in detail with two other methods (OrthoMCL and MultiParanoid) in terms of orthologous clustering. The comparisons were on the basis of common data sets (for the *H. sapiens*, *D. melanogaster* and *C. elegans* species) available for these methods, leading to a series of striking observations and conclusions, with potential far-reaching consequences for orthology detections.

Reflecting the higher-order merging involved in SPOs, it is not surprising to observe that SPOs lead to more compact representations than the other methods, with a given SPO clustering together a more or less large number of sub-clusters associated with orthologous classes in the other methods. Such an observation is not conclusive by itself, without assessing the potential validity of the different levels of clustering. The validation of the overall scheme is then with the following rationale:

- 1) It appears, to some extent, that various orthology methods (OrthoMCL, MultiParanoid, and also eggNOG, even though in this case only a few examples were considered) lead to coherent clustering, at least for the simplest cases (for which the relatedness links should be the most obvious). However, for many cases, with relatedness links increasingly weaker, the different methods display little coherence in the clustering. A striking observation relates then these results with SPO analyses: the level of clustering associated with the various methods appears to correspond to the

level of the mere P_{RBH} classes in the SPOs, with, as a matter of fact, a detailed correspondence in the simplest cases between the simple P_{RBH} clusters and the clusters with the various other methods.

- 2) The observed splitting of a given SPO into variable number of sub-clusters in other methods is easily accounted for on conceptual grounds by the fundamental difference in the handling of RBH-related information. The SPO approach is designed to preserve the original RBH connections, in order to avoid the dispatching of related paralogs into distinct orthologous clusters.

Do practical analyses confirm the effectiveness of this design principle?

Several converging arguments tend indeed to validate the clustering scheme with this respect, notably through the detailed analysis of motifs compositions (detection of common ancestral motifs), or case studies relative to functional annotations. On a more trivial level, it is also possible to mention the observation that sub-clusters, within a given SPO, associated with other methods appear to split in various examples obviously related proteins into different classes.

- 3) Beyond the possible dispatch of relatively distant paralogs into distinct classes, it was highlighted (OrthoMCL: Li et al., 2003) that clustering schemes handling together putative paralogs and orthologs could lead to the somehow opposite bias in presence of very close paralogs: ‘the high similarity of “recent” paralogs relative to orthologs can bias the clustering process’. Accordingly, in order to minimize such biases, ad hoc procedures were introduced in these formulations in terms of more or less sophisticated weighting and normalization schemes. With practical analyses, it however appears that such biases are nevertheless observed in the corresponding methods, as illustrated earlier by various very large clusters concerning essentially variants of a single protein. With respect to this problem, it is interesting to point that the present implementation of SPO method can miss close paralogs corresponding to such situations. A possible solution out of this situation would be to “choose” one appropriate representative for such classes, keeping track of the associated members for further processing. At present rather a conservative solution was adopted, notably in view of different potentially delicate situations (following database annotations) such as those concerning the presence of isoform proteins, proteins resulting from alternative splicing.

In conclusion the framework presented here allows the versatile detection and clustering of orthologs and in-paralogs, with conservation profiles providing further possibilities for downstream in-depth evolutionary analyses.

Acknowledgments

We thank Bernard Dujon and Roland Brosch for insightful discussions and Michael Nilges for constant support. FT acknowledges financial support from the Institut Pasteur (PTR No 264 and 370).

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.gene.2011.10.027.

References

Alexeyenko, A., Tamas, I., Liu, G., Sonnhammer, E.L., 2006. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* 22, e9–e15.

Altenhoff, A.M., Dessimoz, C., 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol* 5, e1000262.

Altschul, S.F., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Bailey, T.L., Elkan, C., 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, California, pp. 28–36.

Catchen, J.M., Conery, J.S., Postlethwait, J.H., 2009. Automated identification of conserved synteny after whole-genome duplication. *Genome Res.* 19, 1497–1505.

Chen, F., Mackey, A.J., Vermunt, J.K., Roos, D.S., 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE* 2, e383.

Cole, et al., 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393, 537–544.

Dehal, P.S., Boore, J.L., 2006. A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics* 7, 201.

Enright, A.J., Kunin, V., Ouzounis, C.A., 2003. Protein families and TRIBES in genome sequence space. *Nucleic Acids Res.* 31, 4632–4638.

Enright, A.J., Van Dongen, S., Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* 30, 1575–1584.

Fang, G., Bhardwaj, N., Robilotto, R., Gerstein, M.B., 2010. Getting started in gene orthology and functional analysis. *PLoS Comput Biol* 26 (6(3)), e1000703.

Fitch, W.M., 1970. Distinguishing homologous from analogous proteins. *Syst Zool.* 19, 99–113.

Fitch, W.M., 2000. Homology a personal view on some of the problems. *Trends Genet.* 16, 227–231.

Forest, F., 2009. Calibrating the Tree of Life: fossils, molecules and evolutionary time-scales. *Ann. Bot.* 104, 789–794.

Gabaldón, T., 2008. Large-scale assignment of orthology: back to phylogenetics? *Genome Biol.* 9, 235.

Gey van Pittius, et al., 2006. Evolution and expansion of the *Mycobacterium tuberculosis* PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. *BMC Evol Biol.* 6, 95. doi:10.1186/1471-2148-6-95.

Goodstadt, L., Ponting, C.P., 2006. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput Biol* 2, e133.

Jensen, L.J., et al., 2008. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.* 36, D250–D254 (Database issue).

Koonin, E.V., 2005. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet.* 39, 309–338.

Kristensen, D.M., et al., 2010. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics* 26, 1481–1487.

Kriventseva, E.V., Rahman, N., Espinosa, O., Zdobnov, E.M., 2008. OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res.* 36, D271–D275 (Database issue).

Kuzniar, A., van Ham, R.C., Pongor, S., Leunissen, J.A., 2008. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.* 24, 539–551.

Li, L., Stoeckert Jr., C.J., Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189.

Linard, B., Thompson, J.D., Poch, O., Lecompte, O., 2011. OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics* 10 (12), 11.

Lynch, M., Conery, J.S., 2003. The evolutionary demography of duplicate genes. *J Struct Funct Genomics* 3, 35–44.

Moreno-Hagelsieb, G., Latimer, K., 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics* 3, 319–324.

Poptsova, M.S., Gogarten, J.P., 2007. BranchClust: a phylogenetic algorithm for selecting gene families. *BMC Bioinformatics* 8, 120.

Remm, M., Storm, C.E., Sonnhammer, E.L., 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol.* 314, 1041–1052.

Salichos, L., Rokas, A., 2011. Evaluating Ortholog Prediction Algorithms in a Yeast Model Clade. *PLoS ONE* 6 (4), e18755. doi:10.1371/journal.pone.0018755.

Schäffer, A.A., et al., 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* 29, 2994–3005.

Sonnhammer, E.L., Koonin, E.V., 2002. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.* 18, 619–620.

Studer, R.A., Robinson-Rechavi, M., 2009. How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet* 25, 210–216 (Epub 2009 Apr 14).

Tatusov, R.L., et al., 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4, 41.

Tekaia, F., Gordon, S.V., Garnier, T., Brosch, R., Barrell, B.G., Cole, S.T., 1999. Analysis of the proteome of *Mycobacterium tuberculosis* in silico. *Tuber Lung Dis* 79, 329–342.

Tekaia, F., Yeramian, E., 2005. Genome trees from conservation profiles. *PLoS Comput Biol* 1, e75.

Tekaia, F., et al., 2000. Methods and strategies used for sequence analysis and annotation. *FEBS* 487, 17–30.

Tekaia, F., Dujon, B., 1999. Pervasiveness of gene conservation and persistence of duplicates in cellular genomes. *J. Mol. Evol.* 49, 591–600.

Tekaia, F., Latge, J.P., 2005. *Aspergillus fumigatus*: saprophyte or pathogen? *Curr Opin Microbiol.* 8, 385–392.

Tekaia, F., Yeramian, E., Dujon, B., 2002. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. *Gene* 297, 51–60.

Tekaia, F., Yeramian, E., 2006. Evolution of proteomes: fundamental signatures and global trends in amino acid compositions. *BMC Genomics* 7, 307.

Zmasek, C.M., Eddy, S.R., 2004. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* 3, 14.