Evolutionary Bioinformatics

METHODOLOGY

IONS: Identification of Orthologs by Neighborhood and Similarity—an Automated Method to Identify Orthologs in Chromosomal Regions of Common Evolutionary Ancestry and its Application to Hemiascomycetous Yeasts

Marie-Line Seret and Philippe V. Baret

Université Catholique de Louvain, Earth and Life Institute (ELI), 1348 Louvain-la-Neuve, Belgium.
Corresponding author email: philippe.baret@uclouvain.be

Abstract: Comparative sequence analysis is widely used to infer gene function and study genome evolution and requires proper ortholog identification across different genomes. We have developed a program for the Identification of Orthologs in one-to-one relationship by Neighborhood and Similarity (IONS) between closely related species. The algorithm combines two levels of evidence to determine co-ancestrality at the genome scale: sequence similarity and shared neighborhood. The method was initially designed to provide anchor points for syntenic blocks within the Génolevures project concerning nine hemiascomycetous yeasts (about 50,000 genes) and is applicable to different input databases. Comparison based on use of a Rand index shows that the results are highly consistent with the pillars of the Yeast Gene Order Browser, a manually curated database. Compared with SYNERGY, another algorithm reporting homology relationships, our method's main advantages are its automation and the absence of dataset-dependent parameters, facilitating consistent integration of newly released genomes.

Keywords: ortholog, synteny, shared neighborhood, hemiascomycete, yeast

Evolutionary Bioinformatics 2011:7 123–133

doi: 10.4137/EBO.S7465

This article is available from http://www.la-press.com.

Introduction

Given the increasing number of large-scale sequencing projects (see http://www.ncbi.nlm.nih.gov/Genomes), comparative genomic approaches are now widely used.[1–7] Indeed, comparison of genome sequences across species offers great potential for studying many aspects of their underlying biology, such as the prediction of gene function. Moreover, it provides insights into the processes of both genome and gene evolution. The reliable identification of orthologs in a one-to-one relationship is critical for many comparative genomics analysis, such as the construction of syntenic blocks,[1,5,8] the reconstruction of accurate species or gene trees, or the automation of the functional annotation of genes.

Orthology is often interpreted as the functional equivalence of proteins across species, while in fact it defines only a particular relationship of homology in which two genes originating from a common single ancestral gene diverged following a speciation event.[9,10] However, orthologs are more likely to have a functional similarity than paralogous genes.[11]

In the process of identifying orthologs, the trend is to assume that if two sequences are significantly similar, they must be homologous; ie, they must share a common origin.[9,10] However, similarity may be a false indication of homology, for example, in cases of convergence and events of duplication and loss that tend to blur the tracing of co-ancestrality.[1] Therefore, the identification of orthologs among the set of homologs defined by similarity requires more specific analysis.

Most approaches for the identification of orthologs may be based on the following evidence: sequence similarity, reconciliation of genes and species phylogenies, and synteny conservation (see[11–16]). The implementation is either manual, semi-automatic (some parameters are defined *a priori* and differ from one dataset to the other), or automatic with constant parameters. To cope easily with the availability of new genomes, the challenge is to develop a simple automated method in which parameters are not modified by the addition of new information.

We developed a program called IONS (Identification of Orthologs by Neighborhood and Similarity). This program relies on two types of evidence: sequence similarity at the protein level and the chromosomal neighborhood (see Algorithm). The method was initially developed for the Génolevures project, a large-scale comparative genomics project for *Saccharomyces cerevisiae* and other yeast species representative of the various branches of the hemiascomycetous class, which notably provides annotated sequence data and classifications of nine complete Hemiascomycete yeast genomes comprising a total of about 50,000 genes. In this context, it was used to identify subsets of orthologs used as anchor points for the construction of syntenic blocks.[1,17] In practice, blocks of conserved synteny are delineated as regions containing numerous orthologous genes that can be separated by a limited number of intervening genes (non-orthologous genes).[18] This preliminary version of the method was also used to study the evolution of families of transporters.[19,20] The purpose of the present paper is to give a full description of the final version of the method and to discuss its advantages in comparison with two other methods, the Yeast Gene Order Browser (YGOB) and SYNERGY.

Comparisons required application of the different methods to a common dataset, in our case the hemiascomycete phylum. The adjusted Rand index (ARI)[21] was used as a quantitative indicator of the equivalence of partitioning for pairwise comparisons of methods.

Material and Methods

Algorithm

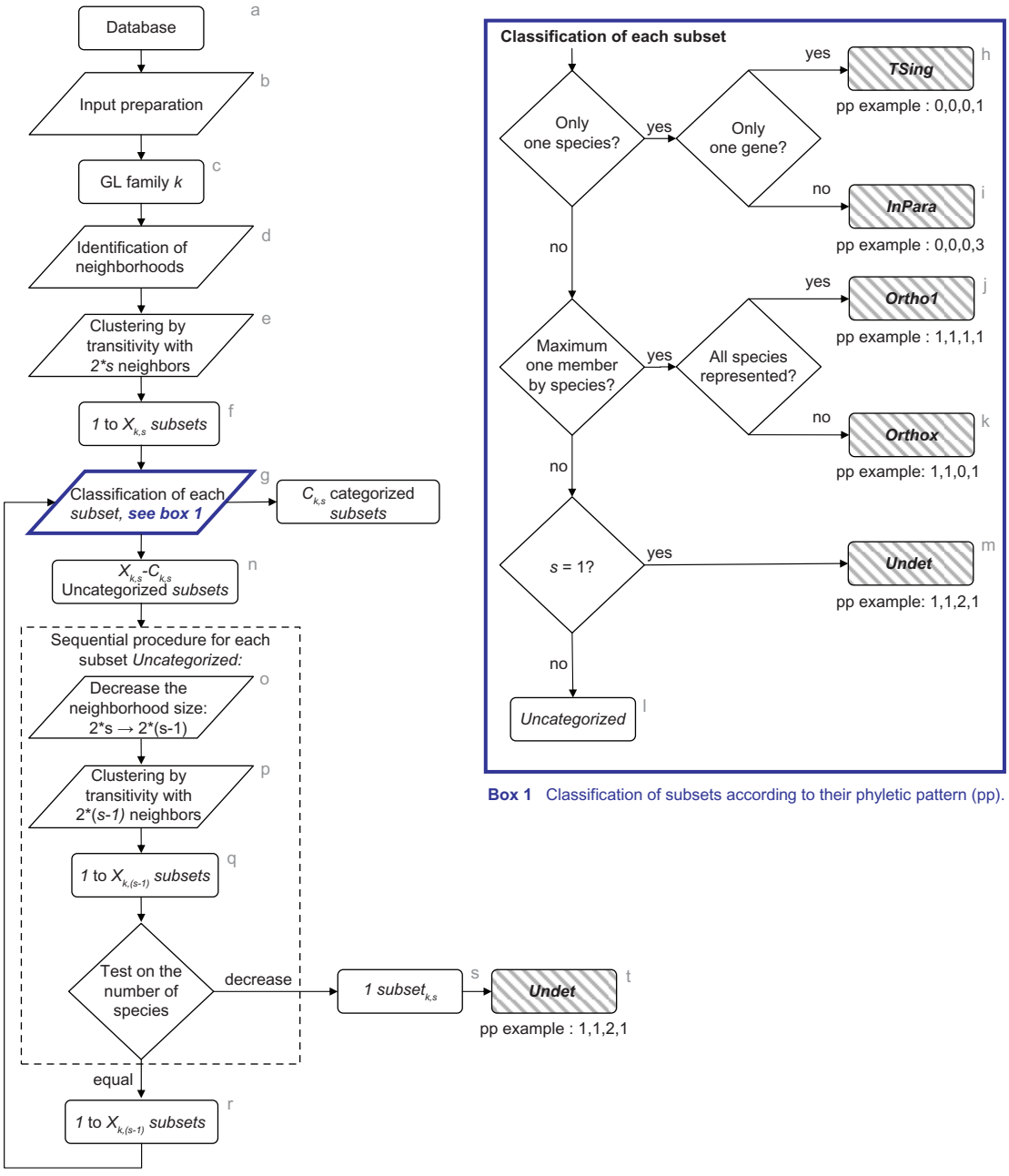
Input data and preparation

The input required for IONS consists of a database of genes belonging to a number of species encompassing a particular taxonomic class or phylum. This database must contain comprehensive information about the relative position of genes as well as assignment to a group of similar gene products or to any other group of putative homologs obtained by any other method (Additional file 1: Sample of input and output files).

For this study, the database (Fig. 1a) consisted of the genomes of nine species covering the Hemiascomycetes class and was based on the assignment to Génolevures families (GL Family) defined by Nikolski et al.[22] These families of similar genes were based on the calculation of pairwise similarities of sequences provided by BLAST and Smith-Waterman and a subsequent clustering. An algorithm was then applied

to construct consensus families from competing clustering computations by an election method (see[22] for details).

Figure 1 presents a flowchart of the algorithm we developed to infer orthology from the combination of similarity (family) and shared neighborhood, defined as the preserved co-localization of some genes on chromosomes of different species (independently of their order). First, the database was read to identify the list of genes belonging to each family (Fig. 1b).



Box 1 Classification of subsets according to their phyletic pattern (pp).

Figure 1. Flowchart of the IONS algorithm. The algorithm is automatically applied to each GL family k. The first clustering by transitivity, with s neighbors taken into account for each side of the query genes, produces 1 to X subsets with the genes belonging to the family k. These $subsets_{k,s}$ are classified into different categories: TSing, InPara, Ortho1, Orthox, and Undet. Subsets that do not correspond to one of these categories are uncategorized and enter the sequential procedure. In the sequential procedure, a test is made to ensure that the new subsets created with the narrow neighborhood (Y $subsets_{k,(s-1)}$) do not result in a reduction of the number of species. A comparison is made between the number of species represented in the subset with the wide neighborhood $subset_{k,s}$ and the maximum number of species represented in the 1 to Y $subsets_{k,(s-1)}$ obtained with the narrow neighborhoods. If the number of species is equal, the 1 to Y $subsets_{k,(s-1)}$ obtained with the narrow neighborhood enter the classification. Otherwise, the subset with the wide neighborhood, $subset_{k,s}$, is validated and labeled as "Undet." Striped frames indicate the end of the analysis for the genes belonging to the labeled subsets. Below these frames is an example of a possible phyletic pattern (pp).

Then all families (Fig. 1c) were sequentially analyzed as described below.

Identification of neighborhoods

For a given family k having i members, the set of s neighbor genes $N_{i,k,n}$ with n comprised between [−s,s] identified on each side of a "query gene" $G_{i,k}$ defines a neighborhood of size 2*s genes (Fig. 1d and additional file 2: Visual representation of main steps of the IONS algorithm leading to a subset of orthologs).

Assignment of genes to subsets through a clustering by transitivity

IONS proceeds by comparisons of the neighborhoods of the i query genes belonging to a particular family k. The translation products of the $2*s*i$ neighbor genes are tested pairwise: if a neighbor of a query gene belongs to the same family as a neighbor of another query gene (ie, if these genes are similar in sequence), the neighborhoods of the two query genes have one neighbor in common.

While analyzing a given family k, two query genes G_{ik} having at least one neighbor in common are assigned to the same cluster. It is noteworthy that, if the neighborhoods of two query genes have at least one neighbor in common with the neighborhood of a third gene, but none with each other, the three query genes are assigned to the same cluster, through transitivity (Fig. 1e and Fig. 2).

The program offers the opportunity to change the number of neighbors required to assign genes to the same subset of orthologs. In all cases, the output of the analysis is either a confirmation of the initial family on the basis of the neighborhood evidence or a splitting of the family into different clusters (Fig. 1f), which are sequentially numbered (eg, GL3R1304_10010, GL3R1304_10020, etc.).

Classification of subsets (Fig. 1g and Box 1)

In a first step, the widest neighborhood size was used to calculate clusters, eg, s = 15 if 15 neighbor genes were identified on each side of the query gene. The number of query genes in each species was called the phyletic pattern (pp) of a subset. According to our classification, subsets with genes in only one species were labeled as "TSing," for technical singleton (Fig. 1h), if they comprised one gene only, and as "InPara" for in-paralogs (paralogs in a given lineage that all evolved by gene duplications that happened *after* the speciation event that separated the lineage under consideration from the other lineages[23]) (Fig. 1i), if they comprised more than one gene. Subsets with a maximum of one gene in different species were called SONS (Subsets of Orthologs defined by Neighborhood and Similarity)[19] and labeled as "Ortho1" (Fig. 1j) if one gene was present in each species (pp: 1,1,1,1) and "Orthox" (Fig. 1k) if at least one species did not have any gene assigned to this subset (pp: 1,1,1,0 eg,). Other subsets comprising multiple genes from different species were considered as "Uncategorized" (Fig. 1l) and further assessed by a sequential procedure.

Sequential procedure

When a subset comprised genes from different species and if some species comprised more than one member (subsets labeled as "Uncategorized"), we progressively diminished the size of the neighborhood taken

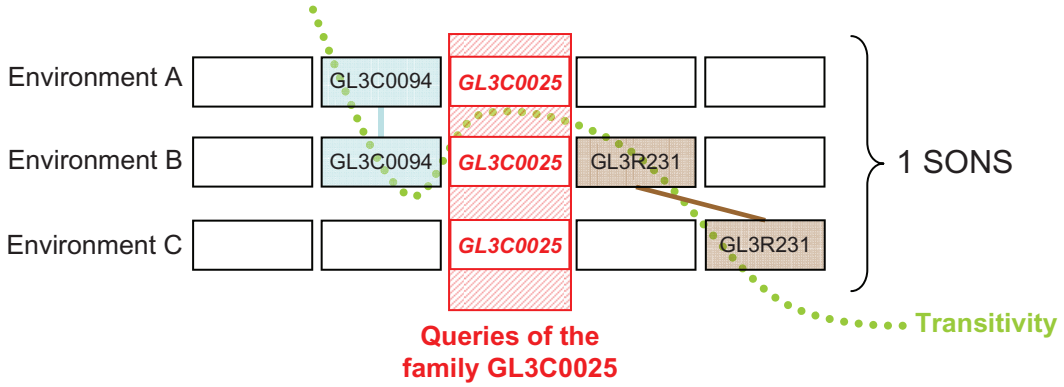


Figure 2. Clustering by transitivity. If the environment (**A**) possesses a homologous protein in (**B**) and if the environment (**C**) possesses a protein that is homologous to another protein of (**B**), then the environments (**A, B**, and **C**) are parts of the same subset of orthologs by neighborhood and similarity.

into account. Subsets in the "Uncategorized" category (Fig. 1n) were tested with a narrower neighborhood size, eg, $s = 14$ for 14 genes each side (Fig. 1o). The clustering step (Fig. 1p) with the narrower neighborhood either confirmed the subset, if all neighborhoods still tied up the different homologs, or split it into a number of new subsets (Fig. 1q). Indeed, if a neighborhood was linked to the others by the 15th neighbor gene only, the subset was split when this part of the neighborhood was no longer taken into account.

If at least one of the new subsets comprised the same number of species as the initial subset in the wide neighborhood, the new subsets were validated (Fig. 1r) and a supplementary suffix was added (ie, GL3R1304_10010_10, GL3R1304_10010_20). In turn, these new subsets were labeled as described in the previous section. "Uncategorized" subsets were recursively tested with a narrower neighborhood size, eg, $s = 13$ (13 genes each side), and so on up to $s = 1$.

In the search of the orthologs in a one-to-one relationship, the algorithm was designed to privilege clusters of orthologs and in-paralogs rather than to split orthologs apart. Thus, if all new subsets comprised fewer species than the initial subset in the wide neighborhood, the procedure was stopped. In this case, genes were assigned to the subset defined in the previous step in terms of neighborhood width (Fig. 1s) and labeled as "Undet" for undetermined (Fig. 1t).

During the sequential procedure, if the neighborhood size (s) was equal to one neighbor, the procedure was halted and the subset considered as "Undet" (Fig. 1m). This "Undet" label means that the subset contained homologous genes for which the relationship (orthology or in-paralogy) could not be assessed using our neighborhood criteria.

A test case on Hemiascomycetes described in the next section illustrates the interest of the sequential procedure. With 15 neighbors, the IONS procedure assigned 33,258 of the 47,874 genes (see Fig. 3) to different types of subsets. Among these, 22,758 genes (47.53% of the total) formed subsets of orthologs (Ortho1 and Orthox). The sequential procedure was then applied to the remaining 14,616 genes. Additional file 3 shows the cumulative results obtained after each step of this sequential procedure at the end of which 6,521 additional genes formed new subsets of orthologs, increasing the total percentage of genes classified into subsets of orthologs (Ortho1 and Orthox) to 61.16% of all 47,874 genes.

Output

The IONS program produces two databases as well as a visual file and a neighborhoods file for each subset (Additional file 1: Sample of input and output files). The first database is ordered by gene and contains its family and the name and status of the subset to which it was assigned. The second database is organized by subsets. Each shows the family, the subset, the status of the subset, the number of genes of each species constituting this subset, and the total number of genes in the subset.

The neighborhoods file reports the presence–absence of the different families in the neighborhoods of the different queries of the subset. The visual file shows the names and families of the neighbors of genes belonging to a particular subset. The raw descriptions of neighborhood relationships available in these visual files may serve as support for ad hoc discussion of gene evolution in complex situations, like the emergence of ohnologs (duplicates arising from the Whole Genome Duplication,[24] abbreviated as WGD).

Results

Test case on nine hemiascomycetes

We applied the IONS method to resolve the homology relationships in the genomes of nine hemiascomycetous yeasts: *Saccharomyces cerevisiae* (SACE), *Candida glabrata* (CAGL), *Zygosaccharomyces rouxii* (ZYRO), *Saccharomyces (Lachancea) kluyveri* (SAKL), *Kluyveromyces (Lachancea) thermotolerans* (KLTH), *Kluyveromyces lactis* (KLLA), *Ashbya (Eremothecium) gossypii* (ERGO), *Debaryomyces hansenii* (DEHA), and *Yarrowia lipolytica* (YALI). These genomes add up to 47,874 proteins, which were classified into 7,927 families (see additional file 4: Results of the IONS method application to the 47,874 CDS of Génolevures), as well as 1,015 proteins being left aside because of ambiguous or complex affiliations[1] (Fig. 4).

The results in Génolevures

Figure 3 shows the distribution of genes into the five categories of subsets (TSing, InPara, Ortho1, Orthox,

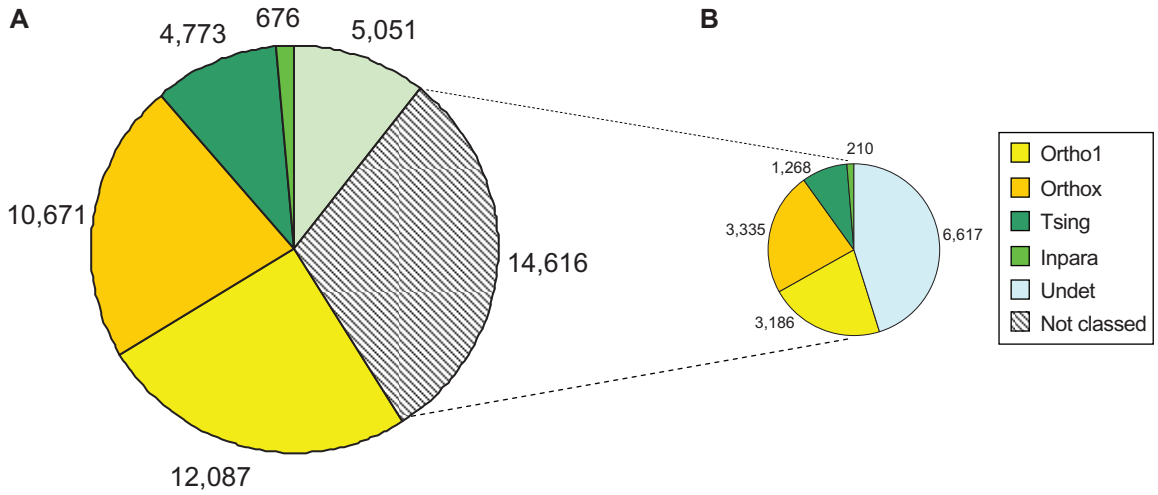


Figure 3. Improvement of the classification by the sequential procedure. (**A**) Distribution of genes in the different types of subsets resulting from the analysis with 15 neighbors taken on each side of the query genes and (**B**) improvement of the classification thanks to the sequential procedure that allows classifying genes not classified in (**A**).
Abbreviations: Ortho1, subset with one gene in each species; Orthox, subset with maximum one gene per species, all species are not represented; TSing, subset with one gene from one species only; InPara, subset with several genes belonging to the same species; Undet, subset with several species represented, at least one species represented by more than one gene.

and Undet). With a maximum of 15 neighbors taken into account, the neighborhood was informative, ie, there was at least one neighbor in common with at least one other query gene, for 43,101 genes (90%). The method confirmed the co-ancestrality of orthologs in a one-to-one relationship of 1,309 families containing one gene in each species as well as identifying 388 new subsets of orthologs of this type (Table 1).

Genes:	A - Total	B - in SONS	C- SONS	D - Pillars	E - Orthogroups
	48,889	***29,279***	***3,761***	***3,675***	***3,639***
S. cerevisiae	5,769	3,448	3,439	3,408	3,276
C. glabrata	5,204	3,319	3,263	3,226	3,113
Z. rouxii	4,998	3,440			
K. thermotolerans	5,104	3,513			
S. kluyveri	5,307	3,603			
K. lactis	5,084	3,449	2,672	2,667	2,610
E. gossypii	4,718	3,325	3,317	3,321	3,201
D. hansenii	6,271	2,870			
Y. lipolytica	6,434	2,312			

Figure 4. Cladogram of the nine hemiascomycetous yeasts. The cladogram is based on the phylogenetic tree in Souciet et al.[1] The first two columns refer to the number of genes. (**A**) Repartition of the 48,889 genes in the different species. (**B**) Repartition in the different species of the 29,279 genes classified in the 4,107 subsets of orthologs (Ortho1 and Orthox) by the IONS method. The last three columns show the number of subsets obtained by the different methods. (**C**) Subsets of orthologs produced by the IONS method; (**D**) pillars of the YGOB; and (**E**) orthogroups produced by SYNERGY for the 12,691 genes involved in the comparison.

Another 1,142 families of orthologs with a maximum of one member in each species were confirmed by shared neighborhood, while 1,268 new subsets of orthologs of this type were created (Table 1). In total, the number of genes classified into subsets of 2 to 9 orthologs (Table 2) increased from 47.71% (22,841 genes, Table 1), as inferred from the families based on similarity, to 61.16% (29,279 genes; see Tables 1 and 2 for details) with the IONS method.

Interests

The subsets of orthologs (Ortho1 and Orthox) produced by the IONS method have already proven to be of particular interest to serve as anchor points for the construction of syntenic blocks[1] and to identify orthologs.[17] Small differences between the results presented in [1] and in this paper result from an improvement in the IONS method by the inclusion of the sequential procedure that progressively decreases the neighborhood by one neighbor at a time rather than the rough iteration made for 15, 10, 5, and 1 neighbors used in the previous version (see additional file 4 for the correspondence between the two classifications). This modification allows identification of more subsets of the Ortho1 type.

The method, especially the Undetermined subsets, is also useful as a starting point for manual dissection of a functional family. For example, SONS was used to suggest a model for the evolution of the hexose

Table 1. Comparison of the 7,927 Génolevures families (similarity) to the subsets produced by the IONS method.

Families		IONS subsets			
		Identical to families	Not identical to families		
			Ortho1	Orthox	Others
One gene per species	**1,689**	**1,309**	–	**397**	**413**
present in 9 species	*15,201*	*11,781*		*3,007*	*413*
One gene per species present	**1,416**	**1,142**	–	**158**	**412**
in fewer than 9 species	*7,640*	*6,439*		*789*	*412*
Others	**4,822***	**3,800***	**388**	**713**	**2,624**
	25,033	*7,554*	*3,492*	*3,771*	*10,216*
Total	**7,927**	**6,251**	**388**	**1,268**	**3,449**
	47,874	*25,774*	*3,492*	*7,567*	*11,041*

Notes: The families that are fully confirmed by the neighborhood analysis are shown in the second column while columns three to five show the subdivision into different types of subsets (Ortho1, Orthox, and Others) of the families that could not be fully confirmed by the neighborhood analysis. Numbers of subsets are in bold. Numbers of genes are in italics. *Among the 4,822 families in "Others," there are 3,343 families containing only one gene in one species. The subset in IONS is identical.

transporters and glucose sensors.[20] The IONS method was also used for analysis of the evolution of the ATP-binding cassette transporters conferring multiple drug resistance in hemiascomycetous yeasts[19] and of the drug: H^+ antiporter 1 family.[25]

Comparison to other datasets of orthologs

The pillars of the YGOB and the orthogroups produced by SYNERGY

To assess the quality of subsets of orthologs (Ortho1 and Orthox) produced by the IONS method, we compared it to two different studies. The first comparison was done with a manually curated database that we assume to be the 'gold standard' for yeast genomes: the pillars of YGOB (v5, January 2011).[26] The second comparison concerned the orthogroups produced by SYNERGY,[27] an algorithm that reports orthology relationships using sequence similarity, synteny, and a given species phylogeny to reconstruct the underlying evolutionary history of genes. This method is automated but, in contrast to IONS, requires an *a priori* weighting of different parameters: protein similarity (α), synteny similarity (β), and probability of duplication and losses when rooting a gene tree (γ). Other methods exist but either have not been applied to the Hemiascomycetes phylum or do not make use of the synteny evidence.

Both the methods (manual vs automated, types of evidence) and results (subsets of different types of homologs) differ. Indeed, the pillars of the YGOB contain groups of orthologous genes that are allowed to contain one ohnolog in each post-WGD species. In contrast, the orthogroups produced by SYNERGY consist of sets of genes from extant species that are descended from a single gene in the species' last common ancestor,[27] which means that they contain orthologs as well as all in-paralogs produced since the most ancestral speciation event of the species studied.

Table 2. Types of subsets produced by the IONS method for the nine Génolevures species.

Type of subset	Number of subsets	Number of genes	Percentage of genes	Mean subset size*	Median subset size	Mean number of species*
Ortho1	1,697	15,273	31.9	9 ± 0.00	9	9 ± 0.00
Orthox	2,410	14,006	29.26	5.81 ± 2.32	7	5.81 ± 2.32
Subtotal	**4,107**	**29,279**	**61.16**			
TSing	6,041	6,041	12.6	1 ± 0.00	1	1 ± 0.00
InPara	272	886	1.85	3.26 ± 3.70	2	1 ± 0.00
Undet	936	11,668	24.37	12.47 ± 7.79	10	7.99 ± 1.78
Total	**11,356**	**47,874**	**100**			

Note: *Mean and standard deviation.

Among the 34,709 genes of the six species studied in Wapinski et al,[27] 17,210 (50%) were classified into groups of orthologs in a one to maximum one relationship (Additional file 5: Comparison of gene classification in different types of subsets by IONS, YGOB, and SYNERGY). In comparison, the manual curation of the YGOB database allowed classification of 42,046 (69%) of the 60,876 genes of the 11 species of the YGOB v5. Of the 47,874 genes of the 9 species studied in our test case, our automatic method IONS classified 29,279 (61%) into groups of orthologs (Ortho1 and Orthox) that can serve as anchor points for syntenic studies.

The difference in percentages between YGOB and IONS can be mainly explained by the evolutionary distance between the species included in our study. Indeed, if we removed the two most ancestral species that are not present in the YGOB database (*D. hansenii* and *Y. lipolytica*), reducing our set of species to a subset of the species included in the YGOB database, the percentage of genes classified by IONS into groups of orthologs increased to 67%. Moreover, it has been shown that for short evolutionary distances, significant sequence divergence occurs before extensive rearrangements of chromosomes. At larger evolutionary distances, however, the number of chromosome rearrangements rises while protein-sequence divergence becomes limited by saturation and functional constraints (see Fig. 5 in[1]).

Comparison of a common dataset

A more precise comparison implies a focus on the same species and on the same subset of genes within these species. This comparison was restricted to genes belonging to four species common to the three studies (Table 3) for which we had no conflict with correspondence of names and that were classified into subsets of orthologs according to IONS (Ortho1 and Orthox).

The comparisons were done on 12,691 genes using the Rand index.[28] This index determines the similarity between two partitions as a function of positive and negative agreements based on the contingency table of the pairwise assignments of data items. The Rand index ranges from 0 to 1. The ARI[21] introduces a statistical normalization to yield values close to zero for random partitions.[29] A value of 1 indicates a perfect identity between the partitions. The Adjusted Rand Index, ARI, (Table 4) was very close to one (0.977–0.996) for the comparison with the pillars of the YGOB, indicating that the orthology assignments were almost exactly identical. The results of the IONS method differed a bit more from the SYNERGY orthogroups (ARI ranging from 0.895 to 0.913). The comparisons of the pillars of the YGOB to the orthogroups produced by SYNERGY also showed more divergent results (the ARI varied from 0.916 to 0.922, see additional file 6: ARI for the comparison between YGOB pillars and SYNERGY orthogroups). These small discrepancies with the SYNERGY orthogroups may be explained by the fact that the YGOB and IONS methods are essentially based on synteny, in contrast to SYNERGY for which synteny is only one of three parameters weighted *a priori* in an automatic assignment. Case studies[19] tend to show that the IONS and YGOB methods are slightly more efficient in identifying orthology relationships in cumbersome contexts. The main advantage of the SYNERGY method vs. YGOB is the automation of the assignment. Nevertheless, the SYNERGY method is based on three parameters (α, β, and γ) to be determined beforehand: weight protein similarity, synteny similarity, and probability of duplication and losses when rooting a gene tree. Moreover, the weighting of components is in part heuristic and will depend on the set of species considered. In this context, the IONS method may be a better option because it is both automated and based on single constant internal parameters: the width of the initial neighborhood and the synteny constraint (number of genes in common in the neighborhood).

Table 3. Comparison of the classification of the 12,691 genes in subsets of orthologs by the IONS method, in the YGOB database, and by SYNERGY.

	IONS (SONS)	YGOB (Pillars)	SYNERGY (Orthogroups)
Number of groups	3,761	3,675	3,639
Mean number of genes per group	3.374	3.453	3.487
Standard deviation	0.899	0.764	1.274
Mean number of genes by species per group	1.000	1.005	1.040

Table 4. Adjusted Rand index of the SONS to the YGOB pillars and SYNERGY orthogroups.

	SACE	CAGL	KLLA	ERGO
SACE		0.9768	0.9810	0.9822
CAGL	0.8954		0.9824	0.9824
KLLA	0.9042	0.9058		0.9961
ERGO	0.9019	0.9030	0.9134	

Notes: ARI with YGOB in upper diagonal and ARI with SYNERGY in lower diagonal. Analysis is restricted to Ortho1 and Orthox.

A subset by subset comparison with YGOB is supplied in additional file 7: Discrepancies between the IONS subsets and the YGOB.

Implementation

The IONS program that finds orthologs subsets according to the method described above was written in Perl. The required input is a csv file (Additional file 1: Sample of input and output files) relating to genomes and that contains, in each line, the Coding DNA Sequences (CDS) name, the species abbreviation, the chromosome letter, the relative position, the family name, and the strand (this last information is optional). The program is available on the mini website: http://web.me.com/philogene/IONS-method/IONS_2011.html.

Discussion

Relevance of the method

The IONS method subdivides precompiled sets of homologs (based on sequence similarity) using gene neighborhood in an iterative process that gradually decreases neighborhood size until a series of homologs with only one gene per species is obtained. The results are equivalent to those of the YGOB[26] or SYNERGY,[27] but the IONS method has several advantages:

1. The method is automated, in contrast to the YGOB method, which requires a time-consuming manual curation for each new genome.
2. There is no dataset-dependent parameter. While the parameters of SYNERGY must be redefined according to a new dataset, which can lead to contradictions between orthologs found in actual and subsequent results, the IONS method is applicable without reconfiguration. The addition of new species will not change the composition of extant groups of Ortho1 and Orthox; it will offer only the opportunity to complement them or to identify new groups.
3. The method is applicable to any predetermined families of homologs and versatile enough either to use with any existing package to define families of homologs or to use an existing database of families as an input. Another option would have been to develop a full package integrating both the delineation of families and the search for orthologs. The limitation of this option is the impossibility of taking advantage of the new development in the definition of families and the difficulty of using pre-determined classifications such as eggNOG[30] on which our method was tested, giving results similar to the Génolevures dataset. Some standard methods of family determination are proposed on the website.
4. The algorithm is based on a conservative approach that favors the most stringent criteria and minimizes the number of false positives.

An originality of the method is to allow some flexibility in parameterization, such as the neighborhood size and synteny constraint.

Neighborhood size

The initial number of neighbors considered on each side of the query gene was arbitrarily set to 15 based on current knowledge of the size of Hemiascomycetes syntenic blocks.[18] This choice also seems to be suitable for novel yeast species because the distribution of mean syntenic blocks size ranges between 14 and 26 genes.[1]

Note that this initial number of neighbors is not critical because an originality of the method is that the process is iterative. Evolutionary mechanisms are not the same in different parts of a genome, so we could not expect that a standard neighborhood size would be appropriate for all gene families. The sequential procedure is a way to circumvent this limitation. The size of the neighborhood used to fix a SONS may vary from subset to subset. In some cases, a subset is defined using 15 neighbors on both sides of the query gene; in other cases, the iterative process leads to the subdivision of the initial subset into smaller units using fewer neighbors (the criteria to stop the subdivision are described in the "Sequential procedure" section).

Synteny constraint

The fact that only one neighbor has to be in common to assign two query genes to the same cluster may

seem a rather low requirement. The algorithm allows modification of this criterion (using more neighbors in common), which may slightly decrease the false-positive rate (7 genes of 12,691 in our comparison with YGOB as the gold standard). However, this increase strongly decreases the number of subsets of orthologs with one gene in each species because extensive chromosome rearrangements may occur for the most evolutionarily distant species. In our test case on Hemiascomycetes, this number decreased from 1,697 identified subsets with a criterion of one neighbor to 948 with a requirement of two neighbors in common and to 428 with three neighbors. The IONS method also will benefit from the intensification of sequencing efforts because the method was designed to integrate new data quickly.

Evolutionary span

The rationale of the method is that the considered evolutionary span of the analyzed species is short enough to retain information on sequence similarity and neighborhood. Otherwise, new species are required. Indeed, at large evolutionary distances, while protein-sequence divergence becomes limited by saturation and functional constraints, extensive chromosome rearrangements may occur,[1] shuffling the traces of co-ancestrality. Any method of orthology detection must confront this limitation. The only solution is to diminish the evolutionary distance between genomes by filling the evolutionary gaps with newly sequenced genomes. The constant diminution of the cost of sequencing will certainly contribute to this objective and, as already mentioned, our method easily accommodates new species.

The IONS procedure reaches its maximum efficiency when families of homologs used as inputs are accurately and comprehensively calculated. If families are not accurate—for example, if a gene product is not present in a family—the current version of our program will not be able to find an ortholog that was placed in a wrong family. Because the method is conservative, a lack of information will never lead to wrong results but will decrease the number of identified SONS.

Using high coverage genome sequences allows avoidance of the problem of false gene losses.[31] High coverage also limits the probability of genome assembly errors that could lead to cases of false negatives in which orthology is not detected between two genes because of a lack of shared neighborhood. The probability of false positives resulting from assembly errors is close to zero because it would require a consistent misalignment of two regions in two different species.

The method was designed to yield a single final result, but the record of intermediate steps allows further analysis. For example, when the phylogenetic history is complicated by a whole genome duplication generating ohnologs, it is possible to easily identify the two SONS corresponding to the same set of ohnologs.

Perspectives

Species that are phylogenetically distant may present considerable sequence and synteny divergence, which makes it difficult to detect similarity at the nucleotide level and thus to classify gene products accurately into families. The addition of new species belonging to the same phylum will probably reduce sequence divergence, allowing a better classification of genes into families and improving results. The quality of the mapping, sequencing, and identification of the coding regions of these new species is crucial: comprehensive identification of genes and of their location relative to each other, as well as an accurate classification into families of homologs, is required to take advantage of the IONS method.

Conclusions

The identification of orthologs is a major issue in comparative genomics. The combination of both similarity and neighborhood evidence facilitates the identification of orthologs. The IONS method was developed using Hemiascomycetes genome sequences carried out by the Génolevures Consortium. The performance of IONS is comparable to that of more labor-intensive methods such as YGOB. The automatic nature of the procedure paves the way for easy application to new genomes.

Acknowledgements

We would like to thank the Génolevures Consortium, coordinated by Jean-Luc Souciet, for access to the database of protein sequences from *Zygosaccharomyces rouxii*, *Kluyveromyces thermotolerans*, *Saccharomyces kluyveri*, and *Eremothecium gossypii* as well as Julie Diffels for her help in the development of the IONS method. We also thank Thomas Rolland and André Goffeau for helpful discussions, Laurence Jassogne for help in language corrections and two anonymous reviewers for helpful comments. With the

support of the Fonds Special de Recherche (FSR) de l'Université de Louvain.

Disclosures

Author(s) have provided signed confirmations to the publisher of their compliance with all applicable legal and ethical obligations in respect to declaration of conflicts of interest, funding, authorship and contributorship, and compliance with ethical requirements in respect to treatment of human and animal test subjects. If this article contains identifiable human subject(s) author(s) were required to supply signed patient consent prior to publication. Author(s) have confirmed that the published article is unique and not under consideration nor published by any other publication and that they have consent to reproduce any copyrighted material. The peer reviewers declared no conflicts of interest.

References

1. Souciet JL, Dujon B, Gaillardin C, et al. Comparative genomics of protoploid Saccharomycetaceae. *Genome Res*. 2009:1696–709.
2. Jackson AP, Gamble JA, Yeomans T, et al. Comparative genomics of the fungal pathogens Candida dubliniensis and Candida albicans. *Genome Res*. 2009;19:2231–44.
3. Wolfe KH. Comparative genomics and genome evolution in yeasts. *Phil Trans R Soc B*. 2006;361:403–12.
4. Dujon B. Hemiascomycetous yeasts at the forefront of comparative genomics. *Curr Opin Genet Dev*. 2005;15:614–20.
5. Dujon B, Sherman D, Fischer G, et al. Genome evolution in yeasts. *Nature*. 2004;430:35–44.
6. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*. 2003;423:241–54.
7. Rubin GM, Yandell MD, Wortman JR, et al. Comparative genomics of the eukaryotes. *Science*. 2000;287:2204–15.
8. Wapinski I, Pfeffer A, Friedman N, Regev A. Natural history and evolutionary principles of gene duplication in fungi. *Nature*. 2007;449:54–61.
9. Fitch WM. Distinguishing homologous from analogous proteins. *Syst Biol*. 1970;19:99–113.
10. Koonin EV. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*. 2005;39:309–38.
11. Hulsen T, Huynen M, de Vlieg J, Groenen P. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol*. 2006;7:R31.
12. Kuzniar A, van Ham RCHJ, Pongor S, Leunissen JAM. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet*. 2008;24:539–51.
13. Salichos L, Rokas A. Evaluating ortholog prediction algorithms in a yeast model clade. *PLoS One*. 2011;6:e18755.
14. Alexeyenko A, Tamas I, Liu G, Sonnhammer ELL. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*. 2006;22:E9–15.
15. Altenhoff AM, Dessimoz C. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol*. 2009;5:e1000262.
16. Chen F, Mackey AJ, Vermunt JK, Roos DS. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One*. 2007;2:e383.
17. Rolland T, Neuveglise C, Sacerdot C, Dujon B. Insertion of horizontally transferred genes within conserved syntenic regions of yeast genomes. *PLoS One*. 2009;4:e6515.
18. Fischer G, Rocha EPC, Brunet F, Vergassola M, Dujon B. Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. *PLoS Genet*. 2006;2:e32.
19. Seret ML, Diffels JF, Goffeau A, Baret PV. Combined phylogeny and neighborhood analysis of the evolution of the ABC transporters conferring multiple drug resistance in hemiascomycete yeasts. *BMC Genomics*. 2009;10:459.
20. Palma M, Seret ML, Baret PV. Combined phylogenetic and neighbourhood analysis of the hexose transporters and glucose sensors in yeasts. *Fems Yeast Res*. 2009;9:526–34.
21. Hubert L, Arabie P. Comparing partitions. *J Classif*. 1985;2:193–218.
22. Nikolski M, Sherman DJ. Family relationships: Should consensus reign?—consensus clustering for protein families. *Bioinformatics*. 2007;23:e71–6.
23. Sonnhammer ELL, Koonin EV. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet*. 2002;18:619–20.
24. Wolfe K. Robustness—it's not where you think it is. *Nature Genet*. 2000;25:3–4.
25. Dias PJ, Seret ML, Goffeau A, Correia IS, Baret PV. Evolution of the 12-Spanner Drug: H+ Antiporter DHA1 Family in Hemiascomycetous Yeasts. *OMICS*. 2010;14(6):701–10.
26. Byrne KP, Wolfe KH. The yeast gene order browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res*. 2005;15:1456–61.
27. Wapinski I, Pfeffer A, Friedman N, Regev A. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*. 2007;23:I549–58.
28. Rand WM. Objective criteria for the evaluation of clustering methods. *J Am Stat Asso*. 1971;66:846–50.
29. Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. *Bioinformatics*. 2005;21:3201–12.
30. Muller J, Szklarczyk D, Julien P, et al. eggNOG v2.0: Extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res*. 2010;38:D190–5.
31. Milinkovitch MC, Helaers R, Depiereux E, Tzika AC, Gabaldon T. 2x genomes—depth does matter. *Genome Biol*. 2010;11(2):R16.