The Quest for a General and Reliable Fungal DNA Barcode

Vincent Robert[*,1], Szaniszlo Szöke[1], Ursula Eberhardt[1], Gianluigi Cardinali[2], Wieland Meyer[3], Keith A. Seifert[4], C. Andre Lévesque[4] and Chris T. Lewis[4]

[1]*CBS-KNAW Fungal Biodiversity Centre, Utrecht, The Netherlands*

[2]*Dipartimento Biologia Applicata- Microbiologia, Università degli Studi di Perugia, Perugia, Italy*

[3]*Molecular Mycology Research Laboratory, CIDM, Westmead Millennium Institute, SEIB, Sydney Medical School - Westmead Hospital, The University of Sydney, Sydney, Australia*

[4]*Biodiversity (Mycology & Botany), Eastern Cereal and Oilseed Research Centre, Agriculture & Agri-Food Canada, Ottawa, Canada*

Abstract: DNA sequences are key elements for both identification and classification of living organisms. Mainly for historical reasons, a limited number of genes are currently used for this purpose. From a mathematical point of view, any DNA segment, at any location, even outside of coding regions and even if they do not align, could be used as long as PCR primers could be designed to amplify them. This paper describes two methods to search genomic data for the most efficient DNA segments that can be used for identification and classification.

Keywords: Genome, molecular, sequences, barcoding, identification, classification, fungi.

1. INTRODUCTION

Since the early days of classification, taxonomists have struggled with the available information and characteristics of their organisms of interest to develop systems that reflect the true phylogeny as closely as possible. Morphological, physiological or chemical characters were used with some success. It is only in recent years, with the development of molecular methods, that we have a better understanding of the evolutionary relationships among species. It was in the morphologically reduced and thus taxonomically difficult groups, such as bacteria or yeasts, where morphology and even physiology were obviously no longer sufficient to distinguish species, that sequencing was first used in identification and phylogenetic analyses. In the early days, because of time constraints, financial issues and technical difficulties, only single gene phylogenies were generated. The choice of the ideal gene to sequence was based on four major criteria: its presence in all organisms to be studied, ease of PCR amplification and sequencing, it's supposed evolution rate, and the absence of pseudogenes, paralogs or orthologs that could complicate amplification and analysis. Choices were never really based on objective and measurable criteria. Ribosomal genes were exploited for many phylogenetic studies, because before the invention of PCR, ribosomes could be isolated and short sequences determined using chemical methodology. The small and large subunits, as well as the Internal Transcribed Spacers (ITS 1 and 2) regions of the rDNA gene cluster were widely used. Mitochondrial genes, like *Cox1* (also widely known as CO1), have also been used by some phylogeneticists.

*Address correspondence to this author at the CBS-KNAW Fungal Biodiversity Centre, Utrecht, The Netherlands; Tel: +31 (0)30 25 12 097; Fax: +32 478 28 57 64; E-mail: v.robert@cbs.knaw.nl

When first large molecular phylogenetic studies were completed, it was obvious that many clades were poorly supported statistically when only one or two genes were used. Recently, several authors explored possibilities for analyzing several genes to obtain the true phylogeny [1-7]. Some [1] suggest that a few genes (5 to 20) randomly selected could be sufficient to obtain the "true phylogeny". Of course, the ideal solution would be to use only one gene for phylogenetic studies, DNA barcoding and identification.

The possibility to sequence the full genome of several organisms introduced additional options to address questions related to evolution or function. In this article, we describe approaches to make an objective choice for the best genes and to define the "ideal" number of genes to be sequenced for a group of interest.

The first approach to find useful genes is the Ideal *Locus* Method (ILM). Such a *locus* would provide a phylogeny as close as possible to the whole genome phylogeny and would distinguish distantly and closely related species.

From our results obtained with ILM, it was clear that finding very good *loci* was possible but that PCR amplification would be extremely problematic to implement. Finding primers that would work across phylogenetically diverse groups, and even within some closely related taxonomic groups, was not a trivial task.

The second approach, called Best Pair of Primers Method (BPPM), was to identify short conserved regions that could be used as forward and reverse primers for any fungi, or a selected taxonomic group of fungi. The variable regions between these two primers were analyzed and subsequently the ability of the amplified regions to be used as reliable phylogenetic representatives and/or as potential barcode candidates for identification was assessed.

1874-1363/11

Both approaches have the same aims, namely to:

1. Define methods to adequately select the minimum number of *loci* needed to generate reliable phylogeny;
2. Find out which genes are the most "suitable" ones in terms of fitting with full genome phyologeny and amenability to PCR amplification;
3. Determine how those best genes perform compared to the currently used genes;
4. Determine if the best genes are "usable" in the laboratory for barcoding or phylogenetic studies.

The Ideal *Locus* Method must also answer the following two questions:

1. Can primers be found that amplify the targeted regions easily for a larger set of species than used in the current study?
2. Is the whole stretch of the gene needed or can only the most informative regions be used?

The results obtained using the two approaches have important implications not only for phylogenetic studies but also for DNA barcoding and identification. There is no doubt that an adequate and reliable selection of genes will have a tremendous impact on the accuracy and the speed at which identifications can be performed.

2. THE QUEST FOR THE IDEAL LOCUS

2.1. Material and Methods

2.1.1. Principles of the method

In this first approach, in contrast to the next method, protein rather than DNA sequences were used. The reason for that is that the alignment of DNA sequences is much more complex than the alignment of proteins.

All genes are compared with the genes of all others species included in the analysis, searching for similar gene segments. The search is performed in forward and reverse-complement directions. The obtained gene sequences where then translated into protein sequences and all the segments found were then aligned and organized into coding regions. All the genes found are assigned to euKaryote Orthologous Groups (KOGs) of proteins, as shown in Fig. (**1**). For the search of the presence-absence of a cluster of orthologous proteins, the method developed for prokaryote was used as described by Snel *et al.* [8]. As many of these genes are not found in all species, it is necessary to filter them out and keep only those found in all species of interest.

At this point one or several copies of the same KOG and for each species/genome were obtained. Another selection step has been introduced to reduce the number of copies to a single copy by keeping the copy that maximizes the

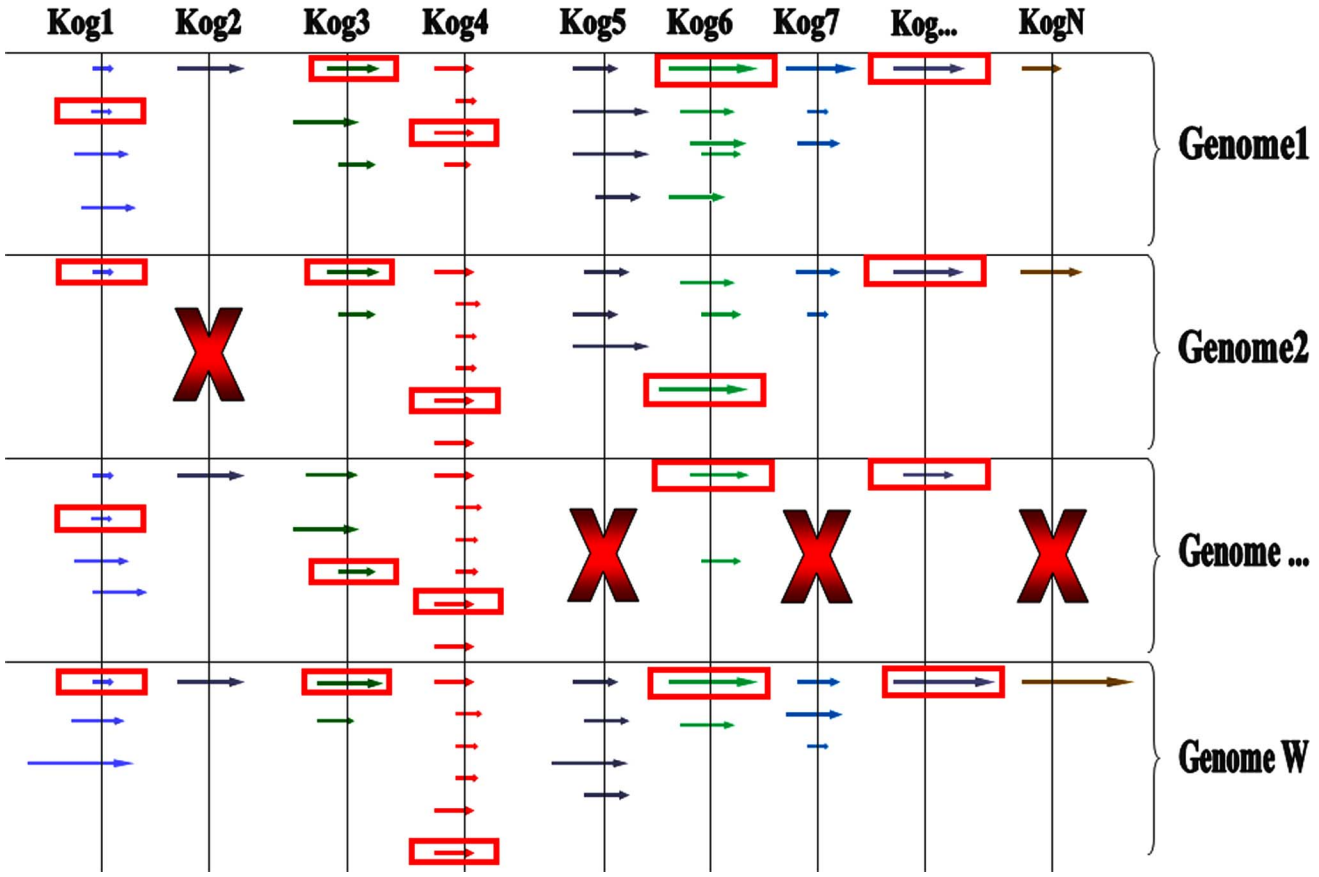

Fig. (1). KOG selection and filtering. KOGs that are not present in all genomes are not retained for the complete analyses (in the example, Kog2, Kog5, Kog7 and KogN are not accounted since they are absent in at least one of the genomes studied). Only one copy per KOG and per genome is kept (see red rectangle around the selected arrows) for further analyses.

similarity with the other copies of the same KOG in the other analyzed genomes.

In the next step, a multiple alignment (using ClustalX [9]) for each KOG was produced using one KOG copy per genome. The resulting multiple alignments could have been transformed into phylogenetic trees using one of the numerous tree reconstruction algorithms (see below). Then, trees could have been compared with each other. However, this route was not followed since the steps between the multiple alignments and the tree reconstruction could introduce possible distortion and unnecessary complication. Instead, the multiple alignments were transformed into distance matrices using the Kimura algorithm [10].

A reference matrix (i.e. the "true matrix") was created by concatenating all multiple alignments of the different KOGs to build a large multiple alignments representing the KOGs present in all of the selected genomes. A distance matrix and a number of phylogenetic trees were then obtained on the basis of the concatenated multiple alignments.

The obtained phylogenetic trees and distance matrix were then considered as references to analyze single gene phylogenies or matrices. All single gene distance matrices were compared to the reference concatenated matrix (RM) *via* the Pearson correlation algorithm (Mantel test) and ranked according to how well they fit with the ideal or reference phylogeny. From there, the best possible genes for phylogenetic analyzes on the group of interest were selected, from which potentially barcoding candidates were then selected.

To rank the *loci* according to their ability to reproduce the "ideal" phylogeny and to decide how many of the *loci* should be used and in which order a so called "gravity center" method [11] was used. The *locus* multiple alignment represented by its distance matrix minimizing the distance with all the remaining distance matrices was selected as the best gene and was considered to be the "gravity center" (GC) of the system. The next best one was the nearest neighbor of the GC. The second best was slightly further away and so on. Similar results were obtained with other methods not described here.

2.1.2. Genomes Analyzed

Twenty five genomes were used for the ILM study. They are listed in Table **1**.

Table 1. Genome Sources, Genome Size (Mb), Number and Percentage of KOGs Used in the Study

Genome	Strain	Number of KOGs	Genome Size (mb)	% KOG Used in this Study	Location
Arabidopsis thaliana		3286	125	18.3	NCBI
Aspergillus nidulans	FGSC A4	2982	31	20.15	Whitehead
Caenorhabditis elegans		4235	100	14.19	Sanger
Candida albicans	SC5314	2636	15	22.8	Stanford
*Candida glabrata (*belongs to teleomorphic genus *Nakaseomyces)*	CBS138	2505	1	24	Genolevures
Cryptococcus neoformans	JEC21	2856	24	21.02	TIGR
Debaryomyces hansenii	CBS767	2760	12-13	21.78	Genolevures
Drosophila melanogaster		4352	120	13.81	NCBI
Eremothecium (Ashbya) gossypii	ATCC 10895	2592	9.2	23.3	EBI
Fusarium graminearum	PH-1 (NRRL 31084)	3063	36	19.62	Whitehead
Homo sapiens		4597	3200	13.07	NCBI
Kluyveromyces lactis	CLIB210	2596	11.4	23.15	Genolevures
Lachancea (Saccharomyces) kluyveri	NRRLY-12651	1747	10.2	30.4	Stanford
Magnaporthe grisea	70-15	2917	40	20.6	hitehead
Naumovia (Saccharomyces) castellii	NRRLY-12630	2390	10.2	25.15	Stanford
Neurospora crassa	N-150	2962	40	20.29	Whitehead
Phanerochaete chrysosporium	RP78	2945	29.9-30	20.41	JGI
Saccharomyces bayanus	MCYC623	2560	12	23.48	Stanford
Saccharomyces kudriavzevii	IFO 1802	1855	10.6	32.4	Stanford
Saccharomyces mikatae	IFO1815	2557	12	23.5	Stanford
Saccharomyces paradoxus	NRRLY-17217	2592	12	23.19	Stanford
Saccharomyces cerevisiae	*S288C*	2668	13	22.53	Stanford
Schizosaccharomyces pombe	Urs Leupold 972 h-	2762	14	21.76	Sanger
Ustilago maydis	521	2850	20	21.09	Whitehead
Yarrowia lipolytica	CLIB99	2699	20-21	22.27	Genolevures

2.1.3. Tree Reconstruction

The concatenated alignment has been analyzed using maximum parsimony, neighbor joining, quartet puzzling using maximum likelihood and Bayesian inference using Markov chain Monte Carlo. Maximum parsimony and neighbor joining analyses have been done using PROTPARS (heuristic search with characters equally weighted) and PROTDIST (Kimura formula) from Phylip [12], respectively. Nonparametric bootstrap support for maximum parsimony and neighbor joining has been calculated from 100 re-sampling rounds. Maximum likelihood trees have been constructed with the TREE-PUZZLE program [13] using the Whelan & Goldman [14] model of amino acid substitution and 1000 puzzling rounds. For Bayesian inference, we have used MrBayes 3.0b4 [15], with four incrementally heated simultaneous Monte Carlo Markov chains over 50 000 generations using random starting trees and a Poisson model of amino acid substitution. Trees have been sampled every 10 generations, resulting in an overall sampling of 5000 trees, of which the first 2000 have been discarded. The remaining 3000 trees have been used to estimate posterior probabilities (i.e. probabilities that groups of taxa are monophyletic, given the data) by computation of a 50% majority-rule consensus tree. Branch lengths have been averaged over the sampled trees, again discarding the first 2000 trees. Stationarity of the process has been controlled using Tracer software, version 1.0 [16]. The Bayesian Markov chain Monte Carlo phylogenetic analysis has been repeated, again using random starting trees, to test the independence of the results from topological priors.

2.2. Results and Discussions on the Ideal *Locus* Method

The "ideal tree" or super tree for the 25 genomes studied based on 531 *loci* is presented in Fig. (**2**). As can be seen, some species belonging to the same genus are not clustering

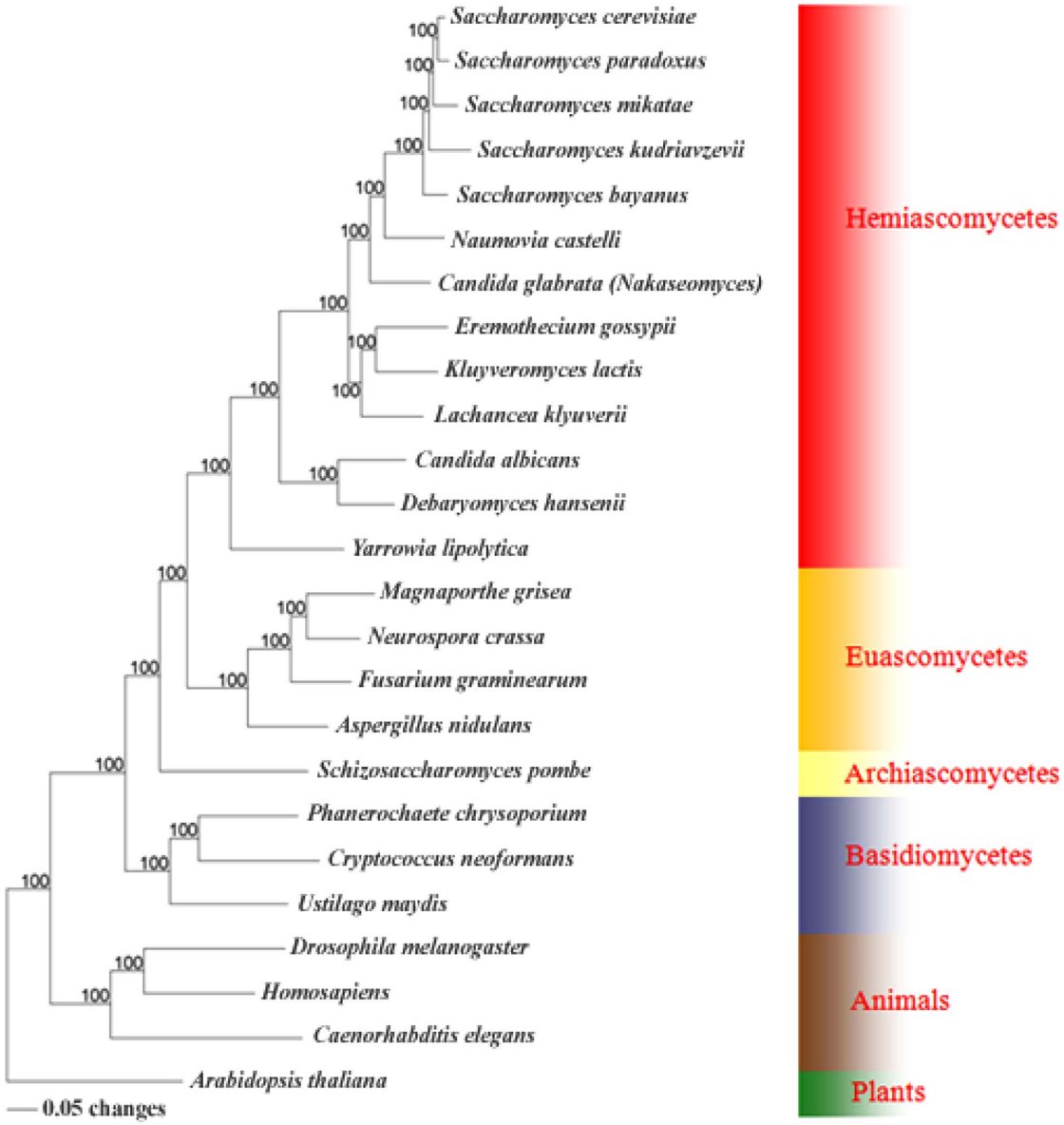

Fig. (2). Phylogenetic super tree based on 531 concatenated proteins present in all 25 genomes studied. The numbers to the left are branch support values (maximum likelihood quartet puzzling support values/posterior probabilities from Bayesian inference). Branch lengths have been estimated using Bayesian inference.

together. Similar observations were made previously and these issues have been addressed by Kurtzman [2] and Kuramae *et al.* [17, 18]. The distance (super) matrix underlying the multiple alignment that was used to reconstruct this super tree, was used as a reference and correlated (Pearson coefficient) with all the individual KOG's distance matrices. Correlation results are summarized in Fig. (**3**) and Table **2**. One third of the genes (29.8%) produce a phylogeny that is highly correlated with the super matrix. Seventy percent of the gene's matrices have a correlation higher than 0.70 with the super matrix. Only a few genes (25) show no or a very low correlation with the reference phylogeny. Neither the length (Pearson's coefficient of correlation = 0.28) nor the evolutionary rates (Pearson's coefficient of correlation = 0.018) of the KOGs are related to or could explain the level of correlation with the super matrix.

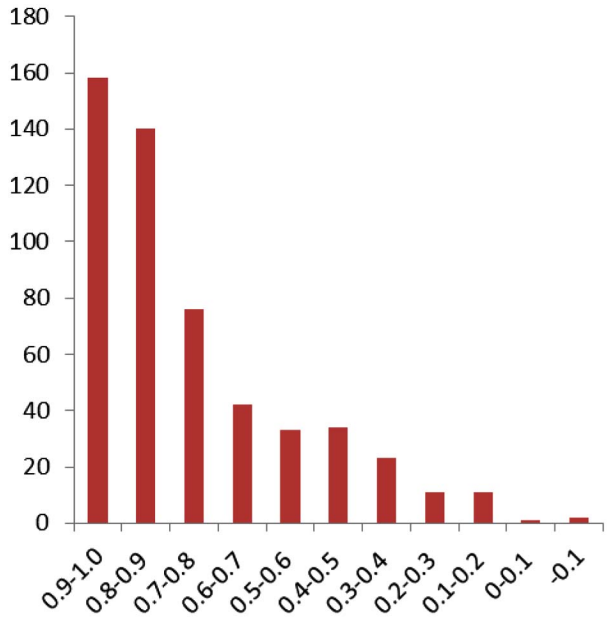

Fig. (3). Histogram presenting the number of KOGs present in each category of correlation level with the distance matrix underlying the super tree.

The *loci* showing the highest level of correlation with the super matrix are presented in Table **3**. Their predicted functions, obtained from the NCBI website, are also provided. KOG 1234 has a correlation of 0.986 with the super matrix (i.e. the reference phylogeny). A large number of *loci* could be used to reconstruct a robust phylogeny from very diverse taxa.

The type, the order and the number of *loci* that should be used to produce reliable phylogenies, depends on the taxa to be classified. However, the high correlation levels obtained by some *loci* indicate that one gene would be highly reliable if well selected. Though, with only one gene and with a high number of taxa, some taxa would be incorrectly placed on the tree because the correlation is not perfect and the number of outliers would increase with sample size. On the herein studied dataset of 531 KOGs and 25 genomes the evolution of the average Pearson correlation levels was computed when comparing all single KOG distance matrices and the distance matrices from an increasing number of the concatenated KOG alignments (see Fig. **4**). Depending on the desired level of correlation with the ideal phylogeny, one could decide to use one, ten or more genes for a given group of organism. In the herein studied dataset, the maximum correlation level found was reached with 190 concatenated KOGs. As expected, using the fungal subset of the dataset of the 25 genomes analyzed more than twice more KOGs were found that are common to all species and showed even higher correlation levels with the super matrix (i.e. the reference phylogeny).

Table 2. The Number of KOGs Present in Each Category of Correlation Level with the Distance Matrix Underlying the Super Tree

Correlation	# of Kogs	% of Kogs
0.9-1.0	158	29.76
0.8-0.9	140	26.37
0.7-0.8	76	14.31
0.6-0.7	42	7.91
0.5-0.6	34	6.21
0.4-0.5	36	6.40
0.3-0.4	24	4.33
0.2-0.3	11	2.07
0.1-0.2	11	2.07
0-0.1	1	0.19
-0.1	2	0.38
Total	**531**	**100**

With the results shown in this study, it seems obvious that a relatively large number of *loci* could be used alone or in combination to reproduce almost perfectly the reference phylogeny. This can be achieved, without having to sequence and analyze the whole genome of the organisms of interest. However, looking at the practical aspects of the results obtained here, we were not able to find good/universal primers (forward and reverse) to easily amplify the KOGs of interest from a large range of different organisms. Multiple alignments of protein coding genes are much easier to handle and probably more relevant and reliable as source of deep phylogenetic information than their DNA counterparts. However, when going back to the DNA multiple sequence alignments of the most interesting KOGs to identify potential primer sequences, it was impossible to find any conserved regions that could be used to amplify the *loci* of interest from a wide spectrum of organisms. It must be noted that degenerate primer design solutions were not tested. We believe this to be a fundamental difficulty despite the fact that for practical reasons not all possible *loci* for all combinations of species were tested. The lack of conserved primer target regions is a serious issue especially in view to a potential DNA barcoding application. Another limitation of the "Ideal *Locus* Method" is that it is only using the protein coding *loci* whereas non-coding regions could be of interest, especially for identification or DNA barcoding. To overcome these issues other approaches were investigated as described below.

Table 3. List of the First 20 KOGs Showing the Highest Correlation with the Super Matrix and Therefore Being the Best Candidate Genes for Phylogeny Reconstruction and Potentially for Barcoding

Ranking	KOG Number	Correlation level	Length	Predicted Function (Source: ftp://ftp.ncbi.nih.gov/pub/COG/KOG/kog)
1	KOG 1234	0.986674	435	ABC (ATP binding cassette) 1 protein
2	KOG 0724	0.981528	435	Zuotin and related molecular chaperones (DnaJ superfamily), contains DNA-binding domains
3	KOG 2472	0.98067	587	Phenylalanyl-tRNA synthetase beta subunit
4	KOG 0714	0.98045	309	Molecular chaperone (DnaJ superfamily)
5	KOG 2002	0.978686	513	TPR-containing nuclear phosphoprotein that regulates K(+) uptake
6	KOG 3844	0.977597	366	Predicted component of NuA3 histone acetyltransferase complex
7	KOG 2369	0.976687	449	Lecithin:cholesterol acyltransferase (LCAT)/Acyl-ceramide synthase
8	KOG 0363	0.976442	523	Chaperonin complex component, TCP-1 beta subunit (CCT2)
9	KOG 0362	0.976349	522	Chaperonin complex component, TCP-1 theta subunit (CCT8)
10	KOG 1450	0.976006	382	Predicted Rho GTPase-activating protein
11	KOG 1439	0.975307	444	RAB proteins geranylgeranyltransferase component A (RAB escort protein)
12	KOG 1113	0.974711	248	cAMP-dependent protein kinase types I and II, regulatory subunit
13	KOG 1156	0.97431	422	N-terminal acetyltransferase
14	KOG 0863	0.973433	218	20S proteasome, regulatory subunit alpha type PSMA1/PRE5
15	KOG 1533	0.973073	277	Predicted GTPase
16	KOG 0340	0.973036	421	ATP-dependent RNA helicase
17	KOG 1273	0.972888	295	WD40 repeat protein
18	KOG 1980	0.972571	546	Uncharacterized conserved protein
19	KOG 1068	0.971993	222	Exosomal 3'-5' exoribonuclease complex, subunit Rrp41 and related exoribonucleases
20	KOG 1485	0.971992	201	Mitochondrial Fe2+ transporter MMT1 and related transporters (cation diffusion facilitator superfamily)

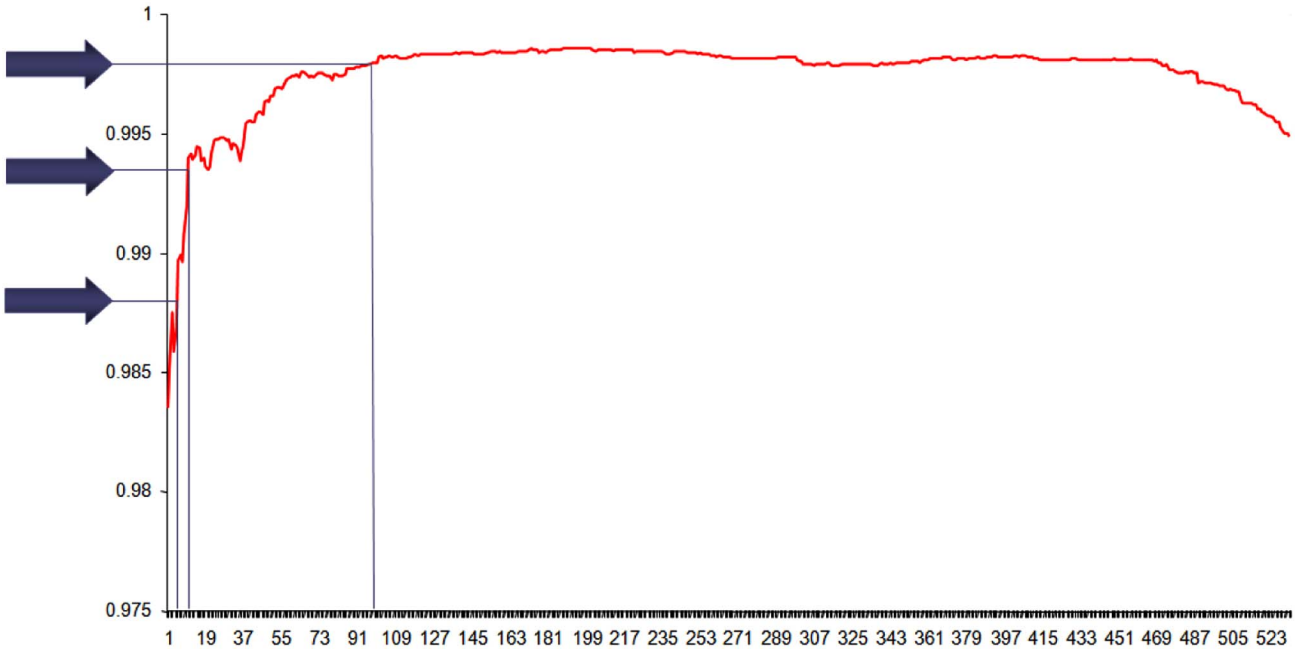

Fig. (4). Evolution of average Pearson correlation levels when comparing all single KOG distance matrices and the distance matrices from an increasing number of the concatenated KOG alignments.

3. SEARCH FOR BEST PAIR OF PRIMERS

3.1. Material and Method

3.1.1. Principles of the Method

This method, called Best Pair of Primers Method (BPPM), was developed to bypass some of the limitations encountered with the KOGs method. Instead of concentrating on the phylogenetic content of the genes like in the ILM, the major aim of the BPPM approach was to find good primer pairs, relatively well conserved for a large group of species (all fungal species in our case study), to amplify highly variable regions. This second approach is therefore more pragmatic than the ILM. Instead of focusing on genes, it focuses on the primers.

The following constrains have been chosen:

1. The whole genome must be used, including DNA without known function, as well as non-coding DNA.
2. The best DNA segments must be easily detected and amplified using relatively well conserved primers.

In a preliminary attempt to find conserved primer pairs, conserved regions longer than 16 nucleotides were searched. A number of algorithms to achieve such a task were developed; the most promising have been optimized to ensure the highest possible speed.

In practice, all sequences are compared pairwise and all segments of at least 24 nucleotides are saved in a table. It is easy to understand that the shorter the minimal size of the segment, the highest the number of records written in that table. For example, searching for all segments of 12 nucleotides found in at least two species would result in many millions of records saved. This is why primers of 24 conserved nucleotides seems to be ideal and the search focused therefore on such length. Although the searching task is extremely computer-intensive, it could be processed on a basic computer in a couple of weeks on a dataset of 52 genomes. The major disadvantage of using such a long primer size is that only 31 genomes (out of 52) were found to have common primer pairs. Even worse, the most common high quality primer pair was present in only 18 of the studied species. When trying the latter method on shorter primers two types of problems were encountered. The number of possible primers increased dramatically and the time needed to search for all primer pairs increased as a square of the number of primers. So it is clearly unrealistic to apply the method to primers much smaller than 24 characters in length. In addition, the most important consideration in primer design is to have a very well conserved section on the 3' end of the primer while the 5' end can show some variability without necessarily hampering the amplification ability of the primer. Hence, an alternative method called the "Short Primer Method" was developed.

In the "Short Primer Method" the following six steps were executed:

1. Assembling a list of all possible primers having a fixed length of 12 nucleotides. This size was selected on the basis of practical and technical considerations. The number of primers is 4^{12} = 16777216. This amount can be kept in a computer memory easily. Using 13 or 14 nucleotides remains feasible on common computers. Over 16 nucleotides, a cluster of computers may become necessary. For example, for 16 nucleotides, the number of possible primers is 4^{16} ≈ 4 billion. A 64 bits operating system become mandatory an managing 4 billion records in a database can be extremely slow if a single hard drive is used.
2. Searching for the presence of these primers in all genomes. This search is much faster than searching for existing primer pairs because no DNA comparison is made. Each DNA record is simply scanned species by species and for each species a single column containing how many times that 12-mers have been found, is saved. The final result is a table with 16777216 rows and one column per species.
3. Keeping primers that have a good quality (see below) and also being present in most species. This operation is straightforward as we already know how many times each primer is found in each species.
4. Using these best primers and search for possible primer pairs.
5. Using the best primer pairs to extract the intermediate DNA sequence. The target length of the intermediate sequence was set to 200 to 1000 nucleotides. These numbers were chosen to ensure a minimum variability and to allow for ease of amplification and sequencing.
6. Compare the amplified DNA of the different species with each other and build distance matrices and trees for comparison with a reference matrix and tree. This last step is to ensure that the amplified DNA would produce a relatively coherent phylogeny but would also be able to discriminate closely related species.

3.1.2. Genomes Analyzed

At the time of this part of the study, seventy seven fungal genomes were available. They are listed in Table **4**. Of the 77 species, 74 have complete DNA sequences. *Podospora anserine, Melampsora laricis-populina* and *Trichoderma reesei* were excluded because only protein data are available. Genome data were stored in 3 tables containing 3 different levels of information: genes, contigs and supercontigs. The 3 different levels have been used and combined to extract the desired information.

3.2. Results and Discussions on the Best Pair of Primers Method (BPPM)

3.2.1. Building the List of Possible Primers

A length of 12 nucleotides was chosen, simply because the total number of primers, given by the following equation, 4^{12} = 16777216, gives a reasonable amount of records. In practice, with the actual computer technology, a maximum value of 14 nucleotides is possible. The length of 12 nucleotides is an interesting value from a practical point of view since all primers with a length greater than 12 are also detected. Therefore, it is possible to construct longer degenerate primers with a core of 12 conserved nucleotides at the 3' end.

Table 4. Genomes Used in the Study

Genome	Strain	Location
Alternaria brassicicola	*?*	*Genome portal*
Aspergillus clavatus	NRRL 1	*Broad Institute*
Aspergillus flavus	AAIH01000000	*Broad Institute*
Aspergillus fumigatus	*Af293*	*Broad Institute*
Aspergillus nidulans	*FGSC A4*	*Broad Institute*
Aspergillus niger	*CBS 513.88*	*Broad Institute*
Aspergillus oryzae	AP007150-AP007177 (DDJB)	*Broad Institute*
Asperguillus terreus,	*NIH2624*	*Broad Institute*
Batrachochytrium dendrobatidis	*?*	*Broad Institute*
Blastomyces dermatitidis	*SLH14081*	
Botrytis cinerea	*?*	*Broad Institute / Syngeta*
Candida albicans	*SC5314*	*Broad Institute*
Candida glabrata	*CBS 138*	*Genolevure*
Candida guilliermondii	*?*	*Broad Institute*
Candida lusitaniae	*?*	*Broad Institute*
Candida parapsilosis	*CDC 317*	*Broad Institute / Trust Sanger*
Candida tropicalis	*?*	*Broad Institute*
Chaetomium globosum	*CBS 148.51*	*Broad Institute*
Coccidioides immitis	*RS*	*Broad Institute*
Coccidioides posadii	*RMSCC 3488*	*Broad Institute*
Cochliobolus heterostrophus	*?*	*Genome portal*
Coprinus cinereus	*Okayama7#130*	*Broad Institute*
Cryphonectria parasitica	*?*	*Genome portal*
Cryptococcus neoformans	*B-3501A*	*Broad Institute*
Cryptococcus neoformans var. grubii	*H99*	*Broad Institute*
Debaryomyces hansenii	*CBS 767*	*Broad Institute*
Fusarium graminearum	*?*	*Broad Institute*
Fusarium oxysporum	*?*	*Broad Institute*
Fusarium verticillioides	*?*	*Broad Institute*
Histoplasma capsulatum	*Nam 1*	*Broad Institute*
Hyaloperonospora parasitica	*Emoy2*	*Genome portal*
Kluveromyces thermotolerans	*CBS6340*	*Genolevure*
Laccaria bicolor	*?*	*Genome portal*
Lachancea (Saccharomyces) kluyveri	*?*	*Genome portal*
Lodderomyces elongisporus	*?*	*Broad Institute*
Magnaporthe grisea	*70-15*	*Broad Institute*
Melampsora laricis-populina	98AG31	*Genome portal*
Microsporum canis	*CBS 113480*	*Broad Institute*
Mycosphaerella fijiensis	*?*	*Genome portal*
Mycosphaerella graminicola	*?*	*Genome portal*
Nectria haematococca	*?*	*Genome portal*

(Table 4) contd…..

Genome	Strain	Location
Neosartorya fischeri	*NRRL 181*	*Broad Institute*
Neurospora crassa	*OR74A*	*Broad Institute*
Neurospora discreta	FGSC 8579 mat A	*Genome Portal*
Neurospora tetrasperma	FGSC 2508 mat A	*Genome Portal*
Paracoccidioides brasiliensis	*?*	*Broad Institute*
Phanerochaete chrysosporium	*?*	*Genome Portal*
Phycomyces blakesleeanus	*?*	*Genome Portal*
Phytophthora infestans	*?*	*Broad Institute*
Phytophthora ramorum	*?*	*Genome Portal*
Phytophthora sojae	*?*	*Genome Portal*
Pichia stipitis	*CBS 6054*	*Genome Portal*
Pneumocystis carinii	*?*	*Pneumocystis* Genome Project
Podospora anserina	S mat+	CNRS and Genoscope
Postia placenta	*MAD-698*	*Genome Portal*
Puccinia graminis	*?*	*(Broad Institute)*
Pyrenophora tritici-repentis	*?*	*Broad Institute*
Rhizopus oryzae	*RA 99-880*	*Broad Institute*
Saccharomyces cerevisiae	*RM 11-1A*	*Broad Institute*
Schizophyllum commune	*?*	*Genome Portal*
Schizosaccharomyces japonicus	*yFS275*	*Broad Institute*
Schizosaccharomyces octosporus	*yFS286*	*Broad Institute*
Schizosaccharomyces pombe	*972h*	*NCBI*
Sclerotinia sclerotiorum	*?*	*Broad Institute*
Sporobolomyces roseus	*?*	*Genome Portal*
Stagonospora nodorum	*?*	*Broad Institute*
Trichoderma atroviride	ATCC 74058, IMI 206040	*Genome Portal*
Trichoderma reesei	*?*	*Genome Portal*
Trichoderma virens	*Gv29-8*	*Genome Portal*
Trichophyton equinum	*CBS 127.97*	*Broad Institute*
Uncinocarpus reesii	*?*	*Broad Institute*
Ustilago maydis	*?*	*Broad Institute*
Verticillium albo-atrum	*VaMs.102*	*Broad Institute*
Verticillium dahliae	*VdLs.17*	*Broad Institute*
Yarrowia lipolytica	*CLIB122*	*NCBI*
Zygosaccharomyces rouxii	*CBS 732*	*Genolevures*

In this step, a table was filled in with these 16777216 primers. For mathematical reasons, a unique key has been assigned to each primer. Each nucleotide is encoded with 2 bits: a = 0, c = 1, g = 2, t = 3. A simple 32 bits key allows encoding at most 16 nucleotides. A simple loop can generate all primer keys and sequences easily. Most computation and comparisons use the 32 bit keys instead of DNA strings. If a primer of more than 16 nucleotides was chosen, the key would had to be stored in a 64 bit integer.

3.2.2. Search for Primers in All Genomes

For each species, one column was added to the primer table, to store the number of times a given primer was found. Primer quality and the number of species where the given primer was found were also stored in the database.

For each species, a map of primers was allocated, with the key of the map being the primer sequence key, the value of the map being the number of occurrences.

Table 5. Primer Table

Sequence Key	Sequence	Quality	Species No	Species 1	Species 2	Species 3
0	aaaaaaaaaaaa	0	72	0	34145	7293
1	aaaaaaaaaaac	0	73	0	173	368
2	aaaaaaaaaaag	0	74	1	199	626
3	aaaaaaaaaaat	0	72	0	322	223
4	aaaaaaaaaaca	0	72	0	73	205
...	...	...	...	...	...	...
16777211	ttttttttttgt	0	72	0	60	165
16777212	tttttttttta	0	73	1	116	384
16777213	ttttttttttc	0	73	0	105	409
16777214	ttttttttttg	0	73	0	107	426
16777215	tttttttttt	0	72	0	998	7295

Of the 16777216 records possible in the results table, only the first 5 and last 5 records are displayed. For example, the first primer "aaaaaaaaaaa" has been found in 72 genomes, never in Species 1, 34145 times in Species 2 and 7293 times in Species 3. It must be noted that Species 1 is *Cryptococcus neoformans*, one of the few genomes that are incompletely published.

The DNA records were loaded, by taking the first 12 nucleotides, generating the corresponding 32 bit key, and keeping the result in the map above : ++mPrimers[key];

The 12-nuclotides floating window was then shifted right by one nuclotide, the key was re-computed and saved in the map. The same operation was repeted up to the end of the DNA sequence.

This procedure was applied to all the DNA sequences and to their reverse-complement DNA sequences. The advantage of generating all 16777216 records in advance in the database is that the primary key of the table, which is the DNA key, is build only once, allowing very fast record search during the operation. In memory, the key is identical to the index of the record in the record table.

At the end of the procedure, the primers map was saved into the database. The beginning and the end of the 16777216 records dataset are given in Table **5**. Column "Species No" gives the number of species where the sequence was found. Column "Species 1" represents the number of copies of the sequences in the genome of species 1.

3.2.3. Primer Quality

One can immediately notice that a primer like "aaaaaaaaaaa" is very common and can be found more than a million times in some species. Such primers were directly removed from the analysis to avoid useless computations. A primer quality ranking algorithm was also designed to filter primers that were not suitable for DNA amplification.

The quality of a primer is difficult to determine. What was needed was a fast and simple algorithm capable of rejecting very poor quality primers. Three bad-quality factors were computed:

- F1 = the number of the most frequent nucleotide / size of the sequence. This factor is close to one for all sequences displayed in Table **5**, and close to 0.25 for a random sequence.
- F2 = Comparison of the sequence with itself with an offset = 3 nucleotides. This factor is very high for sequences containing a long list of nucleotides triplets like the following one:

```
actgaccaccaccaccaccaccaccgt
:::::::::::::::::::::::::::
actgaccaccaccaccaccaccaccgt
```

- F3 = Comparison of the sequence with its reverse complement, with an offset greater than half the sequence size. This is in fact the detection of hairpins.

All these factors take a value of 0.25 for a random sequence, and a value of 1.0 for a very bad primer. As such the factors F2 and F3 have been rescaled with the following equation: F = (F - 0.25) / 0.75.

The final quality is given by:

Quality = 1.0 - max (F1, F2, F3)

3.2.4. Keeping Good Primers

After all DNA sequences had been analyzed, a simple query was applied to compute the number of occurrences of each primer in the database, as shown in column "Species No" (Table **6**).

At this step, the primer results were analyzed and only the most interesting ones were kept. For practical purposes, it was decided to keep primers with a quality above 0.5 and present in 73 species as shown in Table **6**.

This was used to decide how to filter the primers before proceeding to the next step, which is the search of possible primer pairs. Such search is not possible using millions of primers because the amount of possible primer pairs increases as a power of the number of primers used in the pairs. Therefore, for practical purposes, it was decided to only keep primers with a quality score of 0.5 or higher.

Regarding the number of species in which the primer must be present, it became clear that taking 74 (which means

present in all species) was a too strict, since having just one incomplete DNA sequence data set for one species might lead to the elimination of good primers. Once again, the choice was quite subjective but since the main aim of the study was to find primers that are as universal as possible, the limit was set to 73 species. So the number of primers used for the next steps is 204472.

Table 6. Number of Primers of Different Quality Levels

Species No	All Quality	Quality >= 0.5	Quality >= 0.75
74	1408	832	0
73	285205	204472	890
72	696666	538084	3194
71	1194830	962861	7560
70	1759956	1458126	13728
69	2376201	2007830	22331
68	3041505	2608803	32526
67	3743013	3250098	44544
66	4475936	3925482	57743
65	5233228	4629363	72256

A total of 1408 primers of any quality were found in 74 genomes while only 832 primers have a quality between 0.5 and 0.75. No primer was present in all genomes with a quality score above 0.75.
Table **6** provides the number of primers found in at least 74, 73, 72 … 65 species and having a quality of at least 0.0, 0.5, and 0.75. For example, there were 204472 primers with a quality >= 0.5 found in 73 species (the red value in Table **6**).

3.2.5. Searching for Possible Primer Pairs

At this step, the whole genome of each species were scanned a second time, and searched for possible primer pairs. The criterion to find an acceptable primer pair was that the intermediate regions between the two primers should contain between 200 and 1000 nucleotides (see Fig. **5**).

Initially all 204472 primer candidates were loaded and transformed into 32 bit keys in a map. Then all DNA records of all species were looped on and the first 12 nucleotides, transformed as a 32 bit key, were used to search in the map of primers candidates. If no sequence was found, the downstream primer A was then shifted by one nucleotide and searched again, etc. If the sequence was found in the genome sequence, the sequence was called primer A. The next 12 nucleotides were then used as a key and the search was performed again 212 nucleotides downstream to look for the next primer in the primers candidates. If the sequence was found, this was considered as a valid primer pair, and both primers were saved in a list of primer pairs. If not found, primer B was shifted by one nucleotide and searched again up to a maximum distance of 1000 nucleotides between primers A and B. If no pair was found, primer A was shifted to the right by one nucleotide and searched again, etc. During the scanning of all the genomes, all primer pairs found were kept in memory. When all species were covered, all primer pairs that were present in at least 37 species (half of the 73 species chosen in Table **6**) were saved in the database.

A partial result is displayed in Table **7**, with all the primer pairs present in at least 48 species out of the 74 genomes studied.

It must be noted that there were no primer pairs found that were present in all 74 genomes studied. The best pair was found to be present in 52 species but its quality was low. Some primer pairs provided in Table **7** were low quality primers and could not be used for DNA amplification. Some of these primer pairs could amplify more than 10000 different homologous and non-homologous DNA segments in a single species. Additional filtering was therefore necessary to keep valid primers. Only the primer pairs for which the sum of the quality of primers A and B was greater than 1.0, as underlined in Table **7** were kept. From this table, one can notice that some primers were extremely similar and in fact, they can sometimes be combined to form a longer possible primers (see example in Fig. **6**).

For each of 3223 primer pairs kept inter-primer regions were stored as Fasta format in primer pair results file.

For each primer pair results file, the following procedure was followed:

1. All sequences were compared with all others and these sequences were sorted in decreasing average similarity order. In other words, the first one was the most similar to all others and was called the gravity center [11].
2. For each species not already present in the results file, the database was scanned for the sequence most similar to the gravity center sequence. The similarity must be above or equal to 0.5. So, at most, one sequence was added to the file.
3. The gravity center sequence was compared in a pairwise alignment against Genbank to determine if the gravity center sequence had a known function. The 10 most similar genes were added as comment on top of the results file. The similarity must be >= 0.8.

All results files were sorted by function. Most sequences were located in the 5.8 S, 18S, 28S, 26S or 60S regions of the rDNA gene cluster, as shown in Table **8**. As the 5.8S, 18S, 28S, 60S regions of the rDNA gene cluster are already

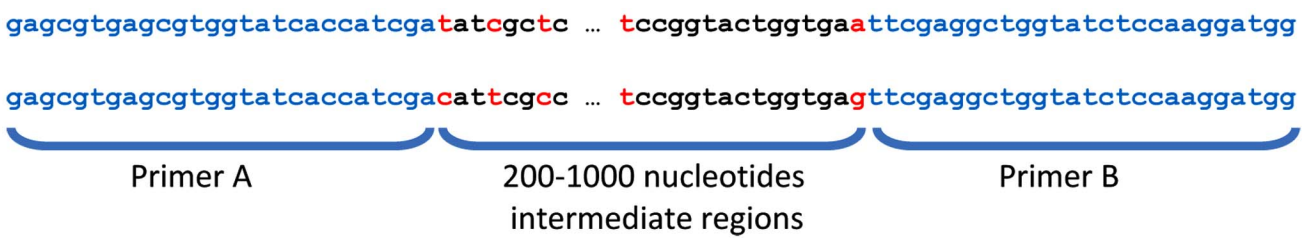

Fig. (5). Search for pairs of primers.

Table 7. Best Primer Pairs, Presence in Species and Associated Combined Quality (Quality Primer A + Quality Primer B). Only Primer Underlined Pairs are Considered of Sufficient Combined Quality

Species No	Primer A	Primer B	Quality A	Quality B
52	ctctctctctct	tctctctctctc	0.5	0.5
52	gagagagagaga	agagagagagag	0.5	0.5
51	agagagagagag	gagagagagaga	0.5	0.5
51	tctctctctctc	ctctctctctct	0.5	0.5
50	catcagacacca	cagatcttggtg	**0.583**	**0.667**
50	gagagagagaga	gagagagagaga	0.5	0.5
50	tctctctctctc	tctctctctctc	0.5	0.5
50	caccaagatctg	tggtgtctgatg	**0.667**	**0.583**
49	catcagacacca	agatcttggtgg	**0.583**	**0.583**
49	attgtactcatt	agacaagcatat	0.5	0.5
49	atatgcttgtct	aatgagtacaat	0.5	0.5
49	gagggcaagtct	gaaattcttgga	**0.583**	**0.667**
49	ggcaagtctggt	gaaattcttgga	**0.583**	**0.667**
49	catatgcttgtc	aatgagtacaat	**0.583**	**0.5**
49	tccaagaatttc	accagacttgcc	**0.667**	**0.583**
49	tccaagaatttc	agacttgccctc	**0.667**	**0.583**
48	ttgtctcaaaga	ctatcaactttc	**0.667**	**0.583**
48	gtcttgaaacac	ctagtagctggt	**0.667**	**0.593**
48	catgcaccacca	gaatttcacctc	**0.5**	**0.667**
48	tgtctcaaagat	ctatcaactttc	**0.667**	**0.583**
48	gaaagttgatag	atctttgagaca	**0.583**	**0.667**
48	gaaagttgatag	tctttgagacaa	**0.583**	**0.667**
48	gaggtgaaattc	tggtggtgcatg	**0.667**	**0.5**
48	ttccagctccaa	gaaattcttgga	**0.583**	**0.667**
48	accagctactag	gtgtttcaagac	**0.593**	**0.667**
48	atatgcttgtct	tggaatgagtac	**0.5**	**0.667**
48	ttcttgacgttg	tccatcttgttg	0.5	0.5
48	gtactcattcca	agacaagcatat	**0.667**	**0.5**
48	caacaagatgga	caacgtcaagaa	0.5	0.5
48	caacaagatgga	tcaacgtcaaga	**0.5**	**0.583**
48	catatgcttgtc	tggaatgagtac	**0.583**	**0.667**
48	tcttgacgttga	tccatcttgttg	**0.583**	**0.5**
48	tccaagaatttc	ttggagctggaa	**0.667**	**0.583**

intensively studied, the "Unknown" group, containing 308 results files, was considered for the rest of the study.

For each results file, corresponding to one primer pair, all sequences were compared with all others and the result matrix was saved. It is clear that some of these primer pairs give very similar results. Grouping these primer pairs so that all pairs of a single group return highly similar or identical DNA sequences was made by computing the covariance matrix between all the similarity matrices. Table **9** gives the 20 first groups that can be made from the primer pairs present in at least 68 of the 74 species studied. These groups also include a number of primer pairs found in less than the 68 species. In each group, there is always a primer pair that is the most similar to all other primer pairs of that group. That is the one displayed in column 2 of Table **9**. Interestingly, most of these primer pairs were also from well-known regions used for phylogeny and molecular taxonomy.

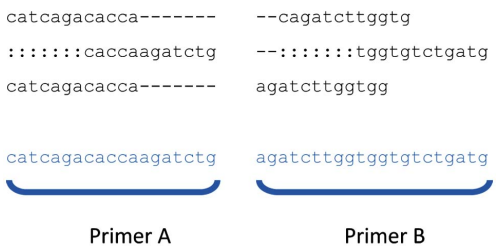

Fig. (6). Combination of a few 12 nucleotide long primers may combine into primers of longer length.

Table 8. Genes Location

Location	File Number
5.8S	10
5.8S - 18S	217
5.8S - 18S - 28S	2198
5.8S - 28S	14
18S	273
26S	9
28S	180
60S	14
Unknown	308
Total :	**3223**

3.3. Comparing amplified DNA

For the best 20 primer pairs, which were all found in at least 68 of the 74 genomes studied, the sequences between primers A and B were used to compute the similarity between all species based on pairwise sequence alignments. The resulting similarities were stored in similarity matrices and a hierarchical clustering tree (UPGMA) was produced to evaluate the relevance of the produced classification (data not shown). Finally, all the trees were manually reviewed and a multiple alignment of the best one was produced (subjectively chosen). The later corresponding the sequences obtained with the primer pair aacaagatggac-ctccaagaacga, was used to produce the (UPGMA) phenetic tree shown in Fig. (**7**).

One problem that occurred is that non-homologous sequences were returned by some of the primer pairs and this is visible in the tree of Fig. (**8**). In this case, primer pairs returned sequences that corresponded to two different (non-homologous) regions in the genome, generating two groups of information, and thus generating two major branches in the tree.

4. WORKING WITH GROUPS OF FUNGI AND PRIMER ANALYSIS FOR EACH TAXONOMIC PHYLUM

The process described above was repeated on a subset of the species (phylum) to:

Table 9. Best Primer Pairs and the Groups they are Belonging to

Group	Primer Pair	Species No	Location	Similarity
1	acaagcgtttct-catcaagttcca	71	?	
2	acatggagaaga-catcaaggagaa	70	actin gene	0.98
3	accttcttgatg-catgttcttgat	72	elongation factor 1-alpha	0.97
4	agtacttgtagg-cttggccttgta	74	60S ribosomal protein L15	0.84
5	ggaacttgatgg-agaaacgcttgt	71	?	
6	ggtatcaccatc-caacaagatgga	71	elongation factor 1-alpha	0.95
7	gtccatcttgtt-gatggtgatacc	70	elongation factor 1-alpha	0.91
8	gtccatcttgtt-tacttgaaggaa	70	elongation factor 1-alpha	1.00
9	gttcttggagtc-gtccatcttgtt	71	elongation factor 1-alpha	0.96
10	aacaagatggac-ctccaagaacga	68	elongation factor 1-alpha	0.99
11	aacaagatggac-tcaccactgaag	68	elongation factor 1-alpha	0.99
12	caacaagatgga-ctccaagaacga	69	elongation factor 1-alpha	1.00
13	cacttcttcatg-atggacgagatg	69	beta-tubulin	0.97
14	catatgcttgtc-gactcgtcatct	69	?	
15	cttcagtggtga-ccatcttgttga	68	elongation factor 1-alpha	1.00
16	gaacttgatggt-agaaacgcttgt	68	?	
17	gttccttcaagt-caacaagatgga	69	elongation factor 1-alpha	1.00
18	tccatcttgttg-acttgaaggaac	69	elongation factor 1-alpha	0.99
19	tcttgacgttga-gtccatcttgtt	68	elongation factor 1-alpha	0.97
20	ttcttgacgttg-gtccatcttgtt	68	elongation factor 1-alpha	1.00

1. Increase the number of candidate primer pairs
2. increase the number of covered species
3. obtain regions that are phylogenetically more informative.

Fig. (7). Phenetic tree (UPGMA) given by sequence group number 10.

Fig. (8). Phenetic tree (UPGMA) - Primer pair returning two groups of DNA information due to the fact that the primer-pair used amplified two different parts with in the genome of the studied species.

The number of primers found for different taxonomic phyla is given in Table **10** for the ascomycetes, basidiomycetes, zygomycetes, oomycetes and chytridiomycetes. The only taxonomic phylum that was studied further in detail was the ascomycetes since our database did not contain enough data for the other groups.

4.1. Ascomycetes

The method described above has been applied to the ascomycetes.

4.1.1. Primer Filtering

Before searching for primer pairs, it was necessary to filter the primers to reduce computation time. Table **10** shows that our database contained 57 ascomycetes with complete genomes. So it was arbitrarily decided to keep all primers of sufficient quality (>= 0.5) and present in at least 56 species.

4.1.2. Primer Pairs Search

From these 739585 primers, all primer pairs were searched in the genome of these 57 species. As before, the DNA segment between the two primers must have a length in the range of 200 to 1000 nucleotides. Primer pairs found in too few species were discarded. All primer pairs found in at least 43 species (75% of the 57 species) were saved. Using these primer pairs, all DNA segments present between the primer pairs were extracted from the database and saved in a DNA file. This step generated 957 DNA files.

4.1.3. Similarity Matrix

For each DNA file, all species were compared with all species and a similarity matrix was generated. The DNA sequences were sorted so that the most similar sequence to all others is located on top of the file.

4.1.4. DNA Identification

Using the first sequence in each file, the gene was identified by pairwise alignments against Genbank. A total of 951 sequences were found in 5.8S, 18S, 26S or 28S regions and 6 sequences were found to belong to the elongation factor alpha-1 region. Not a single sequence of interest was found in an unknown area.

4.1.5. Covariance Matrix

The covariance matrix was computed between all of the 957 matrices. All primer pairs (each represented by its matrix) were organized in groups giving similar results. In other words, all primer pairs of the same group return DNA sequences that give similar results when comparing species. Eleven groups were found. They are given in Table **11**.

Compared to the global analysis, the number of groups was limited. The first group included 158 primer pairs, meaning that many primers overlapped with others and that longer primers were easily generated. For others groups, like group 10, only three primer pairs were found, meaning that increasing the primer size rapidly decreased the number of species covered by these primer pairs. For example, if only three pairs were overlapping each other but for one base, they can be merged all into a 14-mers primer without losing any coverage. To the contrary, if 158 primer pairs have been found in a group, much longer merged primers may be built.

Table 10. Number of Primers Found for Different Taxonomic Phyla

Ascomycetes Primers

Species No	All Quality	Quality >= 0.5	Quality >= 0.75
57	394532	289033	1230
56	952720	739585	4292
55	1602268	1289638	9779
54	2316677	1916687	17753

Basydiomycetes Primers

Species No	All Quality	Quality >= 0.5	Quality >= 0.75
11	0	0	0
10	12795	10813	149
9	6037904	5462077	105124
8	10612145	9695509	196368

Zygomycetes Primers

Species No	All Quality	Quality >= 0.5	Quality >= 0.75
2	11424673	10479663	216074
1	15221608	13999790	311144

Oomycetes primers

Species No	All Quality	Quality >= 0.5	Quality >= 0.75
4	14335633	13131564	281088
3	16376861	15010710	323991

Chytridiomycetes Primers

Species No	All Quality	Quality >= 0.5	Quality >= 0.75
1	11219756	10324479	226664

Computing the average value of a row (or column) in the covariance matrix gives a good idea of how well the corresponding primer pair was representing the group of primers. A value close to 1.0 means that this primer pair gave very similar results to all other primer pairs of the group. If the group was not consistent (i.e. non homologous sequences), some primer pairs had a lower average value.

Some groups were highly homogeneous and all primer pairs of the group extracted very similar DNA segments. For example, group 4 is made of 36 primer pairs giving nearly identical results. The average values row by row in the covariance matrix are in the range 0.966 - 0.979.

Some other groups are much more heterogeneous. For example, groups 3, 5, 6 and 9 seem to contain subgroups

Table 11. Best Primer Pairs Groups

Group	# Primer Pairs in Group	# Species Covered	Primer Pair Examples
1	158	57	aaagttgatagg-aatgagccattc cgaaagttgata-cattcgcagttt
2	62	56	atacaaaccatg-accctactgatg
3	25	56	cttgtctcaaag-attcaaatttct cttgtctcaaag-cattcaaatttc
4	36	56	atttctgcccag-gttgagcttgac gtcaagctcaac-ctgggcagaaat
5	48	55	aaactgcgaatg-aaggcagcaggc gaatggctcatt-caaattacccaa
6	130	55	aattgttcctcg-agaaatttgaat attcaaatttct-agggcaagtctg caattgttcctc-agaaatttgaat
7	65	54	tgacattcagag-tagagccaatcc agaaatcacatt-tagagccaatcc ggattggctcta-aatgtgatttct
8	6	54	cgatgaagaacg-cttaagcatatc tgatatgcttaa-cgttcttcatcg
9	148	54	ctgccagtagtc-aaggcagcaggc aagattaagcca-acatccaaggaa
10	3	50	aacaagatggac-ggtgactccaag
11	28	50	aggcatttggct-acctgctgcggt tagatgacgagg-ggagacctgctg

The second column gives the number of primer pairs found in the group. The third column gives the number of species covered by this group. A few DNA primer pairs are given as examples in the last column.

extracting different information (i.e. non homologous sequences).

The next step is to test all these primers in the laboratory, starting with primer pairs covering a wide range of species, and being representatives of the groups they belong to. Each group should be tested for DNA amplification, and tested as identification and classification candidate.

CONCLUSIONS AND FUTURE WORK

Finding primer candidate pairs with the potential to amplify DNA for species belonging to completely distantly related groups of organisms and providing high phylogenetic information content seems very difficult with limited computation resources. Reducing the species number or reducing the diversity of the species involved, increase significantly the coverage of the primer pairs. In the global search for all true fungi, only seven primer pairs were covering 72 species or more out of the 74 species studied. For the ascomycetes, six primer pairs were found and compatible with all 57 species studied and 186 primer pairs with 56 species.

The result of the Best Pair of Primers Method is dependent on several subjective choices. A critical one is the number of taxa that should be compatible with any selected primer pair. A cut-off value of 73 species (out of 74) and 56 species (out of 57were used for the global and the ascomycetes respectively. Using a lower cut-off value may significantly increase the size of the primer list used to search for primer pairs, and therefore give more options. Considering all primers is not feasible, as the total number of primer pairs to search for would be 16777216 * 16777216 = 281475 billion primers. Thus, primer list reduction cannot be avoided.

Finding new genome regions of high interest and that are not located in 5.8S, 18S, 26S, 28S regions of the rDNA gene cluster or elongation factor 1-alpha has not been successful so far. More research is necessary to enlarge the primer list and apply the method again. However, EF1 alpha has been suggested as a second barcode region after ITS and our results would support this overall approach. The accompanying paper of Lewis *et al*. [19] proposes an alternative approach that exploits annotated protein families from the Pfam protein families database.

Degenerate primers have not been considered in this study and must certainly be envisaged. One could start from the best 12-nucleotide primers that we have found and search for longer degenerate primers that can then cover a much larger set of species.

The “Ideal *Locus* Method” showed that a number of genes seem to produce reliable phylogenies fitting almost perfectly with complete genome phylogenies. Such *loci* would not only be good for phylogenetic studies but would be potentially excellent barcode regions since they are capable of separating closely related species (see

Saccharomyces genus for example). The only and major problem at this stage is the lack of universal primers that would make them the really ideal loci. New generation sequencing methods might allow us to bypass such problem since the need for conserved primers could become obsolete.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7, 2007-2013), Research Infrastructures action, under the grant agreement No. FP7-228310 (EMbaRC project). It was also supported by the Alfred P. Sloan Foundation project "Barcoding the Indoor Mycota".

CONFLICT OF INTEREST

None declared.

REFERENCES

[1] A. Rokas, B. L. Williams, N. King, and S. B. Carroll, "Genome-scale approaches to resolving incongruence in molecular phylogenies", *Nature*, vol. 425, pp. 798-804, 2003.

[2] C.P. Kurtzman, "Phylogenetic circumscription of Saccharomyces, Kluyveromyces and other members of the Saccharomycetaceae, and the proposal of the new genera Lachancea, Nakaseomyces, Naumovia, Vanderwaltozyma and Zygotorulaspora", *FEMS Yeast Res.,* vol. 4, pp. 233-245, 2003.

[3] C.P. Kurtzman, C.J. Robnett, "Phylogenetic relationships among yeasts of the Saccharomyces complex determined from multigene sequence analyses", *FEMS Yeast Res.,* vol. 3, pp. 417-432, 2003.

[4] V. Daubin, M. Gouy, G. Perriere, "A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history", *Genome. Res.,* vol. 12, pp. 1080-1090, 2002.

[5] F. Lutzoni, F. Kauff, C.J. Cox, *et al.*, "Assembling the fungal tree of life: progress, classification and evolution of subcellular traits", *Am. J. Bot.,* vol. 91, pp. 1446-1480, 2004.

[6] J.W. Taylor, M.C. Fisher, "Fungal multilocus sequence typing - it's not just for bacteria," *Curr. Opin. Microbiol.,* vol. 6, no. 4, pp. 351-356, 2003.

[7] J.R. Dettman, D.J. Jacobson, J.W. Taylor, "A multi*locus* genealogical approach to phylogenetic species recognition in the model eukaryote Neurospora", *Evolution,* vol. 57, no. 12, pp. 2703-2720, 2003.

[8] B. Snel, G. Lehmann, P. Bork, M.A. Huynen, "STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene", *Nucleic Acids Res.,* vol. 28, pp. 3442-3444, 2000.

[9] J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin, D.G. Higgins, "The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools", *Nucleic Acids Res.,* vol. 25, pp. 4876-4882, 1997.

[10] M. Kimura, *The Neutral Theory of Molecular Evolution*, Cambridge Univesity Press, p. 75, 1983.

[11] L. Antonielli, V. Robert, L. Corte, L. Roscini, A. Bagnetti, F. Fatichenti, G. Cardinali, "Searching for related descriptors among different datasets: a new strategy implemented by the R package "Dadi"", *Open Appl. Inform. J.*, vol. 4, pp. 15-27, 2010.

[12] J. Felsenstein, "Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods", *Methods Enzymol.,* vol. 266, pp. 418-427. 1996.

[13] H.A. Schmidt, K. Strimmer, M. Vingron, A. von Haeseler, "TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing", *Bioinformatics,* vol. 18, pp. 502-504, 2002.

[14] S. Whelan, N. Goldman, "A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach", *Mol. Biol. Evol.,* vol. 18, pp. 691-699, 2001.

[15] J.P. Huelsenbeck, F.R. Ronquist, "MrBayes: Bayesian inference of phylogenetic trees", *Bioinformatics,* vol. 17, pp. 754-755, 2001.

[16] A. Rambaut, A. Drummond, *Tracer. MCMC Trace Analysis Tool. 1.0*, University of Oxford, UK, 2003.

[17] E. Kuramae, V. Robert, B. Snel, T. Boekhout, "Conflicting phylogenetic position of Schizosaccharomyces pombe", *Genomics*, vol. 88, no. 4, pp. 387-393, 2006.

[18] E. Kuramae, V. Robert, B. Snel, M. Weiss, T. Boekhout, "Phylogenomics reveal a robust Fungal Tree of Life", *FEMS Yeast Res.*, vol.6, no. 8, pp. 1213-1220, 2006.

[19] C.T. Lewis, S. Bilkhu, V. Robert, U. Eberhardt, S. Szoke, K.A. Seifert, and C.A. Lévesque, "Identification of Fungal DNA Barcode targets and PCR Primers based on Pfam protein families and taxonomic hierarchy", *Open Appl. Inform. J.*, vol. 5, pp. 72-86, 2011.

Received: December 10, 2010 Revised: January 20, 2011 Accepted: April 22, 2011