PCR primer design: primers4clades tutorial

Pablo Vinuesa and Bruno Contreras-Moreira

primers4clades web logo

Welcome to the primers4clades (primers for clades) documentation and tutorial page. This document explains how to use this tool and, more importantly, how to read the results. For a quick start try the tutorials, which are based on the provided demo sequences. They will provide a convenient overview of the server's user interface and capabilities. Feel free to contact us if you come accross bugs or find it difficult to understand the results. We aim to use your feedback to keep this documentation up to date.

Table of contents

Back to table of contents

Overview

This section provides a short description of the server capabilities. Practical examples and the technical details are given below (table of contents).

What can this server do?

primers4clades (primers for clades) is an easy-to-use web server developed for researchers interested in the design of PCR primers for cross-species amplification of novel sequences from metagenomic DNA or from uncharacterized organisms belonging to user-specified phylogenetic clades or taxa. It implements an extended CODEHOP primer design strategy based on the analysis of both DNA and protein multiple sequence alignments of coding sequences. The input for the server is a set of aligned or non-aligned protein coding genes with or without introns (Fig. 1). If introns are present and their coordinates provided, the server excises them and joins the CDSs, translates them with a user-selected translation table and aligns the proteins. The underlying codon structure is computed guided by the protein alignment and NJ trees based on maximum likelihood distances are computed. The user can interactively select clusters of sequences on which to focus the search for oligonucleotides, guided by the phylogeny displayed on screen (Fig. 4). It evaluates a comprehensive set of thermodynamic parameters for each primer pair, as well as the phylogenetic information content of the aligned amplicon sets that would be theoretically amplified by each primer pair for a given input data set. A comprehensive table of the primer thermodynamic properties can be downloaded in a convenient tabular format, along with the aligned amplicon sets and the corresponding maximum likelihood trees. These trees can also be graphically displayed by the server. A non-redundant set of primer formulations is returned, ranked according to a quality criterion based on their thermodynamic properties. Each aligned amplicon set is then evaluated for its phylogenetic information content. This is a very valuable piece of information to take into account when making the choice between alternative primer formulations with equal quality. A graphical overview of the results is provided in the form of an amplicon map. These features greatly aid the user in making an informed choice between alternative, non-redundant primer pair formulations.

primers4clades input page displaying default parameters

Figure 1. The primers4clades "required input" page displaying default parameter values and using the fungal alpha-tubulin demo sequences.

The primers4clades extended CODEHOP algorithm

The server has two run modes. By default the server runs in the basic, non-interactive "get primers" mode, as shown in Fig. 1, performing the computational steps 1-5 listed below, forcing the server to search for primers using all input sequences. A neighbor-joining (NJ) tree is computed from maximum-likelihood (ML) distances (WAG+G for proteins, HKY85 for DNA) and displayed on screen (steps 1-2). The server then executessteps 3-5, resulting first in the display

an amplicon distribution map and secondly

the corresponding list of primer pairs ranked according to their theoretical thermodynamic quality parameters and the phylogenetic information content of each amplicon set. This run mode won't let you select clades or sequences for primer design, but you can select a particular substitution model from a drop-down list in the data input page for the ML tree search (Fig. 1) and evaluation of the phylogenetic information content of the corresponding amplicons. ML trees are calculated for each amplicon (unless the allocated CPU time for the process is surpassed), which can be displayed on screen or downloaded.

The "cluster sequences" run mode is interactive. After steps 1-2 have been computed and the NJ tree is displayed on screen, you can select a particular clade or a list of sequences to focus primer design specifically on this data subset. Even after a first set of primer pairs have been found, they can be further "refined" by exclusion (or inclusion) of further sequences until the user is satisfied with the results.

- **Computational steps in the default, non-interactive "get primers" run mode**

1) Translates the input CDSs to protein using selected translation tables, collapsing redundant sequences into single haplotypes. If the FASTA header contains the coordinates of the exon boundaries of eukaryotic genes, these are spliced and joined before translation. The translation products are finally aligned using muscle, alignment that is used to compute the underlying codon alignment.

2) Reconstructs a NJ distance tree using neighbor from the PHYLIP package, using either the protein or codon alignments, from which WAG+G or HKY85+G distances are estimated under the maximum likelihood (ML) criterion using TREE-PUZZLE. The resulting NJ tree is displayed on screen along with information about the min. and max. pairwise distances found between sequences.

3) The oligonucleotides are designed for the aligned set of input sequences using an extended CODEHOP (COnsensus DEgenerate Hybrid Oligonucleotide Primer) design strategy. The original CODEHOP algorithm [Rose et al. (1998) and Rose et al. (2003)] is based on the identification of highly conserved regions within protein BLOCKS and the use of a particular codon usage tables (CUT) and position specific scoring matrix (PSSM) to derive the CODEHOP formulation. The extended CODEHOP algorithm implemented in primers4clades comprises:

1. The automatic evaluation of a non-redundant set of codon usage tables (nrCUTs) for all organisms recognised in the input file FASTA header, as well as the computation of analignment-specific CUT (see details below).
2. In addition to the CODEHOP formulations derived from the different CUTs, the server computes what we call a corrected CODEHOP in which the degeneracy level is corrected considering the target codon alignment, which is displayed to screen.
3. The server also computes a so-called relaxed corrected CODEHOP which has an extended degenerate region as compared to the corrected CODEHOP in case that the latter has a degeneracy level < 24.
4. A fourth, fully degenerated oligonucleotide formulation is also computed based on the codon alignment.
5. A comprehensive set of thermodynamic parameters is calculated for each oligo. The primers4clades pipeline therefore provides the user with 4 primer formulations, which are displayed on screen, aligned with the corresponding codon multiple sequence alignment.
6. The coordinates of each oligonucleotide in a primer pair are used to cut the "theoretical amplicon sets" out of the original protein and codon alignments. A non-redundant set of primer pairs is computed in such a manner that amplicons dereived from the same codon usage table don't overlap more than 80% of their lengths. The distribution of this non-redundant set of primers and theoretical amplicon sets is mapped on the protein coordinates of the first sequence in the multiple sequence alignment of the input data.
7. Maximum likelihood (ML) trees are then inferred for each aligned amplicon set using PhyML 2.4.5 and a default or user-selected substitution matrix or model. The phylogenetic information content of each amplicon is calculated from the corresponding ML phylogenies by computing the mean and median Shimodaira-Hasegawa-like branch support values of the ML phylogeny, as explained in the technical information section.
8. The CODEHOP pairs and corresponding theoretical amplicons found for the input data set are filtered by the user-provided length range of the amplicons and ranked according to the thermodynamic attributes of the former. An amplicon distribution map is provided as a convenient overview of the coverage of the target locus, which also depicts the thermodynamic quality attributes of each primer pair.
9. The phylogenetic information content of each aligned amplicon set is computed and is displayed on screen along with the four primer formulations aligned with the underlying codon alignment.

- **Computational steps in the interactive "cluster sequences" run mode**

In this run mode the server executes steps 1-2 indicated in the previous section. However, after these steps have been performed, the user can select a particular clade on the displayed NJ tree for the program to search for potential PCR primers considering only that sequence subset. After hitting the get primers button, the pipeline will run through steps 3-5 indicated above. After a first set of primer pairs have been found, they can be further "refined" by exclusion (or inclusion) of further sequences by selecting them on click boxes displayed along with the NJ tree, which will correct the primer formulations accordingly.

What is customizable about the server's behaviour?

The input data has to be a DNA protein-coding sequence that can be translated by any of the current NCBI translation tables. You can select whether you want to work at the DNA or protein levels to estimate a first reference NJ tree for the set, and among a set of DNA and protein substitution models to estimate the maximum likelihood phylogenetic trees used to calculate the phylogenetic information content of the predicted amplicon sets Fig. 1. You can further impose a size range for the desired amplicon lengths that are best suited for your specific needs, interactively select or exclude sequences for primer design and select a suitable Tm for the consensus clamp, as shown in Fig. 1. In the interactive "cluster sequences" run mode, you can first choose to reconstruct a NJ phylogeny based on the translation products or on the corresponding codons and thereafter interactively select a particular clade from the NJ tree to instruct the server to focus the search for oligos specifically on the target group of interest. After a first set of primer pairs have been found, they can be further "refined" by exclusion of divergent or otherwise unwanted sequences by selecting them on click boxes displayed along with the NJ tree, which will correct the primer formulations accordingly.

Back to table of contents

Input data and run mode settings

Input data

To run **primers4clades** in the interactive "cluster sequences" mode you need a **FASTA-formatted** file containing at least 4 CDSs (protein-coding DNA), such as that obtained from the ENTREZ nucleotide service of the NCBI or other sequence databases. The aligned or un-aligned sequences can be pasted into the sequence window or uploaded to the server from your local machine (Fig. 1). The maximum number of sequences you can upload is 30. You can get primer formulations for a minimum of two input sequences, but only running in the interactive, "refine primers" mode, which won't evaluate the phylogenetic information content of the new amplicon subset.

- **Translation tables**

As explained above, the pipeline is primarily based on the CODEHOP strategy to find potential primer binding sites on multiple alignments of protein sequences. It is therefore critical that the input DNA sequences are in frame and that they are translated using the proper translation table (genetic code). The server can use any of the translation tables currently defined by the NCBI.

- **Bracketed taxon names and codon usage tables (CUTs)**

Ideally the taxon names should be enclosed in square brackets, as shown in the example below, which corresponds to intronless bacterial sequences. This will allow the server to use the codon usage tables (CUTs) for all organism names found in the FASTA header of the input file. The position of the bracketed text in the FASTA header is irrelevant. If the taxon names are not bracketted, then a default list of CUTs will be used which contains a diverse array of Bacteria, Archaea, Eukarya and viruses.

```
>640461706 RecA protein [Corynebacterium glutamicum R: NC_009342]
ATGGCTCCCAAGAAGACAGCAACAAAGGCAACTGCCGCCAAGGGGAATGA
TCGTCAGAAGGCACTTGATGCCGCACTAGCCCTGATTGAGAAGGATTTCG
...
>640603512 RecA protein [Mycobacterium tuberculosis H37Ra: NC_009525]
ATGACGCAGACCCCCGATCGGGAAAAGGCGCTCGAGCTGGCAGTGGCCCA
GATCGAGAAGAGTTACGGCAAAGGTTCGGTGATGCGCCTCGGCGACGAGG
...
```

- **Outgroup sequence**

The current version of the primers4clades server will not allow to re-root the NJ trees displayed on screen, which will be arbitrarily rooted using the first sequences in the input file. Therefore, if you know a priori that a certain sequence in the dataset would represent a good outgroup, place it as the first one in the input FASTA file. We suggest that you use at least two outgroup sequences in order to properly define in- and outgroup clades on the reference NJ tree. Furthermore, this can be of great value if you are interested in the design of ingroup-specific primers.

- **Eukaryotic genes containing introns**

For eukaryotic genes containing introns, the exon boundary coordinates should be indicated in the FASTA header between << >> symbols, as indicated below. In this case there are 7 exons.

```
>[Fusarium sp.] <<1..19,114..154,211..237,301..376,428..958,1007..1430>>
GAGGTCATTAGCATCAACGGTAAGCTATTGCTCGCTCCCTTTGTCTACTCTGGCCTTTCACCGCATCTCCTTTTCCGCCAT ...
>[Gibberella zeae] <<1..25,120..160,217..243,307..382,435..965,1016..1439>>
ATGCGTGAGGTCATTAGCATCAACGGTAAGCTATTGCTCGCTCCCTTTGTCTACTCTGGCCTTTCACCGCATCTCCTTTTCC ...
```

If exon boundaries are provided in the header, the server will splice and join the exonic regions into the corresponding CDSs before translating them.

- **Reading frames**

If you are uploading unaligned sequences, then the first nucleotide of each sequence should correspond to the first codon position (i.e. the CDS should be coded by the reading frame +1) so that they can get translated correctly. The primers4clades server will cut trailing nucleotides if the sequence length is not divisible by 3 in order to generate non-truncated CDSs. Furthermore, the server will skip sequences in the input file that contain in frame STOP codons. The server will automatically recognize aligned DNA input sequences.

- **Your email address vs. "browser time out", the stored jobs utility, privacy policy and major update notifications**

We strongly recommend that you provide your email in the input page while submitting jobs to primers4clades, as this allows confidential storage of your data and results pages. By associating an email address to primers4clades jobs, users can make use of the convenient "stored jobs" search facility. Notice also that for longer computations your browser could easily "freeze" due to a browser time out. To avoid loosing results due to this potential browser problem, provide your email! This will ensure that you get the results of your analyses automatically sent to you mail. Your address will be strictly protected and won't be passed to any organization or individuals. The developers of primers4clades would only use it to notify registered users of major upgrades of the system (max. 2 messages per year).

- **Limits: number of sequences and CPU usage time**

The current limits imposed on the size and CPU time to process an input dataset are max. 35 haplotypes (distinct sequences) and max. 7 minutes of ML analysis for the phylogenetic evaluation of amplicon sets.

Back to table of contents

Parameter settings for the two primers4clusters run modes

The primers4clades server has two run modes: **cluster sequences** and **get primers**. The latter is the default mode, as explained below.

- **Then non-interactive "get primers" run mode**

If you want to find primers that would target all sequences in the original alignment, you could use the default "get primers" run mode (Fig.1). This will instruct the server to run the pipeline steps 1-5 in non-interactive mode. Key parameters that can be set by the user to control the behaviour of the default "get primers" run mode are the amplicon length range, primer clamp Tm and the substitution matrix or model to use for protein- or DNA-based evaluation of the phylogenetic information content of the amplicons, as shown in Fig.1. Although several matrices and models can be chosen, the server does not provide a model selection function (Read the FAQ section on How should I decide which substitution matrix or model to use for the evaluation of the phylogenetic information content of the amplicons? for more information on model selection in phylogenetics. See the section on testing the server using the demo sequences below for a detailed description of the output generated by this run mode.

- **The interactive "cluster sequences" run mode**

Alternatively, if you want to guide primer formulation by phylogenetic criteria you should run the interactive "cluster sequences" mode of primers4clades (Fig. 3). This will make the server execute the pipeline steps 1 and 2, which will display a NJ tree with numbered leaves, which can be used to select a specific clade on which to target the primer design. You can choose to reconstruct the NJ tree using either the protein (default) or codon sequences, using the WAG+G or HKY85+G models, respectively. After the target clade or group of sequences has been identified (Fig. 4), you can start the interactive "get primers" run mode, which will allow you to further refine the oligonucleotides found by excluding more sequences (Fig. 7). See the section on customized run case with user-selected parameter values, clades and sequences below for a detailed description of the output generated by this run mode.

Back to table of contents

Tutorials

A brief description of the three demo data sets

Three demo data sets are provided for your convenience to test the interface and functionality of the primers4clades web server:

1. The "alphaproteobacteria atpD" sequence set is intended for those of you who want to get a quick impression of the server's interface, and is explained in the submitted manuscript. The tutorials presented in this document are based on the following two data sets.
2. The "fungal alpha-tubulin" sequences with introns. We will use this data set to describe in detail the server's functionality and output under the default, non-interactive, get primers run mode.
3. The actinobacterial rpoB data set will be used to describe the server's functionality and output run under the interactive, get primers mode, which will allow you to specify a phylogenetic clade or set of sequences on which to target the primer design.

Testing the server using the demo fungal alpha-tubulin sequences and default settings

In order to get familiarized with the server interface and basic functionality, we're first going to use the "fungal alpha tubulin genes with introns" demo sequences which you can upload by hitting the the demo button (Fig. 1). Hitting now the submit button will start the analysis pipeline in the "get primers" run mode, executing the pipeline steps 1-5 in non-interactive mode, using the default settings: 1) protein sequences and a WAG+G maximum likelihood distance matrix to reconstruct the NJ tree that will be displayed to screen based on the whole alignment and dataset; 2) report only primer pairs amplifying fragments between 450 and 950 pb; 3) the consensus clamp of the oligos should have a Tm of ~55 °C and 4) evaluate the phylogenetic information content of the amplicons found using the WAG matrix with gamma correction for among-site rate variation across sites. We recommend that you hit also the check server load button on the two mirror sites (CCG/UNAM , EEAD/CSIC) of primers4clades in order to decide where to submit your job.

- **Output of the non-interactive "get primers" run mode.**

After hitting the submit button (Fig. 1) you will get a page that looks as shown below:

primers4clades job demo_fungal_a-tubulins_with_introns (ecae5167)

```
_gencode : universal
_evaluation : protein_WAGG
_cluster distance metric : protein
_Tm : 55
_amplicon_length : 450 , 950
_email : vinuesa@***
```

Clustering sequences...

```
# read_FASTA_sequence : chopped CDS sequence (3' 1 nts) >[GIBBERELLA ZEAE] STRAIN NRRL ...

# read_FASTA_sequence : skipped CDS sequence (inframe STOP codon) >[ASPERGILLUS ORYZAE] ...

# read_FASTA_sequence : chopped CDS sequence (3' 1 nts) >[FUSARIUM CORTADERIAE] ...

# read_FASTA_sequence : chopped CDS sequence (3' 1 nts) >[FUSARIUM SP.] ...

# read_FASTA_sequence : chopped CDS sequence (3' 1 nts) >[FUSARIUM BRASILICUM] ...


# number of recognised taxa = 9
# aligning translated sequences...
# computing distance matrix...
# multiple alignment FASTA file
# alignment stats: length = 452 %gaps = 1.35 %constant = 77.2

# mean distance = 0.08392
# max distance = 0.20029 ( 007 <=> 008 )
# min distance = 0.00000 ( 001 <=> 011 )


# NJ tree with cluster labels :

NJ tree for fungal tubulins, computed from a ML WAG+G distance matrix
```

Calculating primers, results will be shortly emailed...

You might return or wait for your results to appear here.

```
# largest cluster (0) selected, please use runmode 'cluster sequences' to make your own selection

# number of intron-spanning primers skipped = 24
# number of intron-spanning primers skipped = 27
# number of intron-spanning primers skipped = 24
# number of intron-spanning primers skipped = 25
# number of intron-spanning primers skipped = 24
# number of intron-spanning primers skipped = 24
# number of intron-spanning primers skipped = 23
# number of intron-spanning primers skipped = 24
# number of intron-spanning primers skipped = 24
# number of intron-spanning primers skipped = 24
## table redundancy vs amplicons stats:
# input_derived_ffb7e816 redundancy = 0.87 amplicons = 2
# Botryotinia_fuckeliana redundancy = 0.91 amplicons = 2
# Fusarium_cortaderiae redundancy = 0.95 amplicons = 2
# Fusarium_oxysporum_f_sp_lycopersici redundancy = 0.92 amplicons = 1
# Fusarium_sporotrichioides redundancy = 0.95 amplicons = 1
# Neurospora_crassa redundancy = 0.95 amplicons = 1
# Aspergillus_niger redundancy = 0.95 amplicons = 1

# evaluating primers...
# primer evaluation TAB file


# min quality for phylogenetic evaluations = 50%

# ranking pairs of primers...
```

amplicon distribution map for fungal alpha-tubulins

download

```
## Amplicon 1 codon_usage_table = input_derived_ecae5167

AGTCCAAGCTCGAGTTCtgygtntaycc 5'->3' N 162 411 (aligned residues)
agagcaagctggagttctgtgtttaccc >001 [GIBBERELLA ZEAE] PH-1 STRAIN PH-1
agtccaagctcgagttctgcgtctaccc >002 [NEUROSPORA CRASSA] OR74A <<116..1465>>
aagtaaagttggagttctgtgtctaccc >003 [BOTRYOTINIA FUCKELIANA] B05.10 PARTIAL MRNA
agtgcaagctcgagttctgcgtctaccc >005 [PHAEOSPHAERIA NODORUM] SN15  <<60..1412>>
agtccaagcttgagttctgcgtctaccc >006 [ASPERGILLUS NIDULANS] FGSC A4 CHROMOSOME VIII
agtccaagctcgagttctgcgtctaccc >007 [NEUROSPORA CRASSA] MRNA FOR ALPHA-TUBULIN B <<19..1374>>
agactaaacttgaattctgtgtctatcc >008 [COCCIDIOIDES IMMITIS] (CIMG_06299) MRNA, COMPLETE CDS
agagcaagctggagttctgtgtttaccc >009 [FUSARIUM SP.] NRRL 29298 ALPHA TUBULIN GENE, PARTIAL CDS <<1..19,114..154,211..237,301..376,428..958,1007..1430,1484..1683>> ==19(94),60(56),87(63),163(51),694(48),1118(53),==
agagcaagctggagttctgtgtttaccc >010 [GIBBERELLA ZEAE] ALPHA-TUBULIN GENE, COMPLETE CDS <<1..25,120..160,217..243,307..382,435..965,1016..1439,1493..1718>> ==25(94),66(56),93(63),169(52),700(50),1124(53),==
agagcaagctggagttctgtgtttaccc >011 [GIBBERELLA ZEAE] STRAIN NRRL 29149 ALPHA TUBULIN GENE, PARTIAL CDS <<1..19,114..154,211..237,301..376,429..959,1010..1433,1487..1686>> ==19(94),60(56),87(63),163(52),694(50),1118(53),==
agagcaagctggagttctgtgtttaccc >012 [FUSARIUM BRASILICUM] STRAIN NRRL 31238 ALPHA TUBULIN GENE, PARTIAL CDS <<1..19,114..154,211..237,301..376,429..959,1010..1433,1487..1686>> ==19(94),60(56),87(63),163(52),694(50),1118(53),==
agagcaagctggagttctgtgtttaccc >013 [FUSARIUM CORTADERIAE] STRAIN NRRL 31185 ALPHA TUBULIN GENE, PARTIAL CDS <<1..19,114..154,211..237,301..376,429..959,1010..1433,1487..1686>> ==19(94),60(56),87(63),163(52),694(50),1118(53),==
.????..??.?..?.....?.....
AGTCCAAGCTCGAGTTCtgygtytaycc codeh_corr c91bf176_intronless_cluster_0_amp2_N162
AGTCCAAGCTCgarttctgygtytaycc relax_corr c91bf176_intronless_cluster_0_amp2_N162
ardbhaarytbgarttctgygtytaycc degen_corr c91bf176_intronless_cluster_0_amp2_N162

AGTCCAAGCTCGAGTTCtgygtntaycc 5'->3' N 162 411 (aligned residues)
agagcaagctggagttctgtgtttaccc >001 [GIBBERELLA ZEAE] PH-1 STRAIN PH-1
agtccaagctcgagttctgcgtctaccc >002 [NEUROSPORA CRASSA] OR74A <<116..1465>>
aagtaaagttggagttctgtgtctaccc >003 [BOTRYOTINIA FUCKELIANA] B05.10 PARTIAL MRNA
agtgcaagctcgagttctgcgtctaccc >005 [PHAEOSPHAERIA NODORUM] SN15 <<60..1412>>
agtccaagcttgagttctgcgtctaccc >006 [ASPERGILLUS NIDULANS] FGSC A4 CHROMOSOME VIII
agtccaagctcgagttctgcgtctaccc >007 [NEUROSPORA CRASSA] MRNA FOR ALPHA-TUBULIN B <<19..1374>>
agactaaacttgaattctgtgtctatcc >008 [COCCIDIOIDES IMMITIS] (CIMG_06299) MRNA, COMPLETE CDS
agagcaagctggagttctgtgtttaccc >009 [FUSARIUM SP.] NRRL 29298 ALPHA TUBULIN GENE, PARTIAL CDS ...
agagcaagctggagttctgtgtttaccc >010 [GIBBERELLA ZEAE] ALPHA-TUBULIN GENE, COMPLETE CDS ...
agagcaagctggagttctgtgtttaccc >011 [GIBBERELLA ZEAE] STRAIN NRRL 29149 ALPHA TUBULIN GENE, ...
agagcaagctggagttctgtgtttaccc >012 [FUSARIUM BRASILICUM] STRAIN NRRL 31238 ALPHA TUBULIN GENE, ...
agagcaagctggagttctgtgtttaccc >013 [FUSARIUM CORTADERIAE] STRAIN NRRL 31185 ALPHA TUBULIN GENE, ...
AGTCCAAGCTCGAGTTCtgygtytaycc codeh_corr ecae5167_intronless_cluster_0_amp2_N162

AGTCCAAGCTCgarttctgygtytaycc relax_corr ecae5167_intronless_cluster_0_amp2_N162

ardbhaarytbgarttctgygtytaycc degen_corr ecae5167_intronless_cluster_0_amp2_N162

AGAACTCACCCTCCtccatnccytc 5'->3' C 162 411 (aligned residues)
agaactcaccctcctccataccctc >001 [GIBBERELLA ZEAE] PH-1 STRAIN PH-1
agaactcgccttcctccataccctc >002 [NEUROSPORA CRASSA] OR74A <<116..1465>>
agaattcaccttcttccataccctc >003 [BOTRYOTINIA FUCKELIANA] B05.10 PARTIAL MRNA
agaattcaccctcctccataccctc >005 [PHAEOSPHAERIA NODORUM] SN15 <<60..1412>>
agaattcacctttcctccataccctc >006 [ASPERGILLUS NIDULANS] FGSC A4 CHROMOSOME VIII
agaactcgccttcctccataccctc >007 [NEUROSPORA CRASSA] MRNA FOR ALPHA-TUBULIN B <<19..1374>>
agaattcaccttcctccataccttc >008 [COCCIDIOIDES IMMITIS] (CIMG_06299) MRNA, COMPLETE CDS
agaactcaccctcctccataccctc >009 [FUSARIUM SP.] NRRL 29298 ALPHA TUBULIN GENE, PARTIAL CDS <<1..19,...
agaactcaccctcctccataccctc >010 [GIBBERELLA ZEAE] ALPHA-TUBULIN GENE, COMPLETE CDS <<1..25,120..160,...
agaactcaccctcctccataccctc >011 [GIBBERELLA ZEAE] STRAIN NRRL 29149 ALPHA TUBULIN GENE, PARTIAL CDS ...
agaactcaccctcctccataccctc >012 [FUSARIUM BRASILICUM] STRAIN NRRL 31238 ALPHA TUBULIN GENE, PARTIAL CDS ...
agaactcaccctcctccataccctc >013 [FUSARIUM CORTADERIAE] STRAIN NRRL 31185 ALPHA TUBULIN GENE, PARTIAL CDS ...
....?..?..?..?.....!.....
AGAACTCACCCTCCtccataccytc codeh_corr ecae5167_intronless_cluster_0_amp2_C411
AGAACtcrccytcytccataccytc relax_corr ecae5167_intronless_cluster_0_amp2_C411
agaaytcrccytcytccataccytc degen_corr ecae5167_intronless_cluster_0_amp2_C411


# primer pair quality = 90%
# expected PCR product length (nt) = 750
# fwd: minTm = 61.0 maxTm = 65.4
# rev: minTm = 62.6 maxTm = 64.3
# quality warnings:
# fwd fulldeg (10368 >= 193)
#--
# phylogenetic amplicon evaluation:
# subs.model = WAGG
# alpha = 0.41
# mean aLRT = 0.40 median aLRT = 0.06
# amplicon FASTA file
# aLRT likelihood tree (newick)
# end_of_amplicon


## abbreviations:
# quality = quality estimation [best = 100%, worst = 0]
# crosspot = potential of cross-hybridization [0-1]
# relaxdeg = relaxed (3' extended segment) degeneracy
# fulldeg = full primer degeneracy
# minTm = minimum Tm for the pool of relaxed primers
# maxTm = maximum Tm for the pool of relaxed primers
# hpinpot = potential of primer hairpin [0-1]
# selfpot = potential of primer self-priming [0-1]
# aLRT = approximate likelihood-ratio test [0-1],
# overall confidence on trees built using this amplicon
```

Figure 2A. Output of the non-interactive "get primers" run mode of the primers4clades server for the fungal alpha-tubulin demo sequences using the default settings.

The header section (Fig. 2A) reminds you which runmode and parameter values are being used. If there are partial CDSs of a length that is not divisible by 3, the read_FASTA_sequence subroutine will chop off the nucleotide(s) of the last incomplete codon. This subroutine further processes the FASTA file, skipping sequences that contain in frame STOP codons (e.g. the proper frame was not provided by the user), parsing the taxon names found within brackets, "splicing" and "joining" exons if their boundaries were provided by the user, and finally collapsing sequences to haplotypes (i.e. keeping only unique sequences). In this case all sequences are distinct and are therefore not collapsed. After this processing, the CDSs get translated using the universal codon table (default translation table: 1). Thereafter, the server will start aligning the sequences at the protein level. Basic alignment statistics such as alignment length, percentage gapped and constant sites are provided. The protein alignment is used to guide the codon alignment of the CDSs. The alignment of the input dataset can be downloaded from the server. These steps are skipped if the user uploads previously aligned codon sequences. By default the cluster analysis will be performed at the protein level, and a maximum likelihood WAG+G distance matrix is calculated for this purpose. A NJ tree with labelled nodes is displayed on the screen. Cluster 0 refers to the currently selected cluster, which in this run mode consists of the entire tree, i.e. the cluster containing all the original sequences uploaded to the server (13 in this case, but one was skipped due to an in frame STOP codon).

The calculating primers header provides information about the codon usage tables used to calculate the oligos that have been found, as well as their relative redundancy and number of amplicons obtained for each of them. A detailed explanation about how these codon usage tables and redundancy parameters are calculated is provided in the section on technical information. After computing the set of nrCUTs the server gets into a computing intensive phase locating potential primer binding sites and calculating a comprehensive set of thermodynamic parameters for each of them. Information about the thermodynamic parameters for each oligo can be downloaded in a convenient tabular format from the server hitting the TAB file link. After finishing the computations on the thermodynamic properties of the primer pairs, the server ranks them according to this criterion in descending order (best receive a quality score of 100%), as explained in the section on technical information. This quality information is displayed on the amplicon sets map that summarizes the coverage of the target region by a set of non-redundant primer formulations. Notice that the maximal overlap between any two amplicons is set to 20%.

After displaying the amplicon sets map, the server enters the second computing intensive phase. Maximum likelihood phylogenies (in this default run using the WAG+G model) will be estimated from each of the theoretical amplicon sets that each primer pair would amplify. It is from these phylogenies that the phylogenetic information content of the amplicons is calculated. The best scoring primers are evaluated first. After the server is done with the ML search for the first amplicon set, the corresponding primer pairs are shown on screen, aligned with the underlying codons. The primers header line indicates the codon usage table used to calculate the corresponding primer pair, in this case a data-specific one computed from the input alignment. The first line after the header shows the sequence of the forward CODEHOP formulation, as calculated by the original CODEHOP program, showing also the coordinates of the fw and rev oligos on the original (full) protein alignment. The lower case letters on the 3'-end of the oligo correspond to the degenerated portion of the CODEHOP, while the upper-case residues towards the 5'-end correspond to the consensus clamp. The lines below the CODEHOP formulation are the corresponding target sites at the DNA level. This codon alignment is used to calculate what we call the corrected CODEHOP formulation, which adjusts the degeneracy suggested by the corresponding codon usage table to that required to match the codon alignment. The '!' and '?'characters just above the corrected CODEHOP formulation line indicate the discrepancies between the CODEHOP formulation and the codon alignment that are considered to formulate the corrected CODEHOP sequence, the '!' indicating non-degenerate sites and the '?' highlighting degenerate sites. The '.' symbol indicates matches between the original CODEHOP formulation and the codon alignment. The relaxed corrected CODEHOP formulation and the fully degenerated oligonucleotide formulations correspond to modifications of the corrected CODEHOP sequence in which the region of the latter, containing degenerate sites, is extended towards the 5'-end until reaching a maximum degeneracy level of 24 or an unlimited level, respectively. The CODEHOP design, thermodynamic primer pair quality and amplicon phylogenetic information content parameters are explained in the section on technical information.

After displaying the alignments of the different oligonucleotide formulations computed by the server, a primer pair quality summary report is displayed on screen, based on computations performed on the corrected CODEHOP formulations. The first line provides a summary score of the "primer quality" in thermodynamic terms for each primer pair. This parameter can range in value from 100% to 0% (best to worst). The best quality score indicates that none of the oligos in the pair had a degeneracy level or thermodynamic parameter value worse than the threshold values for each parameter (shown in computing primer quality score). The next three lines indicate the expected amplicon length in bps and the computed Tm ranges for the fw and rev corrected codehop formulations. If the quality score is < 100%, this means that at least one oligo has a thermodynamic parameter value worse than the cut-off values used by the server (shown in Table 2). In our case we see that the primer pair quality score is = 90%, because of a single quality warning, which indicates that the full degeneracy of the fwd primer is > 193, the upper cutoff value for this parameter, as shown in (Table 2).

Finally we get a summary of the phylogenetic evaluation of the corresponding amplicon, which indicates the substitution matrix/model used (WAG matrix plus gamma correction for among-site rate variation in our default run) for the evaluation, and the value of the corresponding shape parameter (alpha). The next line is very useful for those of you interested in using the new molecular markers for phylogenetics, since it provides the mean and median Shimodaira-Hasegawa-like bipartition support values computed by the PhyML maximum likelihood algorithm. The closer to 1, the higher the resolution level of the tree inferred from the theoretical amplicon set and therefore the higher its phylogenetic information content. Notice that a FASTA-formatted file containing the aligned amplicon sets can be downloaded (in this case at the protein level). The ML trees for each amplicon, estimated under the WAG+G model, can be either directly visualized or downloaded.

If multiple amplicons are found for a given input file, these are presented in a sorted fashion, based on the primer quality criterion. The computation of the phylogenetic information content of the corresponding amplicon alignments is performed on the sorted primer pairs based on their thermodynamic quality (see detailed explanations in the section on technical information).

Back to table of contents

Additional output files

After completing the primer design, the user can download a maximum likelihood tree calculated with phyML v. 2.4.5 (ref1, ref2) under the protein or nucleotide substitution model chosen by the user in the phylogenetic evaluation box (Fig. 1). The tree can be displayed by the server. Finally, a tab-formatted text file containing the information for all thermodynamic parameters computed for each CODEHOP pair can be downloaded using the primer evaluation TAB file link found in the header of the compute primers window. This file can be easily uploaded into spread sheets like excel or oocalc.

A very convenient feature of the primers4clades server is its stored jobs page, shown and explained below:

the stored jobs page allows you to retrieve previously uploaded data for further processing

Figure 2B. The stored jobs page allows you to organize your results, keeping them temporarily on our servers, or to select stored results from previous runs for further processing.

The stored jobs page is only available if you provide your email address while submitting your jobs to primers4clades. From this page, as shown in Fig. 2B, you have access to previously submitted primary data and their corresponding results pages. Furthermore, after retrieving your data, you can continue processing them using perhaps different parameter settings, skipping the multiple sequence alignment steps. Data are deleted from the server one week if not requested otherwise.

Back to table of contents

Customized run case with user-selected parameter values, clades and sequences

Lets now use the server in an interactive fashion by changing the run mode to "cluster sequences" in the corresponding drop box on the input page (Fig. 3). We are going to upload now the actinobacterial demo dataset, which contains sequences from two genera, but we want to target the primer search specifically to the genus Mycobacterium. Because we are going to focus on a single genus, we will make the phylogenetic evaluation of the amplicons at the DNA level. We also change the translation table to 11, which corresponds to the bacterial code. After providing our email, job name and setting the amplicon size range to 550-1250, we hit the submit button (Fig. 3). This will execute the pipeline steps 1-2 and the server will stop making calculations after displaying on screen the NJ tree. At this time the user is presented with several selection boxes to manually select clades or sequences, as shown in Fig 4, before proceeding with the primer design and selection steps.

primers4clades cluster sequences input example

Figure 3. Customized selection of parameters to run the server in the interactive "cluster sequences" mode using the actinobacterial demo sequences.

- **First output page of the interactive "cluster sequences" run mode.**

After submitting the job using the parameter selection show in Fig. 3, we got a first output page that looked like this:

primers4clades job Actinobacteria_demo_sequences (5011270d)

```
_gencode : 11
_evaluation : dna_HKYG
_cluster distance metric : protein
_Tm : 55
_amplicon_length : 550 , 1250
_email : vinuesa@******
```

Clustering sequences...

```
# number of sequences read = 13
# number of recognised taxa = 13
# aligning translated sequences...
# computing distance matrix...
# alignment stats: length = 1210 %gaps = 3.22 %constant = 69.4

# mean distance = 0.16809
# max distance = 0.43419 ( 001 <=> 003 )
# min distance = 0.00000 ( 002 <=> 011 )

# NJ tree with cluster labels :

NJ tree based on a ML WAG+G distance matrix for rpoB sequences from Mycobacterium and Corynebacterium genome sequences
```

re-cluster sequences selection options

Figure 4. Example of the first output page produced by the primers4clades server run in the interactive "cluster sequences" mode.

The header and clustering sequences sections contain essentially the same information fields as described previously (Testing the server using the fungal alpha-tubulin demo sequences and default settings). However, there is a new section entitled Custom clustering parameters found below the NJ tree. If the user wants to find primers that would target all sequences in the original alignment, she/he should simply click the get primers button. The FAA and FNA links allow the user to retrieve the full protein and codon alignments, respectively.

To focus the primer design on the Mycobacterium clade, defined by the subtending branch 9, the user has to indicate the clade boundaries by typing the corresponding tree leave IDs or terminal node numbers (002__0,013__0) that delimit the target clade into the cluster boundaries box, as shown in Fig. 4. This can be conveniently done by simply highlighting the corresponding leave IDs with the mouse and pasting them into the corresponding box (both IDs separated by a coma). Particular sequences of the target clade could be removed by specifying their leave number identifiers. Alternatively, the user can decide to use only those sequences for primer design that are below a chosen distance threshold. As a guide, the user should look at the mean and max distances reported by the server in the header information found on top of the tree. After making the desired choices, the user needs to hit the re-cluster button for them to be taken into account.

Once the target cluster has been defined and returned to the server by hitting the re-cluster button, the user should then hit the **get primers button** in order to calculate the clade-specific primers. In our example, the server returned the following output:

- **Output page produced after having defined the target cluster on which to focus primer design.**

primers4clades job my_sequences (9945ff96)

Received input data:

```
_gencode : 11
_distance matrix file : 9945ff96_intronless_aln.phy.dist
```

primers4clades job Actinobacteria_demo_sequences (9945ff96)

Received input data:

```
_gencode : 11
_distance matrix file : 9945ff96_intronless_aln.phy.dist
_Tm : 55
_amplicon_length : 550 , 1250
_evaluation : dna_HKYG
_cluster distance metric :
_cluster params : -b 002__0,013__0
```

Clustering sequences...

```
# number of sequences read = 13
# number of recognised taxa = 13
# number of valid cluster boundaries = 2
# size of user-selected cluster = 11

# max distance = 0.43419 ( 001 <=> 003 )
# min distance = 0.00000 ( 002 <=> 011 )

# NJ tree with cluster labels :

NJ tree reconstructed from WAG+G ML distances computed from Mycobacterium rpoB sequences
```

cluster sequence and get primers buttons

Figure 5. Example of the second output page produced by the primers4clades server run in the interactive "cluster sequences" mode after having defined the target cluster on which to focus primer design, as shown in Fig. 4.

As can be seen from the output shown in Fig. 5, the current cluster contains only 11 sequences now. We could again change our sequence selection, or hit the get primers button to let the server search for primers for our target clade. The output of the header and a first primer pair found for this clade are shown in Fig. 6, shown below.

- **Output of the interactive "get primers" option entered within the "cluster sequences" run mode.**

primers4clades job rpoB_mycos_tutorial : cluster 0 (9945ff96)

Received input data:

```
_runmode : primers from selected cluster
_tm : 55
_amplicon_length : 550 , 1250
_evaluation : dna_HKYG
```

Calculating primers...

```
## table redundancy vs amplicons stats:
# Mycobacterium_ulcerans_Agy99 redundancy = 0.97 amplicons = 6
# Mycobacterium_sp_KMS redundancy = 0.97 amplicons = 5
# Mycobacterium_leprae redundancy = 0.95 amplicons = 5
# Mycobacterium_gilvum redundancy = 0.92 amplicons = 5
# Mycobacterium_bovis_BCG_str_Pasteur_1173P2 redundancy = 0.97 amplicons = 4
# Mycobacterium_smegmatis_str_MC2_155 redundancy = 0.97 amplicons = 4
# input_derived redundancy = 0.91 amplicons = 4
# Mycobacterium_tuberculosis_CDC1551 redundancy = 0.97 amplicons = 3
# Mycobacterium_vanbaalenii_PYR-1 redundancy = 0.97 amplicons = 3
# Mycobacterium_avium_104 redundancy = 0.97 amplicons = 2

# evaluating primers...
# primer evaluation TAB file

# min quality for phylogenetic evaluations = 50%

# ranking pairs of primers...

Amplicon distribution map generated by primers4clades; Mycobacterium rpoB gene


## Amplicon 1 codon_usage_table = input_derived_

9945ff96 :
CGAGTGCAAGGACAAGgayatgacnta 5'->3' N 141 459 (aligned residues)
cgagtgcaaagacaaggacatgacgta >002 637025853 [MYCOBACTERIUM TUBERCULOSIS H37RV-NC_000962]
agagtgcaaagacaaggacatgacgta >003 637072850 [MYCOBACTERIUM LEPRAE TN-NC_002677]
cgagtgcaaagacaaggacatgacgta >004 637137115 [MYCOBACTERIUM AVIUM SUBSP. PARATUBERCULOSIS K-10-NC_002944]
cgagtgcaaagacaaggacatgacgta >006 639737776 [MYCOBACTERIUM AVIUM 104-NC_008595]
cgagtgcaaagacaaggacatgacgta >007 639739905 [MYCOBACTERIUM SMEGMATIS STR. MC2 155-NC_008596]
tgagtgcaaagacaaggacatgacgta >008 639759327 [MYCOBACTERIUM ULCERANS AGY99-NC_008611]
cgagtgcaaagacaaggacatgacgta >009 639790664 [MYCOBACTERIUM SP. KMS-NC_008705]
cgagtgcaaagacaaggacatgacgta >010 639804758 [MYCOBACTERIUM VANBAALENII PYR-1-NC_008726]
cgagtgcaaagacaaggacatgacgta >011 639829616 [MYCOBACTERIUM BOVIS BCG STR. PASTEUR 1173P2-NC_008769]
cgagtgcaaagacaaggacatgacgta >012 640138311 [MYCOBACTERIUM SP. JLS-NC_009077]
tgagtgcaaagacaaggacatgacgta >013 640459324 [MYCOBACTERIUM GILVUM PYR-GCK-NC_009338]
?........!........!.....!..
CGAGTGCAAAGACAAGgacatgacgta codeh_corr 9945ff96_intronless_cluster_0_amp1_N141
hgagtgcaaagacaaggacatgacgta relax_corr 9945ff96_intronless_cluster_0_amp1_N141
hgagtgcaaagacaaggacatgacgta degen_corr 9945ff96_intronless_cluster_0_amp1_N141

CGGGTTGTTCTGGtccatraaytg 5'->3' C 141 459 (aligned residues)
cgggttgttctggtccatgaattg >002 637025853 [MYCOBACTERIUM TUBERCULOSIS H37RV-NC_000962]
agggttgttctgatccatgaactg >003 637072850 [MYCOBACTERIUM LEPRAE TN-NC_002677]
cgggttgttctggtccatgaactg >004 637137115 [MYCOBACTERIUM AVIUM SUBSP. PARATUBERCULOSIS K-10-NC_002944]
cgggttgttctggtccatgaactg >006 639737776 [MYCOBACTERIUM AVIUM 104-NC_008595]
cgggttgttctggtccatgaactg >007 639739905 [MYCOBACTERIUM SMEGMATIS STR. MC2 155-NC_008596]
cgggttgttctggtccatgaactg >008 639759327 [MYCOBACTERIUM ULCERANS AGY99-NC_008611]
cgggttgttctggtccatgaactg >009 639790664 [MYCOBACTERIUM SP. KMS-NC_008705]
cgggttgttctggtccatgaactg >010 639804758 [MYCOBACTERIUM VANBAALENII PYR-1-NC_008726]
cgggttgttctggtccatgaattg >011 639829616 [MYCOBACTERIUM BOVIS BCG STR. PASTEUR 1173P2-NC_008769]
cgggttgttctggtccatgaactg >012 640138311 [MYCOBACTERIUM SP. JLS-NC_009077]
cgggttgttctgatccatgaactg >013 640459324 [MYCOBACTERIUM GILVUM PYR-GCK-NC_009338]
?..........?.....!.....
CGGGTTGTTCTGGtccatgaaytg codeh_corr 9945ff96_intronless_cluster_0_amp1_C459
mgggttgttctgrtccatgaaytg relax_corr 9945ff96_intronless_cluster_0_amp1_C459
mgggttgttctgrtccatgaaytg degen_corr 9945ff96_intronless_cluster_0_amp1_C459

# primer pair quality = 100%
# expected PCR product length (nt) = 957
# fwd: minTm = 62.0 maxTm = 62.0
# rev: minTm = 63.3 maxTm = 64.7
#--
# phylogenetic amplicon evaluation:
# subs.model = HKYG
# alpha = 0.16
# mean aLRT = 0.96 median aLRT = 0.97
# amplicon FASTA file
# aLRT likelihood tree (newick)
# end_of_amplicon

# many primer pairs omitted ... until reaching the last (worst one)

  ...

## Amplicon 41 codon_usage_table = Mycobacterium_vanbaalenii_PYR-1 :
GGGCCTCGCTGGACgtncarayngga 5'->3' N 72 425 (aligned residues)
gggactccttgacgtccagaccga >002 637025853 [MYCOBACTERIUM TUBERCULOSIS H37RV-NC_000962]
ggggctacttgatgtgcagactga >003 637072850 [MYCOBACTERIUM LEPRAE TN-NC_002677]
cggcctgctggacgtgcagatcga >004 637137115 [MYCOBACTERIUM AVIUM SUBSP. PARATUBERCULOSIS K-10-NC_002944]
cggcctgctggacgtgcagatcga >006 639737776 [MYCOBACTERIUM AVIUM 104-NC_008595]
ggggctgctcgacgttcagacgga >007 639739905 [MYCOBACTERIUM SMEGMATIS STR. MC2 155-NC_008596]
cgggttgctggacgttcagaccga >008 639759327 [MYCOBACTERIUM ULCERANS AGY99-NC_008611]
ggggcttctcgacgtgcagacgga >009 639790664 [MYCOBACTERIUM SP. KMS-NC_008705]
ggggcttctcgacgttcagaccga >010 639804758 [MYCOBACTERIUM VANBAALENII PYR-1-NC_008726]
gggactccttgacgtccagaccga >011 639829616 [MYCOBACTERIUM BOVIS BCG STR. PASTEUR 1173P2-NC_008769]
ggggcttctcgacgtgcagacgga >012 640138311 [MYCOBACTERIUM SP. JLS-NC_009077]
ggggcttctcgacgttcagaccga >013 640459324 [MYCOBACTERIUM GILVUM PYR-GCK-NC_009338]
?..??.?..?..?..?..?..?..
GGGCCTCGCTGGACgtbcagaybga codeh_corr 9945ff96_intronless_cluster_0_amp1_N72
GGGCCTCGCTGGACgtbcagaybga relax_corr 9945ff96_intronless_cluster_0_amp1_N72
sggvytnctbgaygtbcagaybga degen_corr 9945ff96_intronless_cluster_0_amp1_N72

GGCCTCCACGTCCtgngtngtcat 5'->3' C 72 425 (aligned residues)
cgcctccacgtcctgggtggtcat >002 637025853 [MYCOBACTERIUM TUBERCULOSIS H37RV-NC_000962]
cgcctcgacgtcctgggtggtcat >003 637072850 [MYCOBACTERIUM LEPRAE TN-NC_002677]
ggcctcgacgtcctgggtggtcat >004 637137115 [MYCOBACTERIUM AVIUM SUBSP. PARATUBERCULOSIS K-10-NC_002944]
ggcctcgacgtcctgggtggtcat >006 639737776 [MYCOBACTERIUM AVIUM 104-NC_008595]
cgcctcgacgtcctgggtggtcat >007 639739905 [MYCOBACTERIUM SMEGMATIS STR. MC2 155-NC_008596]
cgcctcgacatcctgggtggtcat >008 639759327 [MYCOBACTERIUM ULCERANS AGY99-NC_008611]
cgcctcgacgtcctgggtggtcat >009 639790664 [MYCOBACTERIUM SP. KMS-NC_008705]
cgcctcgacgtcctgggtggtcat >010 639804758 [MYCOBACTERIUM VANBAALENII PYR-1-NC_008726]
cgcctccacgtcctgggtggtcat >011 639829616 [MYCOBACTERIUM BOVIS BCG STR. PASTEUR 1173P2-NC_008769]
cgcctcgacgtcctgggtggtcat >012 640138311 [MYCOBACTERIUM SP. JLS-NC_009077]
cgcctcgacgtcctgggtggtcat >013 640459324 [MYCOBACTERIUM GILVUM PYR-GCK-NC_009338]
?.....?..?.....!..!.....
GGCCTCCACGTCCtgggtggtcat codeh_corr 9945ff96_intronless_cluster_0_amp1_C425
sgcctcsacrtcctgggtggtcat relax_corr 9945ff96_intronless_cluster_0_amp1_C425
sgcctcsacrtcctgggtggtcat degen_corr 9945ff96_intronless_cluster_0_amp1_C425

# primer pair quality = 50%
# expected PCR product length (nt) = 1062
# fwd: minTm = 67.8 maxTm = 73.3
# rev: minTm = 71.5 maxTm = 71.5
# quality warnings:
# fwd fulldeg (5184 >= 193)
# fwd hpinpot (0.63 >= 0.61)
# fwd selfpot (0.62 >= 0.61)
# rev selfpot (0.62 >= 0.61)
# crosspot (0.70 >= 0.61)
#--
# phylogenetic amplicon evaluation:
# subs.model = HKYG
# alpha = 0.16
# mean aLRT = 0.98 median aLRT = 0.99
# amplicon FASTA file
# aLRT likelihood tree (newick)
# end_of_amplicon


#--
```

refine primers selection buttons

Figure 6. Output for the interactive "get primers" output generated by the "cluster sequences" run mode. At the end of the "get primers" analysis it is still possible to refine primer formulations by excluding particular sequences using the click boxes on the left side hand of the NJ tree.

If the amount of predicted amplicons to evaluate phylogenetically is very large, the limit of CPU time for this process may be exceeded at some point and the remaining amplicon sets won' t be evaluated. In our example set, although 41 amplicons were recovered, all get evaluated below 5 min of total CPU time.

As shown the output presented in Fig. 6, at the end of the "get primers" analysis run by the server in "cluster sequences" mode, it is still possible to refine primer formulations by excluding particular sequences using the click boxes on the left-hand side of the NJ tree. In the example above, we will exclude the marked sequences in order to refine or correct the primer formulations for the new target group, in this case the "slow-growing mycobacteria". The new, refined primer formulation of the target group is based on the primers found in the previous step for the Mycobacterium clade. This is a very fast

computation, since the primers are "refined" based simply on the codon alignment, without searching for new primer binding sites or evaluation. This will produce the following self-explanatory output page:

- **Output for the refine primers run mode:**

primers4clades job my_sequences(cluster=0) (9945ff96)

Received input data:

```
_runmode : primer refinement
_jobname : my_sequences(cluster=0)
_excluded_taxon: >007 639739905 [MYCOBACTERIUM SMEGMATIS STR. MC2 155-NC_008596]
_excluded_taxon: >009 639790664 [MYCOBACTERIUM SP. KMS-NC_008705]
_excluded_taxon: >012 640138311 [MYCOBACTERIUM SP. JLS-NC_009077]
_excluded_taxon: >010 639804758 [MYCOBACTERIUM VANBAALENII PYR-1-NC_008726]
_excluded_taxon: >013 640459324 [MYCOBACTERIUM GILVUM PYR-GCK-NC_009338]
```

Refining primers...

```
## Amplicon 1 :
CGAGTGCAAGGACAAGgayatgacnta 5'->3' N 141 459 (aligned residues)
cgagtgcaaagacaaggacatgacgta >002 637025853 [MYCOBACTERIUM TUBERCULOSIS H37RV-NC_000962]
agagtgcaaagacaaggacatgacgta >003 637072850 [MYCOBACTERIUM LEPRAE TN-NC_002677]
cgagtgcaaagacaaggacatgacgta >004 637137115 [MYCOBACTERIUM AVIUM SUBSP. PARATUBERCULOSIS K-10-NC_002944]
cgagtgcaaagacaaggacatgacgta >006 639737776 [MYCOBACTERIUM AVIUM 104-NC_008595]
tgagtgcaaagacaaggacatgacgta >008 639759327 [MYCOBACTERIUM ULCERANS AGY99-NC_008611]
cgagtgcaaagacaaggacatgacgta >011 639829616 [MYCOBACTERIUM BOVIS BCG STR. PASTEUR 1173P2-NC_008769]
?........!........!.....!..
CGAGTGCAAAGACAAGgacatgacgta codeh_corr 1_N141 ( 1 -> 1 )
hgagtgcaaagacaaggacatgacgta relax_corr 1_N141 ( 3 -> 3 )
hgagtgcaaagacaaggacatgacgta degen_corr 1_N141 ( 3 -> 3 )
--------------------------- >007 639739905 [MYCOBACTERIUM SMEGMATIS STR. MC2 155-NC_008596]
--------------------------- >009 639790664 [MYCOBACTERIUM SP. KMS-NC_008705]
--------------------------- >010 639804758 [MYCOBACTERIUM VANBAALENII PYR-1-NC_008726]
--------------------------- >012 640138311 [MYCOBACTERIUM SP. JLS-NC_009077]
--------------------------- >013 640459324 [MYCOBACTERIUM GILVUM PYR-GCK-NC_009338]

CGGGTTGTTCTGGtccatraaytg 5'->3' C 141 459 (aligned residues)
cgggttgttctggtccatgaattg >002 637025853 [MYCOBACTERIUM TUBERCULOSIS H37RV-NC_000962]
agggttgttctgatccatgaactg >003 637072850 [MYCOBACTERIUM LEPRAE TN-NC_002677]
cgggttgttctggtccatgaactg >004 637137115 [MYCOBACTERIUM AVIUM SUBSP. PARATUBERCULOSIS K-10-NC_002944]
cgggttgttctggtccatgaactg >006 639737776 [MYCOBACTERIUM AVIUM 104-NC_008595]
cgggttgttctggtccatgaactg >008 639759327 [MYCOBACTERIUM ULCERANS AGY99-NC_008611]
cgggttgttctggtccatgaattg >011 639829616 [MYCOBACTERIUM BOVIS BCG STR. PASTEUR 1173P2-NC_008769]
?...........?.....!.....
CGGGTTGTTCTGGtccatgaaytg codeh_corr 1_C459 ( 2 -> 2 )
mgggttgttctgrtccatgaaytg relax_corr 1_C459 ( 8 -> 8 )
mgggttgttctgrtccatgaaytg degen_corr 1_C459 ( 8 -> 8 )
----------------------- >007 639739905 [MYCOBACTERIUM SMEGMATIS STR. MC2 155-NC_008596]
----------------------- >009 639790664 [MYCOBACTERIUM SP. KMS-NC_008705]
----------------------- >010 639804758 [MYCOBACTERIUM VANBAALENII PYR-1-NC_008726]
----------------------- >012 640138311 [MYCOBACTERIUM SP. JLS-NC_009077]
----------------------- >013 640459324 [MYCOBACTERIUM GILVUM PYR-GCK-NC_009338]



 ...

 # many primer pairs omitted ... until reaching the last (worst one)

 ...

## Amplicon 41 :
GGGCCTGCTGGACgtncaraynga 5'->3' N 72 425 (aligned residues)
gggactccttgacgtccagaccga >002 637025853 [MYCOBACTERIUM TUBERCULOSIS H37RV-NC_000962]
ggggctacttgatgtgcgactgca >003 637072850 [MYCOBACTERIUM LEPRAE TN-NC_002677]
cggcctgctggacgtgcagatcga >004 637137115 [MYCOBACTERIUM AVIUM SUBSP. PARATUBERCULOSIS K-10-NC_002944]
cggcctgctggacgtgcagatcga >006 639737776 [MYCOBACTERIUM AVIUM 104-NC_008595]
cgggttgctggacgttcagaccga >008 639759327 [MYCOBACTERIUM ULCERANS AGY99-NC_008611]
gggactccttgacgtccaagaccga >011 639829616 [MYCOBACTERIUM BOVIS BCG STR. PASTEUR 1173P2-NC_008769]
?..??.?..?..?..?..!..?..
GGGCCTGCTGGACgtbcagayyga codeh_corr 41_N72 ( 18 -> 12 )
GGGCCTGCTGgaygtbcagayyga relax_corr 41_N72 ( 18 -> 24 )
sggvytvctkgaygtbcagayyga degen_corr 41_N72 ( 5184 -> 1728 )
--------------------g-- >007 639739905 [MYCOBACTERIUM SMEGMATIS STR. MC2 155-NC_008596]
--------------------g-- >009 639790664 [MYCOBACTERIUM SP. KMS-NC_008705]
--------------------g-- >010 639804758 [MYCOBACTERIUM VANBAALENII PYR-1-NC_008726]
--------------------g-- >012 640138311 [MYCOBACTERIUM SP. JLS-NC_009077]
----------------------- >013 640459324 [MYCOBACTERIUM GILVUM PYR-GCK-NC_009338]


GGCCTCCACGTCCtgngtngtcat 5'->3' C 72 425 (aligned residues)
cgcctccacgtcctgggtggtcat >002 637025853 [MYCOBACTERIUM TUBERCULOSIS H37RV-NC_000962]
cgcctcgacgtcctgggtggtcat >003 637072850 [MYCOBACTERIUM LEPRAE TN-NC_002677]
ggcctcgacgtcctgggtggtcat >004 637137115 [MYCOBACTERIUM AVIUM SUBSP. PARATUBERCULOSIS K-10-NC_002944]
ggcctcgacgtcctgggtggtcat >006 639737776 [MYCOBACTERIUM AVIUM 104-NC_008595]
cgcctcgacatcctgggtggtcat >008 639759327 [MYCOBACTERIUM ULCERANS AGY99-NC_008611]
cgcctcacgtcctgggtggtcat >011 639829616 [MYCOBACTERIUM BOVIS BCG STR. PASTEUR 1173P2-NC_008769]
?.....?..?.....!..!.....
GGCCTCCACGTCCtgggtggtcat codeh_corr 41_C425 ( 1 -> 1 )
sgcctcsacrtcctgggtggtcat relax_corr 41_C425 ( 8 -> 8 )
sgcctcsacrtcctgggtggtcat degen_corr 41_C425 ( 8 -> 8 )
----------------------- >007 639739905 [MYCOBACTERIUM SMEGMATIS STR. MC2 155-NC_008596]
----------------------- >009 639790664 [MYCOBACTERIUM SP. KMS-NC_008705]
----------------------- >010 639804758 [MYCOBACTERIUM VANBAALENII PYR-1-NC_008726]
----------------------- >012 640138311 [MYCOBACTERIUM SP. JLS-NC_009077]
----------------------- >013 640459324 [MYCOBACTERIUM GILVUM PYR-GCK-NC_009338]
```

Figure 7. Output of the "refine primers" run mode. Only the first and last primer pairs are shown, corresponding to formulations adjusted to the "fast growing mycobacterial clade" selected in the screen shown in Fig. 6.

Back to table of contents

Technical information

This section will explain in greater detail some of the key concepts and statistical parameters on which **primers4clades** is based.

- **The CODEHOP (COnsensus DEgenerate Hybrid Oligonucleotide Primer) structure.**

The CODEHOP design strategy was first published by Rosen and colleagues in 1998 and 2003 (ref1, ref2). It represents a novel PCR primer design strategy devised to amplify distantly related sequences of a protein family. The primers are derived from amino acid BLOCKS within a protein multiple sequence alignment that don't contain gaps, which are excised from the alignment using the BlockMaker tool. Each hybrid primer consists of a short (11-12 nucleotides-long) 3' degenerate core region that binds to the codons encoding 3-4 highly conserved amino acids, and a longer 5' consensus clamp region that contains the most probable codon predicted on the basis of a user-selected codon usage table. The basic structure of a CODEHOP can be seen here. Amplification initiates by annealing and extension of primers in the pool with the highest similarity in the 3' core region, which confers the specificity of the amplification, and their stabilization by the consensus clamp, which only partially matches the target template, as shown here. Because all primers are identical at their 5' consensus clamp region, they all will anneal at high stringency during subsequent rounds of amplification. This increases the efficiency of the PCR amplification, making it in principle more efficient than PCR with standard consensus or degenerate oligonucleotides.

- **The extended CODEHOP primer design strategy: corrected CODEHOP, relaxed corrected CODEHOP and fully degenerate oligonucleotides computed by the primers4clades server.**

In addition to the standard CODEHOP, the primers4clades server computes three additional primer formulations: the corrected CODEHOP, relaxed corrected CODEHOP and fully degenerate oligonucleotides. The former two are based on the original CODEHOP formulation, but adjusting the degeneracy of the latter by taking into account the sequences of the underlying codon alignment, as shown in Fig. 2. The relaxed corrected CODEHOP extends the 3' core region of the corrected CODEHOP towards the 5' end until it reaches a degeneracy level of 24. If the corrected CODEHOP already has a degeneracy equal or greater than 24, then the latter two formulations will be identical. The fully degenerated oligonucleotide formulation is computed to reflect the full nucleotide sequence variation of the targeted binding site. Table 1 summarizes the recommended uses of the four oligonucleotide formulations computed by the server:

Table 1. Summary of recommended uses of the different oligonucleotide formulations returned by primers4clades.

standard CODEHOP	Diverged family members or rapidly evolving genes. May be particularly useful if only a limited set of reference sequences are available for the target locus.
corrected CODEHOP	New members of a phylogenetic clade of sequences with a moderate to low evolutionary rate, for which a reasonable number of reference sequences are available as input for primers4clades.
corr. relax. CODEHOP	As above, if the degeneracy level of the degenerated core region is low (< 8 or 12), but the 5' end region is highly variable, and you want to maximize the clade-specificity of your oligos in the PCR experiments. You may probably whish to manually adjust the extension of the degenerated 3' region according to your specific needs.
fully degen. oligos	Use only if the overall degeneracy is low (< 8-16) and evenly distributed across the target region. As mentioned above, you may probably whish to manually adjust the extension of the degenerated 3' region according to your specific needs.

- **Computation and selection of non-redundant codon usage tables (CUTs).**

Codon usage tables contain the observed frequency of use of codons in a genomic context. If we consider codon tables as vectors, in which components are relative codon frequencies, then the similarity of any two tables can be computed as the angle between two vectors. Identical tables will have an angle of 0° (cosine=1), whilst orthogonal tables will have an angle of 90° (cosine=0). It is easy to use this formulation to calculate the redundancy of a set of codon tables. In particular, the redundancy of a table with respect to a set of tables can be estimated as the mean similarity with when compared to all of them. When looking for appropriate primers matching a multiple protein alignment, primers4clades will rank the set of suitable codon tables (inferred from taxa in FASTA headers) in terms of redundancy, giving more priority to less redundant tables. In addition, the server will stop trying different tables when two successive tables in the ranking yield no valid primers.

- **Thermodynamic primer pair quality score**.

The server computes the following thermodynamic parameters for the corrected CODEHOP formulation by using python subroutines from the amplicon software, and signals a quality warning message if the parameter values exceed the threshold values indicated in Table 2.

Table 2. Oligonucleotide thermodynamic and degeneracy level parameter threshold values to signal a quality warning.

parameter	threshold (greater than)
3' core degeneracy corr. CODEHOP	8
3' core degeneracy relaxed corr. CODEHOP	24
full degeneracy	192
hairpin formation potential	0.61
self hybridization potential	0.61
cross-hybridization potential	0.61

Five parameters are computed for each individual oligo, whereas the cross-hybridization potential parameter is calculated for each oligo pair. Therefore, there are **11 potential quality warnings** for each oligo pair and hence the server ranks the oligo pairs in descending quality score values, from 100% to 0%. These parameter cut-off values are moderately liberal and have been empirically determined from the evaluation of the performance of 20 corrected CODEHOP primers developed in the laboratory of P. Vinuesa (Figueroa-Palacios and Vinuesa, unpublished data), and 10 standard degenerate primer pairs reported in the literature.

As mentioned before, the TAB file that can be downloaded contains the actual values for each of these parameters of the corrected CODEHOP oligo formulation and several other ones, listed below.

```
# crosspot = potential of cross-hybridization [0-1]
# deg = primer degeneracy in 3' degenerated segment
# relaxdeg = relaxed (3' extended segment) degeneracy
# fulldeg = full primer degeneracy
# minTm = minimum Tm for the pool of codehop primers
# maxTm = maximum Tm for the pool of codehop primers
# hpinpot = potential of primer hairpin [0-1]
# selfpot = potential of primer self-priming [0-1]
```

In addition to the above mentioned parameters, the primer design and selection pipeline removes al primer pairs that have:

- a Tm difference > 3°C
- a CODEHOP score difference > 4
- a degeneracy difference > 3 [in log(2) scale]
- a length difference > 4 nucleotides
- an overall primer match < 65% with the corresponding codon alignment

Finally, in order to minimize primer/amplicon redundancy, only amplicons that have a sequence coverage < 20% are considered.

- **Quantification of the phylogenetic information content of aligned amplicon sets**

A unique feature of the primers4clades server is its ability to compute a robust phylogenetic information content parameter for each theoretical amplicon, which is based on a recently developed Shimodaira-Hasegawa (SH)-like test for the significance of branches in a maximum likelihood tree, implemented in PhyML v2.4.5 (ref). In brief, the test assesses whether the branch being studied provides a significant likelihood gain, in comparison with the null hypothesis that involves collapsing that branch, but leaving the rest of the tree topology identical. We chose the SH-like procedure for assessing bipartition significance because the test is nonparametric and much less liberal than the diverse (parametric) approximate-likelihood ratio tests that are also implemented in that program. The resulting SH-like branch support values therefore indicate the probability that the corresponding split is significant.

The phylogenetic information content of each amplicon is calculated as the mean and median SH-like branch support values for the corresponding maximum likelihood tree inferred under the user-specified substitution model or matrix. The closer to 1, the higher the phylogenetic information content of the targeted sequence region. This strategy of calculating the phylogenetic information content of different sequence alignments was recently reported by Vinuesa et al. (2008).

- **Computing a non-redundant set of primer pairs yielding informative amplicon sets with high coverage of the target locus**

A unique and very useful feature of primers4clades is that the server returns a non-redundant set of primer pair formulations, ranked according to their thermodynamic properties. Furthermore, the phylogenetic information content of the aligned amplicon sets each primer pair would theoretically amplify, given the input sequences, is also computed. This is a key piece of information to take into account when making the choice between alternative primer formulations with equal quality. The server checks that the resulting amplicon sets for the primer pairs with a quality score equal or above 50% do not overlap more than 80%, when derived from the same codon usage table, ensuring a high coverage of the target locus but filtering excessive redundancy, as shown on the amplicon distribution maps.

- **Error messages and their meaning**

The primers4clades server will often issue messages during the execution of jobs. There are two types of messages that are of special interest: warnings and errors. While warnings do not prevent the completion of jobs, errors must be handled by the user in subsequent subsmissions.

- The most usual warning is:

"# WARNING : failed parsing taxa in FASTA headers, using a set of representative codon usage tables".

This warning appears when primers4clades fails to recover any taxon names in the FASTA headers of input nucleotide sequences, expected to be bracketed.

- There are two common errors:

"ERROR: Please paste nucleotide sequences instead of amino acid strings"

The server expects nucleotides sequences and will complain whenever protein sequences are used as input.

"ERROR: Translated protein sequences contain internal STOP codon(s). Make sure that the first nucleotide in your DNA frames corresponds to the first codon position or/and that an appropriate translation table is applied."

Users are expected to input coding sequences in which the first codon corresponds to the first 3 nucleotides.

Back to table of contents

Genome-scale benchmark analysis of the primers4clades pipeline

- **Genomic NJ-median distance tree computed from 19 Rhizobiales genomes selected for the benchmark analysis**

It is important to acknowledge key observations and parameters that affect the value of results generated by primers4clades. In order to identify those parameters and their critical cut-off values, we performed a genome-wide benchmark analysis using the 983 orthologous gene families shared by 19 fully sequenced rhizobial genomes used for the analysis and listed on the tree shown in Fig. 8. This is a genomic NJ-median distance tree calculated from a subset of 264 of the above mentioned distance matrices that yielded congruent phylogenies, and depicts the phylogenetic relationships between the 19 rhizobial genomes used for the benchmark analysis. Colours represent the three families analyzed: Bradyrhizobiaceae, Phyllobacteriaceae and Rhizobiaceae (blue, green, red).

median distance genome tree for 19 rhizobiales computed from 983 orthologous gene family alignments using ML WAG+G distances

Figure 8. Genomic median distance tree computed from 264 orthologous and congruent gene families shared by 19 fully sequenced rhizobial genomes used for the benchmark analysis. The bar scale represents the expected number of substitutions per site based on the WAG substitution matrix with gamma-corrected among-site rate variation.

- **Computation of a set of 983 orthologous gene families and the corresponding in silico amplicons**

The orthologous gene families were identified by the bidirectional best hit procedure imposing a conservative BLASTP E-value cut-off level of 10exp(-5), a minimum pairwise sequence coverage of 70% and the presence of the gene in all of the 19 genomes analyzed. Each of these gene families was aligned at the protein level using Muscle with three refinement rounds, and a WAG+G maximum likelihood distance matrix was computed for each alignment using Tree-Puzzle.

- **Analysis of five alignment properties on the numbers and quality of predicted primer pairs per locus**

For the benchmark analysis we specifically tested the influence of the following parameters on the numbers of primer pairs per locus predicted by the primers4clades pipeline:

1. Protein alignment length.
2. Percentage of gaps in the alignment.
3. Maximum WAG+G ML distance between pairs of sequences in a gene family multiple sequence alignment (protein level).
4. Among site rate variation in the protein alignment, measured as a function of the alpha (shape) parameter of the gamma distribution, estimated under ML using the WAG+G model with 8 discrete rate categories.
5. Number of codon tables used per alignment.

The results of these analyses are summarized in Fig. 9A-E.

Genomic benchmark analysis to test the influence of diverse alignment parameters on the numbers of predicted primer pairs per locus

Genome-wide benchmark analysis to evaluate the effect of codon tables on the number of predicted primer pairs per locus

Figure 9. Genome-scale benchmark analysis to test the influence of diverse alignment parameters on the numbers of predicted primer pairs obtained per locus. The analyses were performed on a set of 983 orthologous gene family alignments (at the protein level) for the 19 rhizobiales genomes used to reconstruct the phylogeny shown in Fig. 8.

- **Conclusions**

From the results presented in Fig. 9 it is clear that the number of predicted primer pairs per locus increases linearly with the alignment length (9A) and with the number of codon usage tables (9E) analyzed, whereas a linear decrease in predicted primer pairs per locus is observed with an increasing percentage of gapped sites (9B). Interestingly, Fig. 9C shows that for alignments containing sequences with a WAG+G ML distance > 2.5 the primers4clades pipeline will have a very low chance of finding suitable primer binding sites. It is also interesting that an among-site rate variation level accommodated by an alpha value in the range of 0.3-0.6 is the optimal one for the system to find primer binding sites (9D).

Back to table of contents

Experimental (validation) examples

In this section we demonstrate the practical utility of the primers4clades server to develop PCR primers for two different microbial ecology research projects that are currently being developed in the group of P. Vinuesa: I) the amplification of mycobacterial rpoB sequences from metagenomic DNA extracted from three contrasting soil types and II) the amplification of 5 loci (dnaE, fusA, lon, pheS and rpoB) from the genomic DNA of a collection of 28 diverse Bradyrhizobium strains from around the world.

Amplification of mycobacterial rpoB sequences from metagenomic DNA

Metagenomic DNA was extracted from three contrasting soil types and geographic sites:

1. Juan López, Natural Park of Teno in Tenerife, Canary Islands, Spain;
2. The tropical humid forest reserve of Los Tuxtlas,Estación Biológica de la UNAM, Veracruz, México;
3. A conserved patch of seasonally dry deciduous tropical forest near "El Limón" in the Biosphere Reserve of Sierra de Huautla, Morelos, México.

Primer design was based on full-length rpoB sequences of Corynebacterium, Nocardia, Nocardioides and Mycobacterium retrieved from the NCBI using the Entrez nucleotide system, as well as from the Integrated Microbial Genomes site. Primer design was targeted to the Mycobacterium clade, using the same approach shown in the tutorial above, although with a larger set of taxa.

- The selected corrected-CODEHOP primer pair was:

oligonucleotide name	length	sequence	Tm	core deg.	full deg	h_pot	c_pot	expected amplicon size
rpoB2_mycos_N326	26	GAGAACCTGTTCTTCaaggagaagcg	63.3	0	0	0.52	0.5	1296
rpoB2_mycos_C757	25	CGTCCTCGTAGTTGtgrccytccca	66.6	4	4	0.00	0.5	1296

The Tm is that of the oligo in the corresponding degenerate pool showing the lowest Tm; core deg. refers to the degeneracy level of the 3' degenerated region of the corrected CODEHOP formulations, while full deg. refers to the full degeneracy of the region targeted by the primer. h_pot and c_pot refer to the hairpin-formation potential and cross-hybridization potential of the primers.

DNA was extracted from fresh, air-dryed soil samples using a commercial kit. Figure 10 shows the PCR amplification obtained from the three soil samples using two template DNA concentration levels, as indicated in the figure legend. A hot-start PCR approach was very useful to optimize the specificity of the protocol. Clone libraries were constructed from the purified PCR products of the three soil samples. Twenty clones from each library were randomly chosen for sequencing. All sequences belonged to Actinobacteria, and over 90% of them clustered within the Mycobacterium clade as judged from a ML gene tree inferred from the sample and reference sequences (not shown). Furthermore, the environmental Mycobacterium rpoB sequences clustered within both the fast- and slow-growing clades of mycobacteria, demonstrating the utility of the primers4clades primer design pipeline to develop clade-specific oligonucleotides for metagenomic and microbial ecology studies. Large scale sequencing and analysis of the libraries is being currently performed.

PCR amplification of an rpoB fragment from environmental mycobacteria using metagenomic DNA extracted from three contrasting soils as template	Figure 10. PCR amplification of an rpoB fragment from environmental mycobacteria using metagenomic DNA extracted from three contrasting soils as template. Lines 1-3 and 4-6 show the amplification products from ~20 and ~40 ng of metagenomic DNA isolated from the following sites: 1. Juan López, Natural Park of Teno in Tenerife, Canary Islands, Spain (lines 1 and 4); 2. The tropical humid forest reserve of Los Tuxtlas, Estación Experimental de la UNAM, Veracruz, México (lanes 2,5); 3. A conserved patch of seasonally dry deciduous tropical forest near "El Limón" in the Biosphere Reserve of Sierra de Huautla, Morelos, México (lanes 3,6). PCR experiments and picture: Bernardo Sachman-Ruiz.

Figure 11 shows the results of a ML tree search on a dataset containing the 14 reference actinobacterial rpoB sequences from our third demo, and 26 non-redundant rpoB sequences from our clone libraries. The search was performed using phyml v2.4.5 under the TrN+G model, as detailed in the figure. Notice that the Mycobacterium clade receives a full support (P=1) in the SH-like branch support test, and that all environmental sequences.

ML phylogeny of environmental rpoB sequences

Figure 11. Maximum likelihood phylogeny under the TrN+G model of 26 non-redundant sequences from our rpoB clone libraries (CAN=Canary Islands, Teno, Spain; TUX=Los Tuxtlas, Ver. Mexico; REB=REBIOSH, Mor. Mexico) and 14 reference sequences retrieved from the IMG site (bracketted long names).

Back to table of contents

Amplification of dnaE, fusA, lon, pheS and rpoB genes from diverse Bradyrhizobium strains

Figure 12 shows the results of amplification experiments performed with the primers shown in Table 3 using genomic DNA purified from a collection of 28 diverse Bradyrhizobium strains from around the globe for which these loci had not previously been amplified.
A touch-down PCR technique was useful in all cases, starting at 3 degrees Celsius above the optimal annealing temperature, using the primers shown in Table 3. The results shown in Fig. 12 were published in the bachelor's degree thesis of Iraís Figueroa-Palacios.

Amplification of dnaE, fusA, lon, pheS and rpoB genes from diverse Bradyrhizobium strains

Figure 12. PCR amplification of dnaE, fusA, lon, pheS and rpoB genes from a collection of 28 diverse Bradyrhizobium strains from around the world using the "corrected CODEHOP" formulations shown in Table 3. PCR experiments and pictures: Iraís Figueroa.

Table 3. "Corrected CODEHOP" formulations inferred from a selection of 8 rhizobia (symbiotic root-nodule bacteria) in the families *Rhizobiaceae*, *Phyllobacteriaceae* and *Bradyrhizobiaceae*.
Table 3. relaxed corrected CODEHOPs used to amplify the dnaE, fusA, lon, pheS and rpoB genes from diverse Bradyrhizobium strains

[1]The primers were derived from the analysis of 8 fully sequenced rhizobial genomes, namely: *Agrobacterium tumefaciens* C58 (UWash), *Bradyrhizobium japonicum* USDA110, *Mesorhizobium loti* MAFFT303099, *Mesorhizobium* sp. BNC1, *Nitrobacter hamburgensis* X14, *Rhizobium etli* CFN_42, *R. leguminosarum* bv. *viciae* 3841, and *Sinorhizobium meliloti* 1021. The Tm is that of the oligo in the

corresponding degenerate pool showing the lowest Tm. Deg. refers to the degeneracy level of the 3' degenerated region of the corrected CODEHOP formulations, while Full deg. refers to the full degeneracy of the region targeted by the primer. h_pot and c_pot refer to the hairpin-formation potential and cross-hybridization potential of the primers.

All amplicons were purified and successfully sequenced with the same oligos used for the primary amplification, revealing that they were indeed the expected loci (BLASTN analysis; data not shown) and that they had a high phylogenetic information content (see Fig. 13 below). A multilocus sequence analysis was performed with these results in the context of the bachelor's degree thesis of Iraís Figueroa-Palacios, and reported at the Genomics of Nitrogen-Fixing Organisms - Workshop I, in the framework of the 8th European Nitrogen Fixation Conference, Gent, Belgium, August 30th - September 3rd.

Figure 13 shows a Bayesian phylogeny using partitoned models and the 5 sequence data partitions generated by our new primers. The phylogeny was obtained by means of an MC^3 analysis with four chains that was run in two replicate experiments for $5x10^6$ generations and a sample frequency of 200 generations. The first 12501 trees from each MC^3 run were discarded as burn-in and the remaining 25000 trees were pooled to compute the majority rule consensus tree shown in Fig. 13. Notice the high overall resolution of the tree, with most bipartitions having maximal posterior probability. This phylogeny supports our recent finding (Vinuesa et al. 2008) that B. japonicum Ia strains such as USDA 110 do not form a monophyletic group with homology group I strains, which contain the type strain of the species, reinforcing our conclusion that group Ia strains form an independent, not yet named Bradyrhizobium lineage.

Bayesian phylogeny of Bradyrhizobium strains

Figure 13. Bayesian phylogeny inferred from the 5 sequence data partitions generated by our new primers for 28 Bradyrhizobium strains from around the globe using partitioned models.

Back to table of contents

FAQs

- How should I cite primers4clades?

The manuscript was accepted on April 27th for its publication in the "server issue 2009" of Nucleic Acids Res. (*Advance Access published on May 21, 2009;*) doi:10.1093/nar/gkp377.) Please see themain reference for further details.

- What are the references for the second-party open-source software run by the primers4clades server?

The server relies on several resources in order to provide service: CODEHOP, BLAST, MUSCLE, Codon Usage Database, TREE-PUZZLE, PhyML, PHYLIP, Amplicon, NJplot and Bioperl.

You should cite the following references to give them credit:

1. Anisimova, M. and Gascuel, O. (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst. Biol.*, 55, 539-352.

2. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32, 1792-1797.

3. Felsenstein, J. (2004). PHYLIP (Phylogeny Inference Package) v3.6. Distributed by the author. Department of Genetics, University of Washington, Seattle.

4. Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, 52, 696-704.

5. Jarman, S.N. (2004) Amplicon: software for designing PCR primers on aligned DNA sequences. *Bioinformatics*, 20, 1644-1645.

6. Rose, T.M., Henikoff, J.G. and Henikoff, S. (2003) CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. *Nucleic Acids Res.*, 31, 3763-3766.

7. Rose, T.M., Schultz, E.R., Henikoff, J.G., Pietrokovski, S., McCallum, C.M. and Henikoff, S. (1998) Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic Acids Res.*, 26, 1628-1635.

8. Schmidt, H.A., Strimmer, K., Vingron, M. and von Haeseler, A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18, 502-504.

9. Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H.*et al.*(2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, 12, 1611-1618.

- I can't upload the demo sequences, what's wrong?

If you click on the demo button and nothing happens, you need to enable JavaScript in you browser configuration settings.

- Is there a limit set to the number of sequences I can upload to the server and/or to the processor time consumed by a job?

Yes, you may upload a data set containing up to 35 haplotypes (distinct sequences). There is also an upper time limit set to 7 min. for the process of ML evaluation of the phylogenetic information content of the amplicon sets found by the server and sorted according to their thermodynamic quality parameter. If many long amplicons are found for a large dataset, it is possible that you exceed this time. All sets that had not been evaluated before this time-out limit (currently set to 7 minutes), won't be evaluated phylogenetically. A corresponding warning message will be issued.

- Why does my browser freeze?

This can frequently happen when larger datasets are uploaded to the server due to a browser time out. To avoid loosing your work, don't forget to provide your email in the input page. You will be able then to get access to your primary data and results stored on our servers through the stored jobs page.

- When should I use DNA or protein sequences to compute the reference NJ tree?

This depends essentially on the evolutionary divergence level of your data set. As a thumb rule, data sets that include sequences from taxa classified in different families, orders, classes or phyla should be analyzed at the protein level. This may be also useful for very fast-evolving sequences such as those from phages and viruses. In all these cases the third codon positions are very likely saturated. When your dataset contains sequences of different closely related species (of the same family or genus) or multiple paralogous sequences of a single organism you probably should use the DNA sequences to compute the evolutionary distances between sequences. This should be always done if there is very little divergence at the protein sequence level.

- Why can't I use less than four sequences for primer design with primers4clades?

One of the unique features of the primers4clades server is that it will calculate the phylogenetic information content for each amplicon based on the Shimodaira-Hasegawa-like branch support values parsed from a maximum-likelihood tree inferred from the aligned theoretical amplicons. A non-trivial tree has to have at least four leaves or terminal nodes.

- How should I decide which substitution matrix or model to use for the evaluation of the phylogenetic information content of the amplicons?

As explained above, the primers4clades server will use maximum likelihood methods to calculate the phylogenetic information content for each amplicon. Distance-matrix and maximum likelihood tree reconstruction methods both rely on explicit Markov models designed to approximate the evolutionary rates of nucleotide or amino acid sites in multiple sequence alignments. The ML method is known to be relatively robust to model misspecifications, but the methods work best when a reasonable substitution model is chosen.
Two strategies have been used to develop probabilistic substitution models: the empirical and the parametric approaches. The former has been used to model aminoacid replacement rates, while the second approach is the one used to model the process of nucleotide substitution, as explained on this webpage, which explains the basics of model fitting in phylogenetics. As a thumb rule, we have found that the JTT and WAG substitution matrices are in most cases the better fitting ones for protein data sets. For DNA sequences, a standard, reasonably complex model is the HKY85 (Hasegawa-Kishino-Yano model published in 1985), which takes into account heterogeneity in base frequency and bias in the transition/transversion substitution rate. If you see many replacements on your codon alignment, you may want to use a more parameterized (complex) model, such as TrN93 or the GTR model.
In addition to the rate and frequency parameters mentioned above, significantly more realism is added to the base model chosen by including additional parameters that model substitution rate heterogeneity across the alignment sites. The standard Markovian process assumes that the rates of evolution are equally distributed and independent over sites, which is a very unrealistic assumption. A discontinuous gamma distribution is most frequently used to model this rate heterogeneity over sites. Prof. Ziheng Yang pioneered and popularized its use in phylogenetics, showing the tremendous impact it has on phylogeny estimation and in molecular evolution (Yang, 1996). A single parameter (alpha or shape parameter) controls the form of the gamma distribution, making it very convenient in a phylogenetic context. When alpha has a value < 1 there is strong among-site variation present in the data set. The higher the value of alpha, the lower the heterogeneity (as shown in Fig. 2 of the model-fitting tutorial).
Due to the importance of acknowledging among-site rate variation, all the models that can be chosen by the user
the discrete gamma distribution parameter. Also the ML distance matrices computed from the alignments to reconstruct the reference NJ trees use the gamma parameter (WAG+G and HKY+G for protein and DNA alignments, respectively).

- Which primer formulation should I use from the four ones returned by the server?

A short guide intended to help you in resolving this question is presented in Table 1.

Back to table of contents

This page was last updated on: April 27th, 2009.