ThermoPhyl: a software tool for designing thermodynamically and phylogenetically optimized quantitative-PCR assays.

This manual provides a quick working guide to ThermoPhyl, a pattern-matching script that rapidly compares and summarizes the number of matches between a list of candidate PCR assays and designated target and non-target sequences. **ThermoPhyl is designed for applications in which a user wishes to specifically and sensitively target a taxonomic group of interest in a complex sample using quantitative-PCR or traditional PCR.** The name ThermoPhyl derives from its central goal which is to test thermodynamically optimal PCR assays for phylogenetic sensitivity and specificity to arrive at an assay which is both **thermo**dynamically and **phyl**ogenetically optimal. In the past, users have typically designed PCR primers and/or probes manually, by visual comparison against multiple alignment files. Even when a phylogenetically suitable (i.e. maximally sensitive and specific) can be determined, empirical tests often produce poor PCR results.

ThermoPhyl is designed to use as input a very large number of candidate assays which should all produce an efficient PCR. ThermoPhyl assesses each of these candidate assays for phylogenetic sensitivity and specificity. **The outputs are summary tables of the number of matches to sequences designated as target and non-target groups by the user.**

Installing and Running ThermoPhyl

Before ThermoPhyl can be run, PERL must be installed. A download is available here: http://www.activestate.com/Products/activeperl/index.mhtml

To install ThermoPhyl, the program file simply needs to be copied into the PERL directory or, if PERL is added to the computer's path, any directory of choice.

To run ThermoPhyl, double-click on the ThermoPhyl file, or call up a command window (Start>Run->'cmd'), navigate to the directory where you have copied ThermoPhyl (for example, 'cd C:\PERL') and type 'ThermoPhyl_v1.4.pl' to start the program.

ThermoPhyl requires three input files:

1) **A fasta file which contains all of the target and non-target sequences you wish to test (has to be called "outside_world.fas")**. The more sequences this file contains, the higher the confidence in distinguishing between target and non-target groups. This file can contain the users own sequences and/or sequences retrieved from public databases. Typically, this file might contain 100 – 50,000 sequences. Be aware that many databases such as GreenGenes and Silva for 16S rRNA genes contain many very similar sequences; users will generally want to reduce these databases to some sort of core set of representative sequences.

2) **A text file containing only the names of target sequences (has to be called "target_list.txt")** . The names must correspond exactly to those in the fasta file above and should be some sort of unique identifier, like a GenBank Accession number or GreenGenesID. A column heading, such as "name" or "Acc No") can be present or not.

3) **A list of candidate assays for traditional PCR or qPCR (has to be called "candidate_assays.txt")**. The recommended approach is to first compile all target sequences in a single fasta file. For traditional PCR, this file can be read directly into BatchPrimer3 (http://probes.pw.usda.gov/cgi-bin/batchprimer3/batchprimer3.cgi) or for qPCR, each sequence can be input into ABI's PrimerExpress. We typically generate 50 candidate assays per target sequence. All possible candidate assays should be compiled into a **single tab-delimited text file (.txt) with columns in the order of Forward Primer, Probe (if present), Reverse Primer, with or without column headings.** The file can contain >10,000 candidate assays and should like either look like this:

FORWARD	PROBE	REVERSE
TGATTGACCACACCCGTATTACC	GCCGTTCACCTCAGCCTTAG	ATCTCTGCTTGTCCGCTC
CGCTGTTCATGCTTCCGATA	GATCGATCATCGGCGGTTT	CCTCGGTGTGCATCG

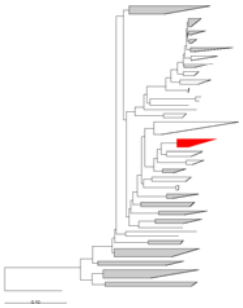
Or like this:

PRIMER_3_OUTPUT
TGATTGACCACACCCGTATTACC
ATCTCTGCTTGTCCGCTC
CGCTGTTCATGCTTCCGATA
CCTCGGTGTGCATCG

In the last example of output from BatchPrimer3 for traditional PCR, the 1st sequence in the list is a forward primer, the 2nd the matching reverse primer, and so on. Each line should have only a line return (paragraph mark) at the end and in the case of multiple columns, each column should be tab-delimited.

Specific Recommendations and Potential Pitfalls:

1) Target groups should form some sort of natural phylogenetic group. Sequences should be properly placed in some sort of a phylogenetic tree to evaluate this and to designate target and non-target sequences in a way which reflects the evolutionary history of the genetic locus or loci in question. In the cartoon to the right, sensitive and specific assays could probably be designed to successfully distinguish between sequences in the red clade and other, non-target, taxa in the rest of the tree. If your target sequences do not form a coherent phylogenetic group, it will obviously be difficult to design an accurate assay, although it is possible that different sequence data (i.e. different portions of the same alignment or another locus) for the same taxa could still be used in such a case.



2) The more sequence data available for both target and non-target groups, the better. The strength of ThermoPhyl, in fact its central goal, is to summarize a very large number of comparisons to arrive at a single 'best' assay.

3) Files must be in the formats described above with the file names given above. It is good practice to keep each analysis in a separate file. Common problems are listed in the FAQ.

FAQ

1. **Q: How is ThermoPhyl different than other primer design programs like PrimerBlast?**

   A: ThermoPhyl differs from other programs in some important ways, summarized in the table below.

Software	Database size	Remote/ Local	FISH Assays (single oligo)	Traditional PCR Assays (two primers)	Quantitative PCR Assays (two primers and probe)	Through put	Summary output sensitivity and specificity
Primer Blast	Large (Genbank)	R	Y	Y	N	Low	N
Probe Check	Various public databases	R	Y	Y	N	Low	N
Primique	Small to Large (user-defined)	R	N	Y	N	High w/ speed limits	N
ThermoPhyl	Small to Large (user-defined)	L	Y	Y	Y	High	Y

2. **Q: What's the best way to design candidate assays?**

   A: For traditional PCR, the most efficient way is to compile all (or a representative set) of your target sequences into a single fasta file, and simply feed this into BatchPrimer3. For qPCR, we have had good results with ABI's PrimerExpress program, although each sequence must be input singly and the output compiled into a single file. We typically generate 50 candidate assays per target sequence, which can be defined in the primer design program.

3. **Q: My output from BatchPrimer3 is formatted strangely. How do I change it?**

   A: No need to worry. ThermoPhyl has an option to use the output from Batch Primer3 as is. Just select the appropriate option in the introductory screen.

4. **Q: If I design so many assays for the similar sequences won't I end up with a bunch of redundant assays?**

   A: Good question. Yes, you probably will, but not to detect these and filter them down to a unique list

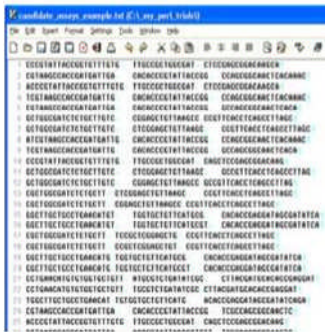
5. **Q: ThermoPhyl tells me I have some discrepancy between my target list and my sequence database and so can't calculate properly the number of matches to targets vs non-targets.**

   A: Common causes of this problem include strange characters (such as &*,!%) in sequence names (in the "outside_world.fas") file and/or the target list ("target_list.txt" file), non-unique identifiers ('AY100' with file of sequence names which include AY100908, AY100909, etc.), and duplicate entries in the sequence database ("outside_world.fas" file). Correcting these errors should resolve the problem. ThermoPhyl will still run despite these errors, but the sensitivity and specificity calculations will be adversely affected. Common problems with sequence databases include duplicate entries, gibberish or inconsistent naming schemes, etc.

Remember, “Garbage In, Garbage Out”.

6. Q: ThermoPhyl doesn’t find any matches in my dataset

A: This can reflect the truth (although matches should always be found among your target sequences!), but more often is an artifact of either having the columns in the assay file in the wrong order (correct order is Forward, Reverse or Forward, Probe, Reverse). Another common problem is having the Reverse Primer sequences reverse-complemented. They should be listed as true Reverse Primers as you would order the oligos (i.e. written 5’ – 3’ for the opposite strand; this is format used by Batch Primer3 and Primer Express to output their assays). The script automatically reverse-complements them to search on the same strand as the Forward Primer. Finally, the primer file (“candidate_assays.txt”) should have no special characters other than tab-delimitations and newline characters as in the example. Finally, make sure you aren’t mixing ‘T’s and ‘U’s in your primers and your outside world file.



7. Q: How long does ThermoPhyl take to run?

A: In tests on moderately fast WinXP machines (e.g. 2 GHz Pentium CPU w/ 3 Gb of RAM), we have tested 5,000 candidate qPCR assays against a database with 5,000 taxa, for 25 million comparisons in about 2.5 hrs. Most users will have many fewer comparison than this and for most applications, ThermoPhyl produces output in seconds to minutes.

8. Q: Why doesn’t ThermoPhyl search for degenerate bases/mismatches in my probe(s) and/or primers?

A: Two reasons:

First, because the premise of ThermoPhyl is to start with thermodynamically optimized assays, any introduction of degenerate bases is considered to be a compromise. For qPCR, ABI recommends against the use of degenerate primers or probes. Ideally, one should be able to find a sensitive and specific assay for a given target group without introducing degenerate bases.

Second, ThermoPhyl is designed to summarize large datasets and the output of searches with varying numbers of mismatches would overwhelm most users and defeat the purpose of the program. We always recommend comparing several of the best assays to your alignment file, and if absolutely necessary, introducing degeneracies at that point.

9. Q: Can ThermoPhyl evaluate FISH probes as well?

A: Yes. Just select the option for a single oligo in the opening screen.

10. Q: What are the output files and how do I view them?

A: ThermoPhyl produces two tab-delimited text files and automatically opens them in a text editor. The first file (‘sorted_search_results.txt’) is a summary table with a sorted tally of the number of target and non-target sequences matched by each candidate assay. The second file (‘raw_search_results.txt’) contains details of each match, including the amplicon length, the position on the DNA strand where each primer/probe matched, etc. These files can also be viewed more conveniently in a spreadsheet like MS Excel.

11. **Q: Is the 'R_motif' sequence printed out by ThermoPhyl ready to order as a Reverse Primer?**

    A: No, it needs to be reverse-complemented. ThermoPhyl prints it out this way to facilitate searching on the same strand by the user when comparing output.

12. **Q Everything worked fine, now can I just order the first assay on the list?**

    A: For some applications, that may be OK, but in general we highly recommend looking at several of the highly ranked assays in a multiple alignment file (the ARB editor is really good for this) to confirm sensitivity and specificity, check for possible degeneracies, etc.

Brian B. Oakley
USDA ARS
Richard Russell Research Center, 950
College Station Road,
Athens, GA 30605
USA
e-mail: brian.oakley@ars.usda.gov