*W516–W520* *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
*doi:10.1093/nar/gki425*

# PriFi: using a multiple alignment of related sequences to find primers for amplification of homologs

**Jakob Fredslund*, Leif Schauser, Lene H. Madsen[1], Niels Sandal[1] and Jens Stougaard[1]**

Bioinformatics Research Center, University of Aarhus, Hoegh-Guldbergsgade 10, 8000 Aarhus C, Denmark and [1]Laboratory of Gene Expression, Department of Molecular Biology, University of Aarhus, Gustav Wieds Vej 10, 8000 Aarhus C, Denmark

Received February 2, 2005; Revised March 15, 2005; Accepted March 23, 2005

## ABSTRACT

**Using a comparative approach, the web program PriFi (http://cgi-www.daimi.au.dk/cgi-chili/PriFi/main) designs pairs of primers useful for PCR amplification of genomic DNA in species where prior sequence information is not available. The program works with an alignment of DNA sequences from phylogenetically related species and outputs a list of possibly degenerate primer pairs fulfilling a number of criteria, such that the primers have a maximal probability of amplifying orthologous sequences in other phylogenetically related species. Operating on a genome-wide scale, PriFi automates the first steps of a procedure for developing general markers serving as common anchor loci across species. To accommodate users with special preferences, configuration settings and criteria can be customized.**

## INTRODUCTION

The development of molecular genetic markers that can be transferred between species in order to exploit syntenic relationships is of importance in diverse research areas ranging from comparative genetics over medical applications to agricultural breeding programs, because such markers can be used to optimize the exploitation of genetic map resources. Typical examples are marker-assisted breeding programs, where marker development is based on the identification of DNA polymorphisms between two mapping parents. These research programs are often marker limited due to lack of DNA sequence information from the species in question. Bioinformatic approaches to exploit information from related species combined with systematic identification of polymorphisms in PCR-amplified fragments of gene orthologs could change this situation.

Such PCR-based strategies for the identification of polymorphisms require the design of primers from multiple sequence alignments, focusing on conserved and variable regions. Automating this task enables high-throughput identification of candidate sites and eliminates the time-consuming and error-prone manual processing of hundreds of alignments. Taking a comparative approach, our web-based primer finder, 'PriFi', evaluates a multiple alignment of phylogenetically related DNA sequences and suggests pairs of primers located in highly conserved regions. These primers are expected to amplify orthologous sequences from related species where sequence information is missing. The pairs are scored according to their quality, and a brief report explaining the score follows each pair. The alignment, suggested primers, and PCR products are displayed graphically, both in a schematic overview and in a letter sequence alignment. The PriFi web site also includes a tutorial. PriFi users should cite this paper and PriFi's URL if they wish to reference the program.

## MATERIALS AND METHODS

The input to PriFi is a multiple sequence alignment. PriFi allows the user to upload either a given alignment file directly (in the Clustalw .aln format) or a file containing multiple sequences (in the FASTA format), from which the program then creates an alignment using Clustalw (Clustalw is used with permission from the European Bioinformatics Institute website: http://www.ebi.ac.uk/clustalw/). PriFi may be run in a general mode with any DNA sequences, but it is designed to find primers in specialized alignments where at least one of the sequences has annotated introns. In this intron mode, a primer pair is only valid if its expected PCR product includes an intron (marked by X'es), in order to enhance the chance of polymorphism discovery. Before uploading any sequences, the user must manually substitute each intronic region with a series of X'es, indicating the approximate length of the intron (e.g. XXX means <20 nt and XXXX means 201–500 nt); see an example in Figure 1. Currently, PriFi cannot automatically identify introns. By default, PriFi runs in intron mode; to switch to the general mode, the user simply clicks the Configure button, scrolls to the last parameter ('Introns in

*To whom correspondence should be addressed. Tel: +45 8942 3125; Fax: +45 8942 3077; Email: jakobf@birc.au.dk

```
  ** ****** ** ******* ******   ***** * ******* ******* ***  ** *  * ** *
CAGCATGCTGACGAAGCCTTGGACCGCCAXXXCAGGAATCAACCGTAGTGGAATCCAGCTAAGGCACACGGAT--
----ATGCTGACGATGCCTTGGGCCGCCA---CAGGACTGAACCGTAATGGAATCTAGCTAAGGCTTACGGAT--
--GCGTGCTGATGAAGCCTTGGACCGCCA---CAGGAATCAACCGTAGTGGAATCCAGCCGAGCCACATGGCTAC
  +-------------------+             +-----------------------+
```

**Figure 1.** For this sample alignment, two primer regions were identified (marked by +——+), using parameters minimum primer length = 18 and maximum number of ambiguities = 4. An asterisk symbolizes a perfect match, i.e. a column with at least 2 nt which are all identical. Gaps are ignored. The X'es in the first sequence indicate an intron of length <200 nt.

sequences') and sets it to 'no', before uploading a data file. In general mode, PriFi does not require that primer pairs span an intron.

From the algorithm's point of view, a primer is initially simply a subsection of the given alignment, i.e. a start index and an end index. Hence, a primer pair is given by four indices. A valid primer pair must fulfill a large set of requirements (see below), and checking a primer pair for these requirements takes the algorithm a small yet non-negligible amount of time. For any alignment of realistic length, there is an astronomical number of ways to randomly pick four indices, and so a naïve algorithm which tests all possible primer pair combinations for all requirements would be very slow.

We, therefore, apply several filters to the full set of potential primer pairs, checking some of the requirements along the way, eventually arriving at a much smaller set of pairs which need to be checked for the remaining requirements. There are three levels of filtering as shown in Figure 2: the first filter operates on the complete alignment by delimiting the regions within which individual primers are searched for; the second filter operates on such individual primer candidates, and the third on candidate pairs of primers.

The first filter identifies the conserved regions of the alignment. Since we only want primers in conserved regions, there is no need to look elsewhere. We identify conserved regions by masking out those alignment columns that cannot be part of a valid primer. In intron mode, the filter first forces a 'safety zone' around the introns by masking those columns that contain intron symbols or that are close to an intron.

Less conserved regions contain many mismatch columns, and the trouble with those is that a primer partly based on a mismatch column will be degenerate; in other words, a mismatch column in the alignment induces an ambiguity in any primer spanning this position. Therefore, we look at mismatch columns to use some of them as delimiters between conserved regions. All valid primers must have a minimum length ($l$) and a maximum number of ambiguities ($a$). For each mismatch column, we check whether it is possible to place a window around it of length at least $l$ such that the part of the alignment covered by this window has at most $a$ mismatch columns. If no such window can be found, the column cannot be part of a valid primer, and it is masked out. After this masking procedure, the conserved regions are identified as those regions which have a length of at least $l$ and contain no masked columns. The resulting sets of alignment regions are called the primer regions (Figure 1).

Within each primer region, all possible primer candidates (i.e. all subsections within the minimum and maximum primer length requirements) are evaluated. The division of the alignment into primer regions dramatically reduces the number of primer candidates. For example, imagine an alignment of

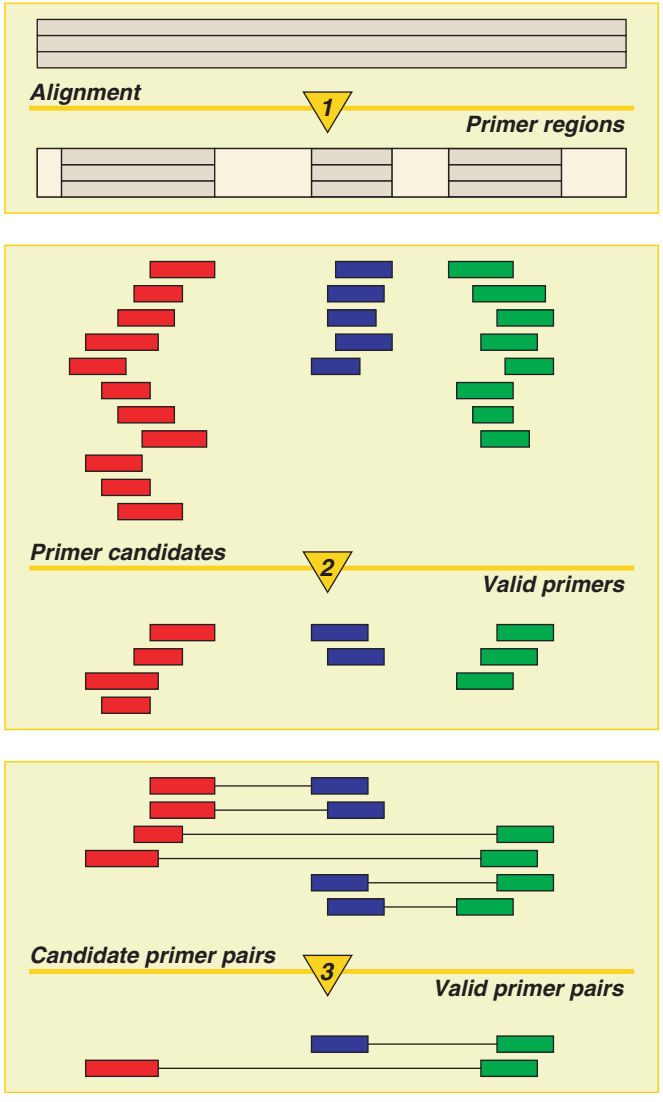


**Figure 2.** PriFi finds valid primer pairs using three filters (shown as three yellow triangles). The first filter operates on the alignment identifying highly conserved primer regions. The second filter identifies the individual primers within the primer regions, evaluates them according to certain criteria and discards those that are invalid. The third filter considers, evaluates and scores all possible pairings between valid primers, discarding invalid pairs.

length 43 and a minimum and maximum primer length of 18 and 35, respectively. This alignment would house 9 different primer candidates of length 35, 10 of length 34, 9 of length 33, etc., all the way down to 26 candidates of length 18, totalling 315 primer candidates to evaluate. If the middle column of this alignment could be masked out leaving two primer regions of

length 21 each, the total number of primer candidates would drop to only 20: in each region, there would be 4 candidates of length 18, 3 of length 19, 2 of length 20 and 1 of length 21 (Figure 3).

The second filter operates on these individual primer candidates to avoid keeping too many for further consideration, while still keeping the best (Figure 2). For example, one of our criteria for valid primers is that they do not end in an ambiguity. At this point, a primer candidate might still serve both as a forward and reverse primer, so we do not yet know its direction, but still we can eliminate those primers which have ambiguities at both ends. Furthermore, we can calculate a primer's estimated melting temperature (1,2), count its number of ambiguities, analyze mismatch diversity (the number of different nucleotides in each mismatch column), check its distance to the nearest intron in both directions and calculate the average number of sequences that it is based on. We have several conditions on this last parameter: for example, if a primer is based on an alignment subsection in which only two sequences are represented, it can have at most two ambiguities, and the subsection must be extraordinarily well conserved. If at least three sequences are represented in (most of) the subsection, more ambiguities are allowed. Thus, PriFi is more likely to find primers in alignments with at least three sequences.

After checking these and other criteria and eliminating all invalid primer candidates, we perform a pruning step on the remaining set of valid candidates, a step which reduces the overall algorithm time by several orders of magnitude. In a long, conserved region, it is possible to suggest a large number of good primers which overlap. Rather than keeping all such overlapping, for a great part essentially identical, primer candidates, we only keep the superior 'representatives', i.e. if two primer candidates overlap by more than some threshold (which is 10 nt by default but may be set by the user), and one is better than the other in all aspects, the inferior one is thrown out. If not, both are kept.

Those primers (alignment subsections) that pass the second filter are turned into actual primer (consensus) sequences with ambiguity codes in any mismatch column positions. The third and final filter now operates on primer pairs by considering all two-primer combinations. Each pair is checked following all remaining criteria for primer pair validity. Pairing two individual primer candidates means assigning to them an orientation, so now we can check the 3′ end tail for degeneracy and high AT content. We can also check that at least one of the primers has at least a certain distance to the closest intron (to ensure unique identification of the PCR product), the estimated PCR product length and that the primer melting temperatures are not too different.

In general, some of the PriFi's criteria are exclusive: a primer (pair) not meeting these criteria is discarded. Other criteria are graduated. Primer pairs are scored (rewarded/penalized) according to a number of such criteria, each with an optimum value that may be modified by the user. They include: primer distance to nearest intron, AT content in the 3′ end tail, average number of sequences in the alignment subsection which the primer is based on, degeneracy, melting temperature and others. Evaluation of self-complementarity is currently not supported. While we took inspiration from (2), PriFi is first and foremost an attempt to capture the, to some extent, intuitive yet successful practice of our laboratory for primer design, and here, self-complementarity is not taken into account (see Results).

The third filter discards all invalid primer pairs (Figure 2), tallies the scores of the valid pairs and ranks them. Four of the top primer pairs are reported. However, the four highest ranking pairs are typically combinations of the same two or three forward and reverse primers, so to avoid redundancy and promote diversity among the suggested primer pairs, we report primer pairs following this strategy. First, the overall best scoring pair is reported. Then, the highest scoring pair whose primers do not overlap by more than a certain number of nucleotides (default 10) with the already reported pair is reported, and so forth. Moreover, if one of the primers in a pair overlaps with two of the primers already reported, this pair is not reported.

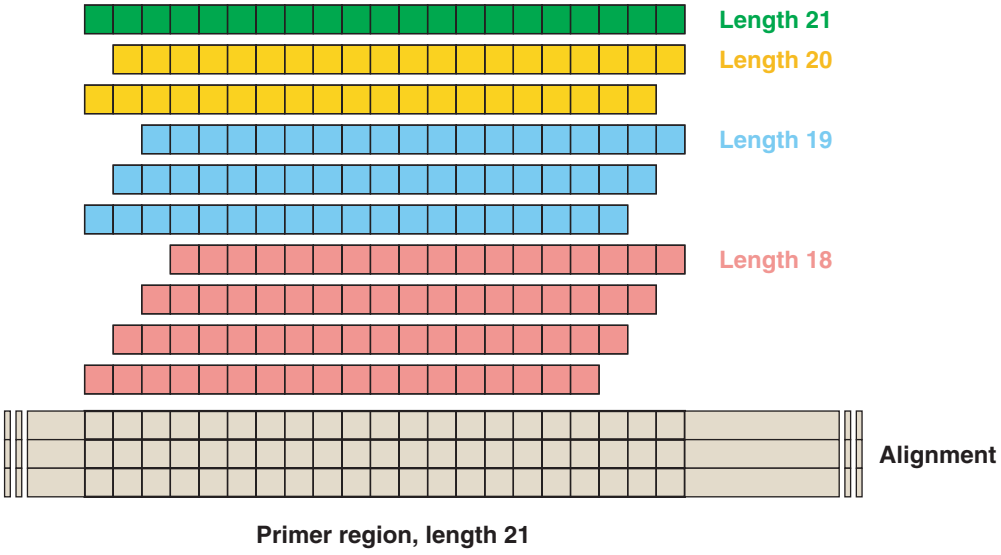


**Figure 3.** With a minimum primer length of 18, one can place 10 different primers in a primer region of length 21: four of length 18, three of length 19, two of length 20 and one of length 21.


**Figure 4.** In the line overview at the top, the input sequences are the black lines while the suggested primer pairs and the corresponding PCR products each have their own, different color. The primers are the thickened ends of the lines. The sequence view below provides a close-up look at the alignment and the primers.

A primer report lists the forward and reverse primers and their characteristics, and specifies the obtained score. Here is an example:

```
Forward: 5′-ATCCGATTTCGAGAAATGCAAACCCTG-
         GTTGATCC
Reverse: 5′-CCCTTCACAGTGGTGATACACTTTCGC-
         TTGTTACG
Tm = 66.4/66.9
Primer lengths: 35/35
Avg. sequences in primer alignments: 3.0/2.0
Estimated product length: 1785
Primer/intron distances: 36/88
A/T's among last 8 bp of 3′ end: 4/5
Ambiguities: 0/0
93.2: High-Tm bonus
6.0: Forward primer length
6.0: Reverse primer length
24.7: Bonus for sequences in primer alignments
3.0: Forward has G/C terminal in 3′ end
3.0: Reverse has G/C terminal in 3′ end
60.0: Good product length
−5.0: Reverse in unconserved region or
based mostly on two sequences
−11.3: Primer/intron distance(s) outside
70–150 bp
−3.0: Too high AT content in 3′ ends
Score: 176
```

The middle part lists certain quantitative traits (the 'average number of sequences in primer alignment' is the average number of nucleotides per column in the alignment subsection corresponding to the primer. The 'primer/intron distance' is the distance from each primer to the closest intron inside the PCR product). The last part gives the score and lists the constituent terms, e.g. a high melting temperature contributes greatly, and a reward is also given to primers based on alignment subsections with many sequences represented (the more sequences that are involved, the more can you trust the conservedness of a conserved region). Expected PCR product length and primer/intron distances also contribute, positively or negatively, to the score. The primer report shows on what grounds the primer pair was selected. Presented with up to four primer pair reports, the user can make an informed choice.

All penalties and rewards may be fine-tuned, and when configuring PriFi, the user can click on each parameter to get an explanation. In intron mode, PriFi requires all primer pairs to span at least one intron. If the user configures PriFi to run in the general mode, PriFi expects no intron symbols in the sequences, and all criteria regarding introns become void. Any X'es in a sequence then has no special meaning.

After the analysis completes (a matter of a few seconds for four input sequences), the suggested primers are displayed in three ways: as lines in the alignment line overview, as letter sequences in the sequence view and as a textual explanation for the obtained score. Each primer pair is shown in its own color (the same color in all three displays), ensuring clarity and easy distinction between them, while input sequences are shown in black. Alignment match columns are highlighted in olive green, and introns are colored lilac (Figure 4). Thus, the user can make an informed choice between the suggested primer sets based on their location in the alignment and their score reports which summarize their characteristics.

## RESULTS

Based on multiple alignments of clustered legume expressed sequence tags from the two model legume species *Lotus japonicus* and *Medicago truncatula* (medic barrel) and the crop *Glycine max* (soybean) extracted from the TIGR database (3,4), we have used PriFi to design primers to amplify orthologous sequences in *Phaseolus vulgaris* (common bean) which is relatively closely related to the 'founder' species, and also in different species of *Arachis* (groundnut) which is phylogenetically much more distantly related. Exhaustive testing of 36 primer sets with scores between 187 and 98 using PriFi's default criteria for primer design identified 24 correct products in bean and 19 in groundnut. Survey testing has identified correct products for primers with scores as low as 59. In most cases, the obtained fragment sizes and default values for distances from primers to introns were sufficient to unambiguously identify the product and discover nucleotide polymorphisms within a single sequence run for each mapping parent. Using the Oligo Calculator by Qing Cao, Warren and Buehler (http://www.basic.nwu.edu/biotools/oligocalc.html), we found that ∼10% of the primers had significant regions of self-complementarity that might in theory result in self-priming during PCR. However, all these primers have worked well in the laboratory.

## DISCUSSION

Other programs exist which automate primer design [e.g. (5–8) (Primers can be found at http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi and DoPrimer at http://doprimer.interactiva.de/)], but usually they require full genomic information from a species, or they simply find optimal primers for a known target in a given sequence. The Codehop program (9) finds primers for amplifying unknown targets but works with protein sequence alignments.

The automated identification of well-conserved regions in sequence alignments and design of primers that are likely to amplify orthologous sequences even in distantly related species eliminate one of the major time-consuming steps in the process of developing PCR-based DNA anchor markers that can be used to interconnect the genetic maps developed for related species (10). The use of such anchor markers will improve our understanding of syntenic relationships between related species and help to optimize the exploitation of genetic map resources by enabling information transfer between species (11). We are currently implementing a general software pipeline based on these ideas.

## SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

## ACKNOWLEDGEMENTS

The authors thank the Danish Agricultural and Veterinary Research Council for support, including the funding to pay the Open Access publication charges for this article.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Sugimoto,N., Nakano,S., Yoneyama,M. and Honda,K. (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res*., 4501–4505.
2. Burpo,F.J. (2001) A critical review of PCR primer design algorithms and cross-hybridization case study. Biochemistry 218 at Stanford University. Available at http://cmgm.stanford.edu/biochem218/Projects%202001/Burpo.pdf.
3. Pertea,G., Huang,X., Liang,F., Antonescu,V., Sultana,R., Karamycheva,S., Lee,Y., White,J., Cheung,F., Parvizi,B. *et al*. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.
4. Quackenbush,J., Cho,J., Lee,D., Liang,F., Holt,I., Karamycheva,S., Parvizi,B., Pertea,G., Sultana,R. and White,J. (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res*., **29**, 159–164.
5. Haas,S., Vingron,M., Poustka,A. and Wiemann,S. (1998) Primer design for large scale sequencing. *Nucleic Acids Res*., **26**, 3006–3012.
6. McKay,S.J. and Jones,S.J. (2002) AcePrimer: automation of PCR primer design based on gene structure. *Bioinformatics*, **18**, 1538–1539.
7. van Baren,M.J. and Heutink,P. (2004) The PCR Suite. *Bioinformatics*, **20**, 591–593.
8. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol*., **132**, 365–386.
9. 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.
10. Lyons,L., Laughlin,T., Copeland,N., Jenkins,N., Womack,J. and O'Brien,S. (1997) Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian sequences. *Nature Genet*., **15**, 47–56.
11. Choi,H.-K., Mun,J.-H., Kim,D.-J., Zhu,H., Baek,J.-M., Mudge,J., Roe,B., Ellis,N., Doyle,J., Kiss,G.B., Young,N.D. *et al*. (2004) Estimating genome conservation between crop and model legume species. *Proc. Natl Acad. Sci. USA*, **101**, 15289–15294.