Published online 28 April 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W209–W213***
doi:10.1093/nar/gkp269

UniPrime2: a web service providing easier Universal Primer design

Robin Boutros[1], Nicola Stokes[1], Michaël Bekaert[2] and Emma C. Teeling[2,*]

[1]UCD School of Computer Science and Informatics and [2]UCD School of Biology and Environmental Science & UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland

Received January 31, 2009; Revised April 3, 2009; Accepted April 11, 2009

ABSTRACT

The UniPrime2 web server is a publicly available online resource which automatically designs large sets of universal primers when given a gene reference ID or Fasta sequence input by a user. UniPrime2 works by automatically retrieving and aligning homologous sequences from GenBank, identifying regions of conservation within the alignment, and generating suitable primers that can be used to amplify variable genomic regions. In essence, UniPrime2 is a suite of publicly available software packages (Blastn, T-Coffee, GramAlign, Primer3), which reduces the laborious process of primer design, by integrating these programs into a single software pipeline. Hence, UniPrime2 differs from previous primer design web services in that all steps are automated, linked, saved and phylogenetically delimited, only requiring a single user-defined gene reference ID or input sequence. We provide an overview of the web service and wet-laboratory validation of the primers generated. The system is freely accessible at: http://uniprime.batlab.eu. UniPrime2 is licenced under a Creative Commons Attribution Noncommercial-Share Alike 3.0 Licence.

INTRODUCTION

The process of identifying conserved stretches of a genomic sequence among phylogenetically distinct taxa, in which to design PCR primers to amplify homologous markers in other species, was developed in the 1990s (1). This has proven to be a highly successful method and has resulted in many comparative studies that address the evolution of key genes in non-model organisms (2,3). However, to use this method a research scientist must have some degree of bioinformatic capability. Firstly, the marker of choice from the appropriate taxa needs to be retrieved from the ever-expanding databases (e.g. GenBank (4)) and then aligned either by eye or by using freely available software, e.g. ClustalW (5), T-Coffee (6) or GramAlign (7). Within this resulting alignment, highly conserved regions must be identified and then some consensus of this alignment is used to design a primer pair, taking melting temperature (T_M), GC content, secondary structure, length of amplicon and primer-self end complementarity into account (1). The success of employing these universal primers depends on the alignment quality, ability to locate conserved regions and a low potential probability of mis-priming (i.e. probability of amplifying a paralogous stretch of the genome).

As molecular biology is now firmly integrated into the fields of ecology, behaviour, conservation science, the ability to design optimal primers for comparative gene studies is becoming more important. As more genomes are annotated and collated, non-computational biologists can identify any system of interest (e.g. echolocation) that they believe to be under selection in their study species (e.g. bats) and design the appropriate primers to further investigate this question (2). Biological studies are becoming highly multi-disciplinary and the tools and skills required to mine these vast databases are increasing in usability and number. However, as the information contained in the databases is expanding at an exponential rate it is now becoming impossible to manually implement all steps required for primer design (8).

To enable non-computational biologists to design universal primers (for any marker or taxa) that have been proven to successfully work in the laboratory we originally designed UniPrime (8). This is an open-source suite of integrated software programs that allows users to automatically design sets of universal primers to amplify regions of suitable inter-specific variation across divergent taxa by simply inputting a NCBI GeneID or accession number. UniPrime is significantly different from other primer design programs in that it uniquely allowed all steps of the process to be saved, including initial data retrieval, multi-species alignment and primer design sites (8). However, the original version of UniPrime required the local installation of BioPerl (9), Primer3 (10,11),

*To whom correspondence should be addressed. Tel: +353 1 716 2263; Fax: +353 1 716 1152; Email: emma.teeling@ucd.ie

T-Coffee (6) and PostgreSQL (www.postgresql.org) plus the UniPrime database. Each program had to interact with the UniPrime database and 'trouble-shooting' the installation and implementation problems of this suite of programs were platform specific. This has proven difficult and frustrating for non-computational biologists, the target audience. Other concerns and limitations of the initial UniPrime software was that the only allowable input format of reference sequence was a GeneID. This limited researchers to use only the annotated databases in their initial search. An additional bottleneck in the original version of UniPrime was the use of the T-Coffee (6) alignment program. Alignments of over eight taxa took a significant amount of time, and thus significantly increased the duration of the primer design process.

To overcome the installation issues and limitations of Uniprime (8), we have now developed UniPrime2 (http://uniprime.batlab.eu), which is a user-friendly web service (rather than a downloadable package) with many novel system enhancements. These enhancements have resulted in an almost 10-fold increase in average efficiency per query, advanced search capabilities and the local installation problems of the early version of UniPrime have been alleviated.

IN SILICO METHODS

The UniPrime2 algorithm

From the original algorithm described in (8), we have improved and added several alternative steps.

Step 1: locus. Users can now input a Fasta file format of any sequence that they wish. Additionally, they can input a GenBank GeneID (unique species specific identifier for genes provided by Entrez Gene) of the target locus.

Step 2: orthologs. The initial input sequence is used as a 'query' for a Blastn (12) search of the NCBI database to identify highly similar homologous sequences. The user can delimit the search at varying phylogenetic levels by incorporating the Entrez Query BLAST option. When a Fasta sequence file has been inputted any NCBI database can be specified and searched by the user. When the user defines a GeneID only the RefSeq mRNA database will be searched for homologous sequences, insuring that the retrieved sequences are annotated.

Step 3: alignment. Users can now choose to align their sequences with the GramAlign (7), or the T-Coffee (6) algorithm. GramAlign is recommended because of its speed, and we have validated the authenticity of the primers designed using this new functionality (see 'Results' section). From the alignment, a consensus sequence is inferred.

Step 4: primer design. All possible primers along the consensus sequence are generated by Primer3. (11).

Step 5: virtual PCR, (optional). The primer sequences are submitted for a Blastn search within the 'Whole-genome shotgun reads' (wgs) database of GenBank to identify sequences that match the forward and reverse primer sequence within a compatible size range for PCR amplification (primers <10 kb apart).

Implementation

The UniPrime2 web service is running on an Apache server (www.apache.org) using PHP (www.php.net) to link HTML forms with Perl scripts. The following modules from BioPerl package (9) were used in the implementation of UniPrime2: bioperl, bioperl-run and bioperl-ext. The GD graphics package (www.libgd.org) is used to view sequences. UniPrime2 uses the PostgreSQL (www.postgresql.org) database management system to temporally store information relating to user requests. T-Coffee (6) and GramAlign (7) are used for alignment, while Primer3 (10,11) is used to generate primers.

WET LABORATORY METHODS

PCR and DNA sequencing

PCR was performed with 2 nM of each primer (see 'Experimental validation' section), 1.5 mM MgCl2, 1 U of Platinum *Taq* DNA polymerase (Invitrogen Corporation, Carlsbad, California, USA) and 10 ng of genomic DNA. Touchdown conditions of amplification were used for all species, as follows: 10 cycles of denaturation at 95°C for 30 s, annealing at 65°C for 30 s –1°C per cycle, extension at 72°C for 60 s; followed by 35 cycles with 95°C for 30 s, annealing at 55°C for 30 s, extension at 72°C for 60 s. The initial denaturation step and the last extension step were 3 min each. The PCR products were separated and visualized in 1% agarose gel. Sequencing reactions were performed in both directions on PCR products, using the same primer set as for amplification.

RESULTS

Experimental validation

We used UniPrime2 to generate primers from four diverse genes (AOF2, EFEMP1, LRP6 and OAZ1) that were originally examined in (8). We found that the same set of optimal primer pairs where designed when either GramAlign or T-Coffee were used by UniPrime2. In a second step, we evaluated the ability of UniPrime2 to design a new set of primers for two other genes ATG16L1 (Autophagy 16-like 1) and APOBEC3G (Apolipoprotein B mRNA-Editing Enzyme, Catalytic Polypeptide-like 3G). We validated the primers designed by amplifying and sequencing fragments from these genes in four divergent Orders across Class Mammalia. The genomic DNA from five divergent mammal species was used (Supplementary Data), for the two genes selected (Table 1). The generated sequences were deposited in GenBank (FJ796956-FJ796961)

Efficiency

As already stated, one of the principal design enhancements of the new UniPrime2 web server is the inclusion of the GramAlign algorithm (7) as an alternative sequence

Table 1. Primers used and estimated product length

Gene	GeneID		Primers (5′ -> 3′)	Size	Location (human)
ATG16L1	55054	F	GGATCTACGCAGCAAAGTCTG	613	Exon 1
		R	NTCGNATGTCCCAGAAACG		
APOBEC3G	60489	F	GACNGCATGAGACTTACCTGTG	2678	Exon 5–6
		R	GNGAAGATGCACANGCTCAC		

For each gene/locus the forward (F) and reverse (R) primers are indicated. The 'GeneID' was the entry used for the initial step. The expected product length is an average value inferred from the multiple alignments but varies between species.

alignment solution to the popular, but computationally expensive, T-Coffee algorithm (6). Table 2 shows typical average runtimes for the six genes described above using two UniPrime2 configurations, T-Coffee and GramAlign and searching the NCBI RefSeq mRNA database. There is approximately a 10-fold decrease in total UniPrime2 execution time per user request when T-Coffee is replaced with GramAlign. There is no significant increase in execution time efficiency when Greenwich Mean Time (GMT) intervals coincide with the peak access times of NCBI tools and databases (GenBank and BLAST) required by UniPrime2 (Table 2).

Figure 1 depicts the variation in runtime per request at each stage in the UniPrime2 processing pipeline when GramAlign is used. Steps 2 and 4 are now the most time consuming components in the execution of a request. This is because at both these steps the system is remotely accessing NCBI resources. The final and optional processing step is the virtual PCR process (not shown in Figure 1). Typically, the virtual PCR of a primer pair will take in the order of hours rather than minutes to complete, because each of the primer sequences are submitted during this process to a time-consuming Blastn search within the GenBank databases. However, this step allows users to retrieve all possible homologous sequences and estimate the potential of mis-priming in the laboratory.

Table 2. Average runtime of primer design request on the UniPrime2 web server

Time (GMT)	UniPrime + TCoffee	UniPrime + GramAlign
Mon. 12 pm	3237 s	322 s
Tues. 3 pm	3617 s	356 s
Wed. 10 am	3728 s	406 s
Thu. 3 pm	3329 s	343 s

Results of the average runtime in seconds per primer design request on the UniPrime2 web server at varying time intervals, when T-Coffee (UniPrime + T-Coffee) or GramAlign (UniPrime + GramAlign) is used for sequence alignment.

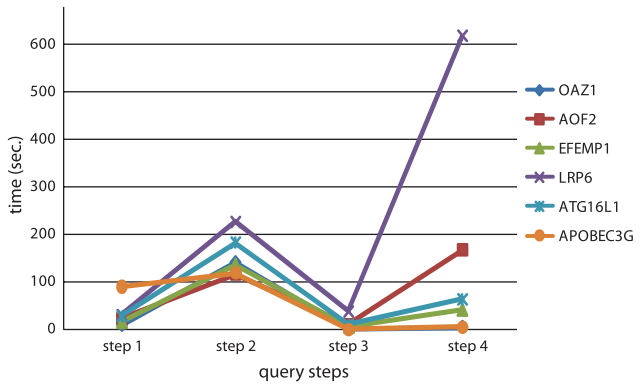

Figure 1. Runtime for each step in the UniPrime2 processing pipeline for different genes.

UniPrime2: the web server

Step 1: locus. All user inputs, such as locus name and other parameters, are now collected though a HTML form on the UniPrime2 web server. Previously, users had to interact with the system through a complicated sequence of command line requests to a local installation of the system. UniPrime2 now accepts a user-defined input sequence in Fasta format, as well as the previously supported GeneID. There are links and tutorials to enable users to find a GeneID and examples of a Fasta file input.

Step 2: orthologs. A species and taxon name database is now used to provide quick and easy spellchecking and species name prompting. When searching with a Fasta input query users can search any of the Genbank databases, which appear as a drop-down menu, this allows for greater flexibility when using non-annotated genomes or taxa. If required, users can select the method of alignment of the retrieved sequences, the threshold of the consensus and amplicon size and directly design primers. They can also easily modify the e-value threshold at this stage, or any of the other default parameters if they feel they have sufficient expertise. If no homologous sequences are retrieved, the user can avail of the primer design 'trouble-shooting' tutorial on the website.

Step 3: alignment. Users are given the option of choosing T-Coffee or GramAlign. Users can review the taxa that have been retrieved from Genbank at this stage and easily click on which taxa to include or not. They can also readily change the consensus threshold by using a drop-down menu at this stage or modify the parameters using the 'advanced parameters' option.

Step 4: primer design. Users can review the consensus sequence used to generate the primers, and can easily download the alignment file. All primers designed can be downloaded as a spreadsheet that includes T_M, penalty scores and location in the alignment. The size of the amplicon for each primer pair is depicted on the website. If no primers are designed advice can be sought in the

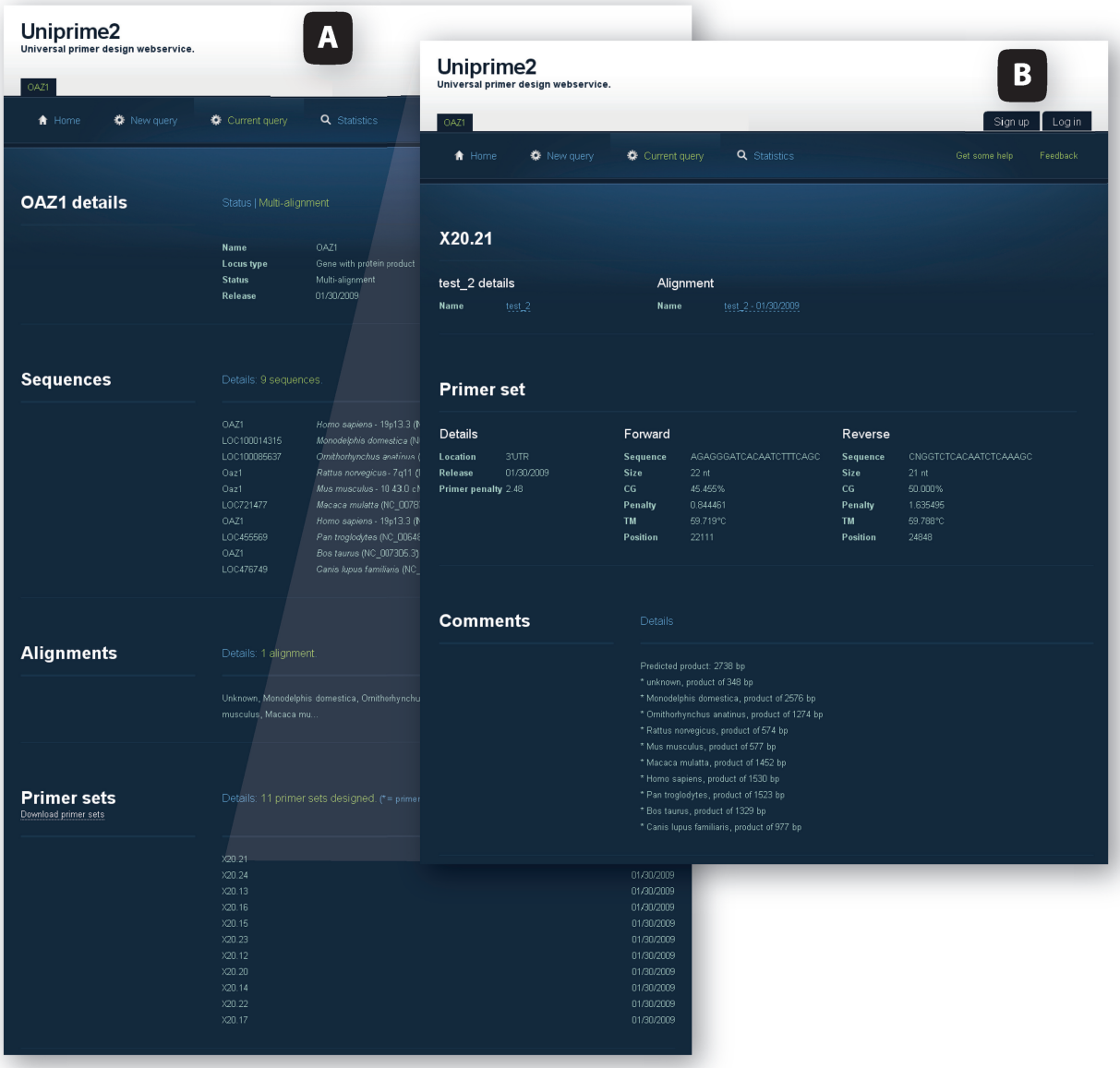

Figure 2. Screenshots of the UniPrime2 interface. Screenshot (**A**) is the main results page for the OAZ1 locus, while (**B**) shows click-through details for one of the generated OAZ1 primer pairs, X20.21.

'trouble-shooting' section of the website. Figure 2 depicts a screenshot of the resulting output from these four steps, where the user can view the gene, the retrieved sequences, the alignment and the primers designed.

Step 5: virtual PCR. User can select which primer pair they would like to use to search the Genbank databases. The results outputted are an aligned file of all possible sequence retrieved by the Blast search.

In all cases, primer generation can now be run in parallel enabling users to execute concurrent requests. Each request can be accessed by clicking on the appropriate tab at the top of the page. Tabs are labelled with the NCBI GeneID or Fasta sequence name used in the request. Users now have access to their previously submitted requests and primer results. There is an ongoing 'ticker-tape' that describes what stage of primer design UniPrime2 is involved in. An e-mail alert is now sent to the user when their request has been processed. Users can now view outputs at various steps of the processing pipeline, and modify a wide variety of parameters where appropriate. This more interactive setup helps users to identify and resolve situations where highly divergent conservation regions result in low primer quality.

CONCLUSION

UniPrime2 represents a new generation of integrated user-friendly, computational solutions to primer design that anyone can use. It seamlessly harnesses the ever-growing masses of genomic data and allows users to design

functional primers without expert bioinformatic skills. It is a tool that will greatly aid the multi-disciplinary modern biologist.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The provision of the server by University College Dublin Research IT Services is gratefully acknowledged.

FUNDING

Science Foundation Ireland PIYRA 06/YI3/B932 awarded to E.C.T. Funding for open access charge: Science Foundation Ireland PIYRA 06/YI3/B932.

Conflict of interest statement. None declared.

REFERENCES

1. Murphy,W.J. and O'Brien,S.J. (2007) Designing and optimizing comparative anchor primers for comparative gene mapping and phylogenetic inference. *Nat. Protoc.*, **2**, 3022–3030.
2. Li,G., Wang,J., Rossiter,S.J., Jones,G., Cotton,J.A. and Zhang,S. (2008) The hearing gene Prestin reunites echolocating bats. *Proc. Natl Acad. Sci. USA*, **105**, 13959–13964.
3. Song,B., Gold,B., O'Huigin,C., Javanbakht,H., Li,X., Stremlau,M., Winkler,C., Dean,M. and Sodroski,J. (2005) The B30.2(SPRY) domain of the retroviral restriction factor TRIM5alpha exhibits lineage-specific length and sequence variation in primates. *J. Virol.*, **79**, 6111–6121.
4. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.
5. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
6. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
7. Russell,D.J., Otu,H.H. and Sayood,K. (2008) Grammar-based distance in progressive multiple sequence alignment. *BMC Bioinform.*, **9**, 306.
8. Bekaert,M. and Teeling,E.C. (2008) UniPrime: a workflow-based platform for improved universal primer design. *Nucleic Acids Res.*, **36**, e56.
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.
10. Koressaar,T. and Remm,M. (2007) Enhancements and modifications of primer design program Primer3. *Bioinformatics*, **23**, 1289–1291.
11. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
12. Ye,J., McGinnis,S. and Madden,T.L. (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res.*, **34**, W6–W9.