PrimerZ: streamlined primer design for promoters, exons and human SNPs

Ming-Fang Tsai[1], Yi-Jung Lin[1], Yu-Chang Cheng[1], Kuo-Hsi Lee[1], Cheng-Chih Huang[2], Yuan-Tsong Chen[1,2] and Adam Yao[1,2,*]

[1]National Genotyping Center (NGC), Academia Sinica, Taipei, Taiwan 11529, R.O.C. and [2]Institute of Biomedical Sciences (IBMS), Academia Sinica, Taipei, Taiwan 11529, R.O.C.

ABSTRACT

PrimerZ (http://genepipe.ngc.sinica.edu.tw/primerz/) is a web application dedicated primarily to primer design for genes and human SNPs. PrimerZ accepts genes by gene name or Ensembl accession code, and SNPs by dbSNP rs or AFFY_Probe IDs. The promoter and exon sequence information of all gene transcripts fetched from the Ensembl database (http://www.ensembl.org/) are processed before being passed on to Primer3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi) for individual primer design. All results returned from Primer 3 are organized and integrated in a specially designed web page for easy browsing. Besides the web page presentation, csv text file export is also provided for enhanced user convenience.

PrimerZ automates highly standard but tedious gene primer design to improve the success rate of PCR experiments. More than 2000 primers have been designed with PrimerZ at our institute since 2004 and the success rate is over 70%. The addition of several new features has made PrimerZ even more useful to the research community in facilitating primer design for promoters, exons and SNPs.

INTRODUCTION

Polymerase chain reaction (PCR) is a commonly used laboratory procedure nowadays for a variety of tasks, such as DNA cloning, sequence determination and SNP detection. Consequently, numerous primers need to be designed for DNA amplification in order to produce enough DNA.

For DNA sequencing, to design primers for the promoter and exon regions of a gene, one needs to retrieve the required sequence information for each single exon and promoter, including their corresponding flanking sequences, convert them to the correct format, switch to a primer design application like Primer 3 (1), import the sequence information and adjust parameter settings if necessary before submitting the request. While this is tolerable for primer design for a gene that has a single exon, a human, mouse or rat gene can easily have more than 10 exons, and each of the above steps may thus need to be repeated 10 times or more. Furthermore, if an exon is too long to sequence properly in one run, several primers have to be designed to map overlapping sections of the exon. The whole process requires many manual steps for repeated window opening, browsing, application switching, copying and pasting, typing and so on, which can be tedious, time-consuming and error prone. In fact, for a gene with 10 exons, the process can easily exceed 300 steps.

To improve the situation, we have developed PrimerZ to replace almost all of these manual steps so that users can easily complete a primer design task using just a few clicks for a gene or batched human SNPs. More than 2000 primers have been designed with PrimerZ at our institute since 2004 and the success rate is over 70%.

PrimerZ accepts either gene names or SNP IDs (2). There are a multitude of user-definable options available including product size, maximum exon length, excluded regions of the query sequence, GC-content and maximum allowable local alignment score. Simply by submitting a candidate gene name, a SNP ID or up to 100 batched SNP IDs, in most cases all the primer sequences should be returned in a minute or two. The generated information, including gene transcript graph, primer data from Primer3, and direct links to UCSC In-Silico PCR (3,4) for PCR product prediction, to NCBI Blast and to Ensembl source data, is integrated and displayed on a single page for convenient viewing. Additionally, all the primer data can be exported in csv format for further processing.

IMPLEMENTATION

PrimerZ takes advantage of the well-developed public-domain database Ensembl, through its API (application programming interface). When a gene query is received,

*To whom correspondence should be addressed. Tel: +886 2 26523091; Fax: +886 2 27824066; Email: adam@ibms.sinica.edu.tw

PrimerZ will access the Ensembl database through ENSj API (5) to retrieve promoter and exon information.

By default but with an adjustable setting, PrimerZ retrieves one region of 1440 bp upstream from the start of 5′-UTR which it treats as the promoter region. The promoter region is thereafter divided into four 360 bp non-overlapping segments plus their respective flanking sequences. The default length of flanking sequence is 240 bp. All exons are directly flanked with 240 bp sequences, except when an exon of the gene is >360 bp, when the sequence will be split into segments of ≤360 bp for a better quality sequencing result. For example, an exon of 1000 bp will produce two 360 bp segments and one 280 bp segment.

The above sequence information, plus parametric settings packaged by PrimerZ are then fed into Primer3 for primer design. All returned primer results, together with a transcript graph, are integrated into a one-page report for final output. The self-explanatory workflow of the whole PrimerZ system is shown in Figure 1.

The software development environment included the following software:

 (i) Java, JDK: j2sdk1.4.2_06, Server VM
 (ii) Struts Framework 1.2
(iii) Red Hat Enterprise Linux Academic Edition
(iv) MySQL 4.1
 (v) Tomcat 5.0 web server
(vi) Ensembl Java API

The API provided by Ensembl is used for gene information retrieval. JAVA language is used to pipeline gene data into Primer 3 and merge the returned results.

Figure 1. PrimerZ workflow.

INPUT

PrimerZ is an easy-to-use tool to design primers for genes and SNPs, using only a few simple steps to design the wanted primers. It currently provides gene primer design for all Ensembl species while SNP primer design is currently only available for human SNPs. For gene primer design, there are some essential parameters required for optimal design of primers, such as maximum exon length, exon flanking region, product size range and excluded region. Users can modify these parameters or even full range Primer 3 parameters under 'Advance Options' according to the particular requirements of their experiments.

Exons with length larger than the 'maximum length' parameter will be fragmented. The 'excluded region' value allows the program to bypass regions with low sequence quality or containing repetitive elements such as ALUs or LINEs for primer design. For SNP primer design, a SNP rsID or an AFFY_ProbID(6) as well as a batch file containing mixed ID types can be accepted as input, where the maximum number of SNPs per batch is limited to 100.

OUTPUT

The result page comprises three major parts: initial input data and parameters, the gene transcript diagram
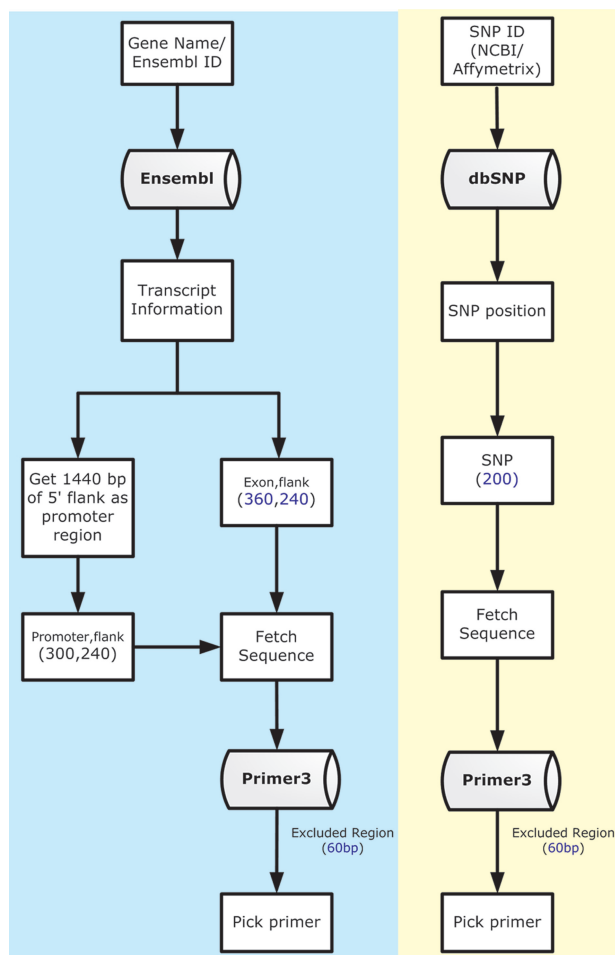
(only available for gene primer design) and designed primer information. The first part lists all the input data and parameters plus Ensembl Database version, for ease of reference. The second part shows each promoter fragment and exon of a transcript diagram, and links to their accompanying original Primer3 output. Finally, the third section presents tabulated primer information and incorporates executable links to NCBI blast and UCSC In-Silico PCR to check specificity and product prediction. In addition, an 'Ensembl Link' button offers the Ensembl Exon Report of the transcript so that a user can trace all primers to their original sequence information. A csv format text file of the results can be downloaded at this results page.

BENCHMARK

In a manual operation benchmarking test, it took a person very familiar with all processes about 1260 s and 380 discrete steps to design the primers of HADHSC (L-3-hydroxyacyl-Coenzyme A dehydrogenase, short chain), our benchmark gene with 10 exons. With automated PrimerZ, the same operation took only 16 s and three steps. This comparison clearly demonstrates

the superior efficiency and ease-of-use of PrimerZ for gene primer design. A benchmark test on batched SNPs showed a similar dramatic reduction in workload from 90 discrete steps to two steps, from 840 to 49 s, and from 40 result pages to a single page.

CONCLUSION

PrimerZ has been designed to obtain reliable primers for PCR experiments and to allow standardized, automated primer design for batch operation. Users can access the UCSC In-Silico PCR directly from the result page to verify their primers to achieve higher accuracy and lower cost. PrimerZ also allows users to modify the conditions of primer design, including the maximum exon size, the flanking region of the target sequence, the exclusion region and the maximum allowed polyA and CA-repeats in the PCR products. In addition, PrimerZ will offer primer design from NCBI transcripts in the near future, which should be of great interest to those users who use NCBI data to design primers. The results from NCBI and Ensembl will be shown in the same page. Following the release of Primer3 Web Interface in November 2006, we are installing and testing a local copy of Primer3 Web to alleviate the burden on the original Primer3 website and the restriction on the number of SNPs allowed.

Primer Z is a simple-to-use program that greatly facilitates and enhances the traditionally time-consuming task of accurate primer design for PCR, and should be an excellent additional tool for the modern molecular biologist.

REFERENCES

1. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Meth. Mol. Biol*, **132**, 365.
2. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*., **29**, 308.
3. Hinrichs,A.S., Karolchik,D., Baertsch,R., Barber,G.P., Bejerano,G., Clawson,H., Diekhans,M., Furey,T.S., Harte,R.A. *et al.* (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*., **34**, D590.
4. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res*., **12**, 996.
5. Curwen,V., Eyras,E., Andrews,T.D., Clarke,L., Mongin,E., Searle,S.M. and Clamp,M. (2004) The Ensembl automatic gene annotation system. *Genome Res*., **14**, 942.
6. Lamy,P., Andersen,C.L., Wikman,F.P. and Wiuf,C.P. (2006) Genotyping and annotation of Affymetrix SNP arrays. *Nucleic Acids Res*., **34**, e100.