Short Technical Report

Express Primer Tool for High-Throughput Gene Cloning and Expression

ABSTRACT

High-throughput approaches for gene cloning and expression require the development of new nonstandard tools for molecular biologists and biochemists. We introduce a Web-based tool to design primers specifically for the generation of expression clones for both laboratory-scale and high-throughput projects. The application is designed not only to allow the user complete flexibility to specify primer design parameters but also to minimize the amount of manual intervention needed to generate a large number of primers for the simultaneous amplification of multiple target genes.

INTRODUCTION

The abundance of genomic sequence data from different organisms provides new opportunities for genome-scale projects that focus on the elucidation of the roles of protein structure and function (4,14). This has led to the development of several tools for the design of DNA amplification primers for DNA sequencing (3), sequence tags (15), and additional design applications (6,8). We extend this molecular toolbox by the addition of a publicly available application to design primers specifically for expression cloning. This capability for the generation of large numbers of clones expressing heterologous proteins will be required to fully utilize the data resource provided by the genomic sequence projects. The development of these tools is a critical component of the Protein Structure Initiative (7,10) (http://www.rcsb.org/pdb/strucgen.html) and other large-scale programs that aim to characterize protein structure and function.

The protein production goals of The Midwest Center for Structural Genomics (MCSG) (12) require the implementation of automated methods for the generation of expression clones. To meet the economic constraints of this process, one must minimize the manual labor required for the design of primers for the amplification of the coding regions. The Express Primer Tool we developed enables the users to specify various experimental parameters (e.g., optimal $T_m$, $T_m$ range and maximum $T_m$ difference) for single or multiple candidate sequence(s) in FASTA format input as a flat text (ASCII) file. The application generates candidate primers, selects optimal primer pairs, and writes the forward and reverse primers pairs to a Microsoft® Excel® file that is suitable for electronic submission to a synthesis facility. The program parameters emphasize high throughput but allow for target attrition at various stages of the project. This high-throughput, plate-based approach, while typical of many genome-scale projects, is fundamentally different than standard target-focused approaches that typically severely limit target attrition but often require significantly more manual intervention.

MATERIALS AND METHODS

System Requirements

The Express Primer Tool requires a Web server with a JSP Engine. Installation requirements are dependent on the Web server and the type of JSP Engine used. The tested and recommended server is Apache Tomcat, integrated into an Apache HTTP server (Apache Software Foundation, Forest Hill, MD, USA). The application was implemented on an Intel® Pentium®-based machine, the Dell Precision 330, running Red Hat Linux 7.2® (Red Hat, Raleigh, NC, USA) with a 1.7-GHz processor and 256 MB of RAM.
User Interface

To maximize the dissemination of this application to the research community, the Express Primer Tool was written to allow for implementation as a Web-based application. The Web-based approach provides a multi-user capability that enables researchers to access the program by logging onto a Web site running on a single server rather than each individual downloading and installing the application locally. The Express Primer Tool was written in Java®, a platform-independent language, so that the application could be installed without modification wherever a Java Virtual Machine (available free of charge along with documentation at http://java.sun.com) is installed (i.e., Windows®-based, Unix®-based, and Linux-based machines). Download and installation instructions with complete documentation and additional reference information can be found in the “Resources” section of the MCSG Web page (www.mcsg.anl.gov).

RESULTS AND DISCUSSION

Description of the Express Primer Tool

The Express Primer Tool incorporates standard primer design features (3) that provide considerable experimental flexibility. Although designed for high-throughput methods, the application provides an option “Optimize for Individual Primers” (Figure 1) for the generation of primer pairs for non-plate-based methods. The application uses sequence input as flat text files and generates primers on the basis of user-specified parameters. Sequences that fail to generate acceptable primer pairs are saved in flat file format to enable the user to repeat the process with new primer design criteria. Specific features of the primer design interface (Figure 1) are as follows.

File input. Sequence data are imported as flat text ASCII files containing coding region sequences of desired target protein genes in the FASTA format. The file input format is compatible with the format of nucleotide files that are downloadable from the genomes section of the National Center for Biotechnology Information (NCBI) Web site (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/allorg.html). The application accepts an input file of sufficient size to enable the design of primers for an entire bacterial genome. The application successfully designed primers for a 5000-target gene set using several different combinations of primer design options. Genome files downloaded from the NCBI Web site use multiple formats for each sequence description. The Express Primer Tool uses the first uninterrupted character string as a connotation for the “Gene ID” field in the program output.

Optimal Tm. This selection allows the user to define a relative Tm for primer design with $T_m$ calculated using the following formula (2).

$$T_m = \begin{cases} 2 \times (\text{sequence length} - \text{number of G+C}) + 4 \times \text{(number of G+C)} & \text{for sequences of 14 or longer:} \\ 64.9 + 41 \times \frac{(\text{number of G+C} - 16.4)}{\text{sequence length}} & \text{for sequences of 14 or longer:} \end{cases}$$

Several alternative formulas (1,9,13) have been proposed for use in primer evaluation and could be substituted for the current formula. In practice, we select the annealing temperature based on experience and preliminary screens with each lot of genomic DNA.

Tm difference. This feature specifies the maximum value of $\Delta T_m$ for a primer pair (default value of 5°C).

Tm range. This selection allows the user to specify the maximum range of $T_m$ values considered for the forward and reverse primers.

Primer length definition. Successful primer generation is dependent on the allowable range of primer lengths. Longer sequences tend to have a higher $T_m$, whereas shorter primers are less costly. If the user does not define mini-
mum and maximum primer lengths, then the default minimum length is 16, and the default maximum length is 30.

**Limit sequence length.** For certain projects, it may be desirable to restrict the size range of expressed proteins. Protein size can dictate analysis methods, and large proteins may require different conditions for amplification and analysis. This feature enables the user to select the desired coding region size range needed for their specific application.

**Extra Cs or Gs at end of primer.** Additional G/Cs at the 3′-end of a primer promote annealing and can increase amplification efficiency (3). This option allows the user to increase the representation of G/Cs if they are components of the sequence, even if the shorter primers with fewer G/Cs at 3′-ends meet the design criteria.

Additional features such as the ability to append cloning handles, force an ATG start codon, or exclude start/stop codons in the design process are incorporated into the software and are applicable at all levels of throughput. Furthermore, the Java context of the application enables individual users to modify the program to incorporate institution-specific preferences such as universal cloning handles or specific links to institutional databases for sequence input. A complete description of the application parameters is available on the application Web site.

**Description of Algorithm**

Following the input of a flat text file containing sequences in FASTA format, the application segregates individual sequences and processes one sequence at a time. After processing for restriction enzyme constraints, the application selects candidate primer sequences (from the 5′-end), starting with the default sequence length of 16 unless the user specifies an alternate length. Candidate primers are screened according to user input and program criteria with successful candidates inserted into a temporary acceptable primer list. This process continues until the length of the primer reaches the default length of 30, unless the user specifies an alternate length. After compiling a list of forward candidate primers, the application reverse-complements the gene sequence and repeats the same process to generate a corresponding temporary list of acceptable reverse primers.

**Acceptable Primer Criteria Utilized by the Algorithm**

**Tm range test.** In the default high-throughput mode, the Tm of the primer must fall within the specified Tm range (optimal Tm ± Tm range) as defined by the user input values. If the calculated Tm for the sequence falls outside the specified Tm range, then the primer is discarded by the application. If the high-throughput approach is NOT chosen (choosing “yes” for the Optimize for Individual Primers feature), then the application does not apply a Tm range criteria for primer selection.

**Consecutive G/C test.** After the Tm range test, the application checks the number of consecutive terminal G/Cs for the primers that passed the Tm range test. The algorithm initially imposes a constraint that the last two nucleotides must be a G or C. If the minimum number of consecutive G/Cs is two or more, then it is considered valid. Sequences that do not meet the two consecutive G/C test must meet a secondary criterion that the last eight nucleotides must contain at least three G/C bases with the last nucleotide being either a G or C. When the “Add extra G/Cs” feature is selected, the application adds as many G/Cs as possible at the 3′-ends of the primers while continuing to meet

---

**Figure 1. Screenshot of the Express Primer Tool user interface.**
the rest of the criteria. The algorithm details for these criteria are available on the Express Primer Tool Web site.

Selecting the Best Forward and Reverse Primers

Upon completion of the acceptable primer tests, the application evaluates the forward and reverse candidate primers lists. If one or more of the lists is empty, then the application generates output indicating the sequence “failed” to produce any primers. When both lists contain one or more candidate primers, the primer pairs are further examined to identify the pair with the smallest Tm difference while checking to see that the Tm difference stays within the specified limits. In case of a numerical tie, the application chooses the pair whose average Tm is closer to the optimal Tm. Equivalent pairs at this level are selected on the basis of the shortest combined length. If none of the primer pairs have a Tm difference within the specified maximum, then the application generates output indicating the sequence failed to produce any primers. Input gene sequences that do not meet the above selection criteria (i.e., genes that are designated as failed) are written to a flat file on the server for additional runs with a new set of parameters. The application tracks sequences that fail to meet the specified criteria and do not generate valid primer sets. Every failure is recorded, and its corresponding gene ID and reasons for failure are displayed on the results page (Figure 2).

Application of the Express Primer Tool to Bacterial Sequences

The Express Primer Tool was used to design primers for amplification from various genome sources including a pool of 192 protein targets from *Haemophilus influenzae* and more than 500 targets from *Bacillus subtilis*. One group of primers (148) was designed using the Optimize for Individual Primers option. A second group of primers was designed using the default setting to optimize the primers for plate-based methods described for the

---

**Figure 2.** Express Primer Tool output. (A) Screenshot of application output illustrating sequence information for the successful generation of primer pairs. (B) Illustration of application output for sequences that fail to generate primer pairs with the user-specified criteria.
Table 1. Application of Express Primer Tool

<table>
<thead>
<tr>
<th>Organism</th>
<th>No. of Targets</th>
<th>Application</th>
<th>Amplification Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixed bacterial</td>
<td>148</td>
<td>Laboratory scale</td>
<td>93%</td>
</tr>
<tr>
<td>H. influenzae</td>
<td>192</td>
<td>High throughput</td>
<td>72%</td>
</tr>
<tr>
<td>B. subtilis</td>
<td>528</td>
<td>High throughput</td>
<td>88%</td>
</tr>
</tbody>
</table>

aDetailed information regarding the targets used in this study is available on the MCGS Web site (www.mcgs.anl.gov). H influenzae target identification numbers, APC345–368 and APC393–560; membrane protein targets, APC791–844; and B. subtilis target identification numbers, APC1001–1870.

generation of expression clones (5,11). An amplification efficiency of 93% (138 genes successfully amplified from a target pool of 148) was observed for the targets designed for individual amplification experiments (Table 1). These amplification fragments were generated using standard methods. Amplification efficiencies of 72%–88% were observed for the plate-based studies (Table 1). The size of the high-throughput target pools was sufficient to fill approximately eight 96-well plates. For this group, amplification was validated by a combination of gel electrophoresis and a PicoGreen® fluorescent assay (Molecular Probes, Eugene, OR, USA). The amplification of the H. influenzae targets was validated by gel analysis of the entire plate, followed by the PicoGreen assay to assess the concentration (data not shown). A portion (16 targets/plate) of the B. subtilis targets was screened by gel analysis (data available on application Web site) in addition to the standard PicoGreen assay. We observed good agreement between the gel electrophoresis and PicoGreen analysis methods.

The generation of an expression clone is often a prerequisite for the initiation of structural and functional studies. The Express Primer Tool is a Web-based application that is user friendly and provides an efficient resource for the generation of successful primer pairs for the DNA amplification of coding region sequences. The ability to design primers specifically for the generation of expression clones is not a component of common primer design applications found on the Web. Although adaptable for laboratory-scale studies, the application is most suited for a systems biology approach that necessitates the design of amplification primers for genome-scale experiments. The utility of the Express Primer Tool is facilitated by the rapid progress in the development of plate-based methods for the generation of expression clones (5). We are currently examining methods to integrate additional tools, such as a protein domain identification application, to further enhance the utility of the Express Primer Tool for eukaryotic proteins. The development of these tools will enhance protein expression capabilities and enable optimal utilization of the abundance of genomic sequence data. It is the integration of these tools and methods that will provide the opportunity to apply a systems biology approach for the understanding of protein structure and function and accelerate our understanding of protein structure and function.

ACKNOWLEDGMENTS

The authors acknowledge Diane Makowski, Fred Stevens, and Brian M. Dennis for their comments, support, and critical review of the manuscript. This work was supported by the U.S. Department of Energy, Office of Health and Environmental Research, under contract no. W-31-109-Eng-38 and by the National Institutes of Health grant no. GM62414-01. The views and opinions of document authors expressed herein do not necessarily state or reflect those of the U.S. Government or any agency thereof, Argonne National Laboratory, or The University of Chicago.

REFERENCES


Received 15 May 2002; accepted 10 July 2002.

Address correspondence to:
Dr. F.R. Collart
Biosciences Division
Building 202
Argonne National Laboratory
Argonne, IL 60439-4833, USA
e-mail: fcollart@anl.gov