8

Genome-Scale Probe and Primer Design with PRIMEGENS

Gyan Prakash Srivastava and Dong Xu

Summary

This chapter introduces the software package PRIMEGENS for designing gene-specific probes and associated PCR primers on a large scale. Such design is especially useful for constructing cDNA or oligo microarray to minimize cross-hybridization. PRIMEGENS can also be used for designing primers to amplify a segment of a unique target gene using reverse-transcriptase (RT)-PCR. The input to PRIMEGENS is a set of sequences, whose primers need to be designed, and a sequence pool containing all the genes in a genome. It provides options to choose various parameters. PRIMEGENS uses a systematic algorithm for designing gene-specific probes and its primer pair. For a given sequence, PRIMEGENS first searches for the longest gene-specific fragment and then designs best PCR product for this fragment. The 2.0 version of PRIMEGENS provides a graphical user interface (GUI) with additional features. The software is freely available for any users and can be downloaded from http://digbio.missouri.edu/primegens/.

Key Words: PCR primer design; cross-hybridization; cDNA microarray; oligo arrays; qRT-PCR; sequence alignment; dynamic programming.

1. Introduction

Various genome-scale sequencing project have generated vast amounts of sequence data. High-throughput data analysis and its study are one of the primary focuses for molecular biologist. Microarray is one of the most common tools for studying gene expressions on a large scale ***(1,2)***. In cDNA microarray, typically each spot on the array contains sequence segment of a specific gene, which is amplified by PCR. The segment is expected to be gene specific to

From: *Methods in Molecular Biology, vol. 402: PCR Primer Design*
Edited by: A. Yuryev © Humana Press, Totowa, NJ

avoid cross-hybridization among genes sharing significant sequence identity. In another case, researchers may simply want to amplify gene-specific segments for a selected group of genes using reverse-transcriptase (RT)-PCR. In both cases, the problem can be formulated to choose a gene-specific segment for a gene in a genome and then design PCR primers according to some specifications. Such an objective is often achieved manually, e.g., using Primer3 ***(3)*** for primer design for a given sequence. Primer3 designs many possible primer pairs for a given sequence, but it does not guarantee their uniqueness in the whole genome. Therefore, a user has to manually run BLAST ***(4)*** for each PCR product against the genome to search to avoid cross-hybridization. Such manual approach cannot be applied to a large scale. PRIMEGENS ***(5)*** does not only fulfill this task but also automate the primer generation on the large scale. Furthermore, PRIMEGENS has a rigorous formulation, which has a much better chance to find gene-specific segment than a manual process.

Figure 1 gives some general idea about PRIMEGENS. The essence of PRIMEGENS is based on searching the sequence-specific fragment for any particular sequence. PRIMEGENS implements this task by finding the fragment of a given DNA sequence, which does not have high-sequence similarity with any other sequence in the given sequence pool (whole genome in general). If the given sequence is unique, then the whole sequence is considered as the

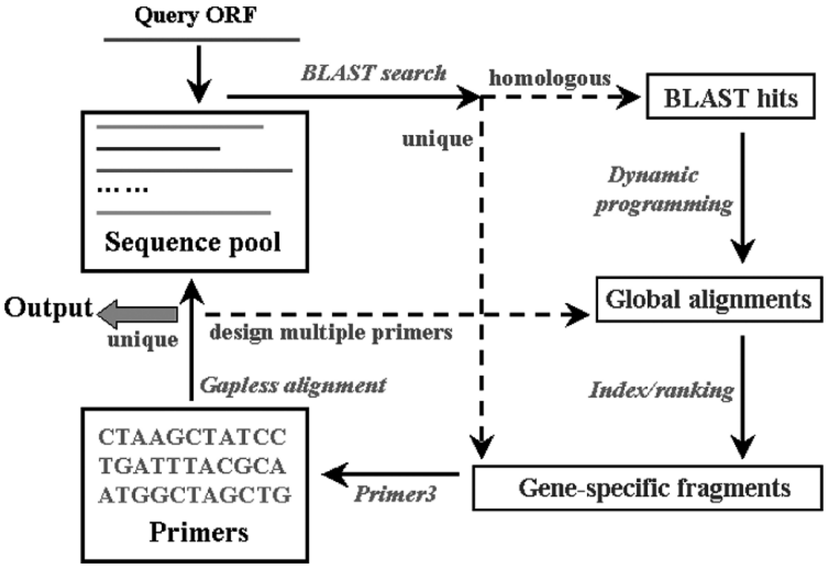


Fig. 1. Basic PRIMEGENS model.

sequence-specific fragment. Otherwise, PRIMEGENS searches for the unique fragment based on the BLAST result for the query sequence. The optimal global alignment between the query sequence and each of its significant BLAST hits is performed ***(6)***. Based on the alignment, PRIMEGENS searches for the longest unique segment for the query sequence. Finally, it designs primers on the selected gene-specific fragment using Primer3. **Figure 2** describes the detailed algorithm of the PRIMEGENS implementation.

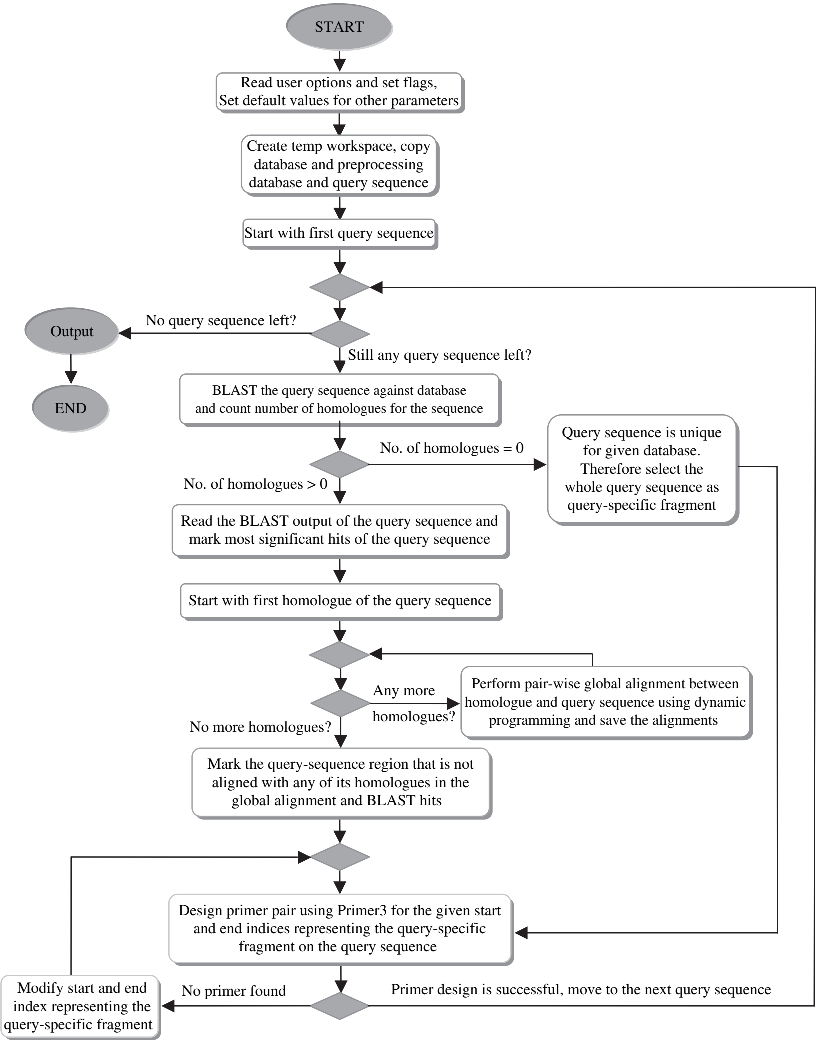

START
Read user options and set flags,
Set default values for other parameters
Create temp workspace, copy database and preprocessing database and query sequence
Start with first query sequence
No query sequence left?
Output
END
Still any query sequence left?
BLAST the query sequence against database and count number of homologues for the sequence
No. of homologues = 0
Query sequence is unique for given database. Therefore select the whole query sequence as query-specific fragment
No. of homologues > 0
Read the BLAST output of the query sequence and mark most significant hits of the query sequence
Start with first homologue of the query sequence
Any more homologues?
Perform pair-wise global alignment between homologue and query sequence using dynamic programming and save the alignments
No more homologues?
Mark the query-sequence region that is not aligned with any of its homologues in the global alignment and BLAST hits
Design primer pair using Primer3 for the given start and end indices representing the query-specific fragment on the query sequence
No primer found
Primer design is successful, move to the next query sequence
Modify start and end index representing the query-specific fragment

Fig. 2. Flowchart for PRIMEGENS implementation.

We recently developed PRIMEGENS version 2.0. In this new version, we improved the main algorithm and added a number of new features. In particular, we developed a Java-based graphical user interface (GUI), which can be used under both Windows and Linux platforms.

2. Description of PRIMEGENS

This section will explain what a user needs to specify for running PRIMEGENS. **Subheading 2.1** describes compositions of PRIMEGENS. **Subheading 2.2** shows various types of inputs. **Subheading 2.3** covers different types of execution features supported by PRIMEGENS. **Subheading 2.4** explains about the *append file* ***(3)***, which is the input for primer design that controls the specifications of primers. **Subheading 2.5** describes the format of primer design results. More detailed description about PRIMEGENS is provided in the software documentation, which comes along with the PRIMEGENS package.

2.1. Composition of the Software Package

PRIMEGENS 2.0 is available in the form of compressed format as *PRIMEGENSv2.zip* for Windows and *PRIMEGENSv2.tgz* for Linux. These packages are freely available for any users and can be downloaded from http://digbio.missouri.edu/primegens/. For installation, user should specify the location of PRIMEGENS folder by setting the environment variable *PRIMEGENS_PATH* with the absolute path of the software location. More detailed description about software installation can be found in *README.txt*.

PRIMEGENS 2.0 (PRIMEGENS as the main folder) consists of following major directories and files:

1. bin/: console application executables
2. blast/: BLAST executables
3. doc/: documentation
4. include/: supporting resources
5. output/: output results
6. primer3/: primer3 executables
7. primerdesign/: graphical interface files
8. test/: testing resources
9. README.txt: instruction manual
10. primerdesign.jar: main Java executable for graphical interface

2.2. Inputs to PRIMEGENS

PRIMEGENS supports various input features according to the user requirements. This section describes each type of input.

2.2.1. Sequence Pool

To start primer design, a user needs to create a database file consisting of all the sequences in the FASTA format. The content format of the database file should look like as shown in **Fig. 3**.

2.2.2. Sequence of Interest

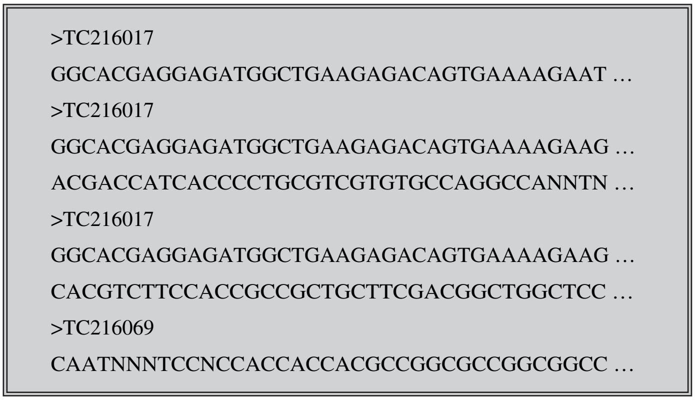
By default, PRIMEGENS searches for the unique sequence-specific fragment and primer pair relative to all the sequences present in database file. Alternatively, if the user is interested in a set of those sequences, a list of these sequences should be provided in to a separate file. This subset file can be either in the FASTA format or a list in which each line gives the name of the gene.

2.2.3. Saving Result Files

PRIMEGENS generates various types of results in different files. This information may be useful subsequently; therefore, a user can specify any location on the local computer to save all the result files.

2.3. Execution Features (Command-Line Options)

Once the input files are selected, PRIMEGENS provides various execution features. Following are some of the useful options provided by PRIMEGENS.



```
>TC216017
GGCACGAGGAGATGGCTGAAGAGACAGTGAAAAGAAT …
>TC216017
GGCACGAGGAGATGGCTGAAGAGACAGTGAAAAGAAG …
ACGACCATCACCCCTGCGTCGTGTGCCAGGCCANNTN …
>TC216017
GGCACGAGGAGATGGCTGAAGAGACAGTGAAAAGAAG …
CACGTCTTCCACCGCCGCTGCTTCGACGGCTGGCTCC …
>TC216069
CAATNNNTCCNCCACCACCACGCCGGCGCCGGCGGCC …
```

Fig. 3. Input database format.

2.3.1. Keeping All the BLAST Results (from GUI)

This option can be controlled by the user to keep all the BLAST results (which otherwise will be deleted by default) for each query sequence against the database. The BLAST result for each query sequence against the database is stored in the TMP directory in the main PRIMEGENS workspace. The file nomenclature is organized in such a way that query sequence name appears as follows *<Query-name>_<PID>_primegens_log*, where "PID" is the process ID of this computational job. The user can open these files in any text editor like Microsoft Word Pad.

2.3.2. User-Defined Expectation Value for BLAST

This optional feature allows user to set any user-defined expectation value ***(4)*** for running the BLAST program. In case the expectation value is not specified by the user, PRIMEGENS sets its default value to *1e-5*.

2.3.3. Using Subset File

One can choose a subset file for designing primers, where the entries in the subset files are selected from the search database file. In this case, a user needs to specify if the file is in the FASTA format or in the non-FASTA format using the command-line execution as follows:

```
$> primegens.exe –lf <fasta-subset-file-name> <database-file-name>
```

or

```
$> primegens.exe –l <non-fasta-subset-file-name> <database-file-name>
```

2.3.4. User-Defined PCR Product Size

A user can define its own PCR product-size range for the primer pair design. These values can be specified in the form of maximum and minimum value for the product size using the −*fz* flag on the command line. For example, the command line defining the product size range of 80–120 is as follows:

```
$> primegens.exe –fz 80 120 –lf <subset-file-name> <database-file-name>
```

In case a user wants to specify product size as a function of sequence length, the command to specify product size should be

```
$> primegens.exe –f <fraction> -lf <subset-file-name> <database-file-name>
```

Here, *fraction* is a value between 0 and 1, which represents the ratio between the minimum length of the fragment and the whole sequence length. By default, PRIMEGENS assumes the maximum product size as the sequence length itself.

2.3.5. Uniqueness of Primers

Primer pair designed for any gene-specific segment still has non-zero probability to amplify a different gene, as the short primer pair itself may not be gene specific. In particular, if the first *K* bases in the left primer and the last *K* bases of the right primer match exactly in some region of another gene in the whole sequence pool, then the designed primer pair may amplify both genes. PRIMEGENS allows user to set a parameter *K*, whose default value is 10. PRIMEGENS uses this option to make sure that primer pair is unique for the first *K* bases (left primer) and the last *K* bases (right primer) for associated query sequence. A user can choose the option like the following command

```
$> primegens.exe –pterm K –lf <subset-file-name> >database-file-name>
```

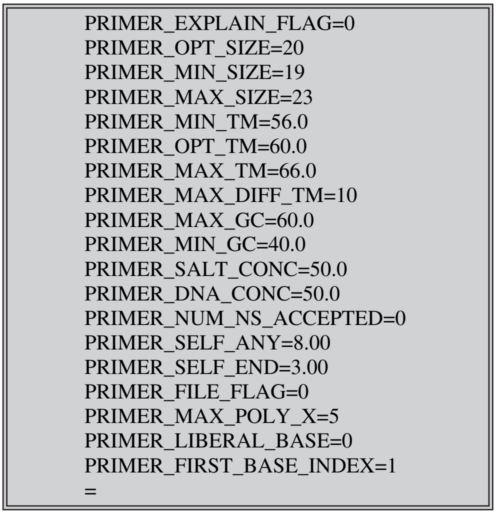
2.4. Primer3 Parameter Input (Append File)

All the optional parameters for Primer3 are stored in a file with a reserved name called *append.txt* ***(3)***. If a user wants to specify his or her own parameter values, he or she needs to modify this file. A sample format for *append.txt* is shown in **Fig. 4**. The location of this file should be the same as the location of the input database file and the optional subset file (for the console version rather than the GUI version). If PRIMEGENS cannot find this file, it will use the default values provided in the “include” directory of the software package.

2.5. Output of PRIMEGENS

PRIMEGENS supports permanent storage of primer design results. To generate organized results, PRIMEGENS creates various files and directories. Here is a brief description of all types of generated files and directories. For generality, it is assumed that a user has selected both *Database.txt* as the database file and *subset.txt* as the subset file for PRIMEGENS. If *subset.txt* is not specified, it will select all the entries from *Database.txt* without sequences (only sequence IDs).

- Database_nohit.txt: This file contains those query sequences from *subset.txt* that are unique in *Database.txt* (no hit is found in BLAST). The file is in the FASTA format with header line containing the sequence name and length.
- Database_nohit.txt_back: This is the backup file of the *Database_nohit.txt* result for the previous run.



```
PRIMER_EXPLAIN_FLAG=0
PRIMER_OPT_SIZE=20
PRIMER_MIN_SIZE=19
PRIMER_MAX_SIZE=23
PRIMER_MIN_TM=56.0
PRIMER_OPT_TM=60.0
PRIMER_MAX_TM=66.0
PRIMER_MAX_DIFF_TM=10
PRIMER_MAX_GC=60.0
PRIMER_MIN_GC=40.0
PRIMER_SALT_CONC=50.0
PRIMER_DNA_CONC=50.0
PRIMER_NUM_NS_ACCEPTED=0
PRIMER_SELF_ANY=8.00
PRIMER_SELF_END=3.00
PRIMER_FILE_FLAG=0
PRIMER_MAX_POLY_X=5
PRIMER_LIBERAL_BASE=0
PRIMER_FIRST_BASE_INDEX=1
=
```

Fig. 4. Format of *append file* to PRIMEGENS.

- Database_nohit_primer.txt: This file contains all the primers successfully designed by the PRIMEGENS for sequences in *Database_nohit.txt*. The file includes the left and right primer sequences along with their locations on the query sequence.
- Database_nohit_primer.txt_back: This backup file contains all the record of *Database_nohit_primer.txt* for the previous primer design results created by PRIMEGENS.
- Database_primer.xls: This plain text file can be opened in an excel spreadsheet, and it contains all the designed primers along with various other information about the primer pairs.
- Database_primer.xls_back: This is the backup file of *Database_primer.xls* for the previous primer design.
- Database_primer_undo.xls: This plain text file can be opened in an excel spreadsheet, and it contains all the query sequences whose primer pairs could not be designed by PRIMEGENS based on the specified parameters.
- Database_primer_undo.xls_back: This is the backup file of *Database_primer_undo.txt* for the previous primer design.
- Database_seg.txt: This file contains those query sequences that are not completely unique in the database but have some sequence-specific fragments that are unique in whole database. It includes the name of the query sequence along with the longest sequence-specific fragments and its location on the query sequence.
- Database_seg.txt_back: This is the backup file of *Database_seg.txt* from the previous execution of PRIMEGENS.

- Database_seg_primer.txt: This file contains all the successfully designed primers for the sequences in *Database_seg.txt*. The file includes the query-specific fragments along with the designed primer pairs and their locations on the original sequences.
- Database_seg_primer.txt_back: This is the backup file of *Database_seg_primer.txt* for the previous run.
- Database_sim.txt: This file contains information about the query sequences that are not unique in database. This file also shows which sequence in the database is the closest to the query sequence along with the sequence identity.
- Database_sim.txt_back: This is the backup file of *Database_sim.txt* for the previous run.
- Primer_plate_left [index]: This file contains query sequence name and its left primer, which should be kept in a 96-well primer plate. The plate number is represented by the index digit.
- Primer_plate_right [index]: This file contains query sequence name and its right primer, which should be kept in a 96-well primer plate. The plate number is represented by the index digit.

Besides the various files, there are some temporary directories that are created by PRIMEGENS. Unlike the above files, these directories are cleaned before and after the software execution. A user may save them in different folders to keep them from deleting. These directories are described below.

- TMP/directory: This directory contains the BLAST output for each query sequence in *subset.txt* against *Database.txt*. The file name is selected on the basis of query sequence name. This directory can also be saved through an option from GUI (*see* **Subheading 2.3.1**.).
- LOGS/directory: This directory contains various logs, which are created during the primer design process when debugging mode is used for execution. The purpose of this option is to provide run-time information of the whole process from the user perspective.
- SEQ/directory: This directory contains individual sequence file for each query sequence in the FASTA format. The purpose of this directory is to retrieve sequence of interest without searching into the database.
- ALIGNMENT/directory: This directory contains global alignments performed for those sequences, which are not unique in the database. The alignment is performed to find the longest sequence-specific fragment for the query sequence.

3. PRIMEGENS Execution

This section will explain how to run PRIMEGENS. We will describe it for both command line and GUI version.

3.1. Using PRIMEGENS as Console Application

To use the software for designing primers, the following basic procedure should be followed.

1. Create a new directory to save all the input and output of the software.
2. Prepare a database file (i.e., *Database.txt; see* **Subheading 2.2.1**.) consisting of all the potential sequences with cross-hybridization in the new directory.
3. If necessary, select a subset of sequences (i.e., *subset.txt; see* **Subheading 2.2.2**.) from above database, whose primers should be designed, and keep it in the new directory.
4. Copy *append.txt* file from include/of the PRIMEGENS folder in to the new location. Modify the primer parameters according to the experiment and save it. If not specified by the user, PRIMEGENS will use the default parameters.
5. Change the directory to the newly created directory and give following command,

```
<PRIMEGENS_PATH>/bin/primegens.exe –option subset.txt database.txt
```

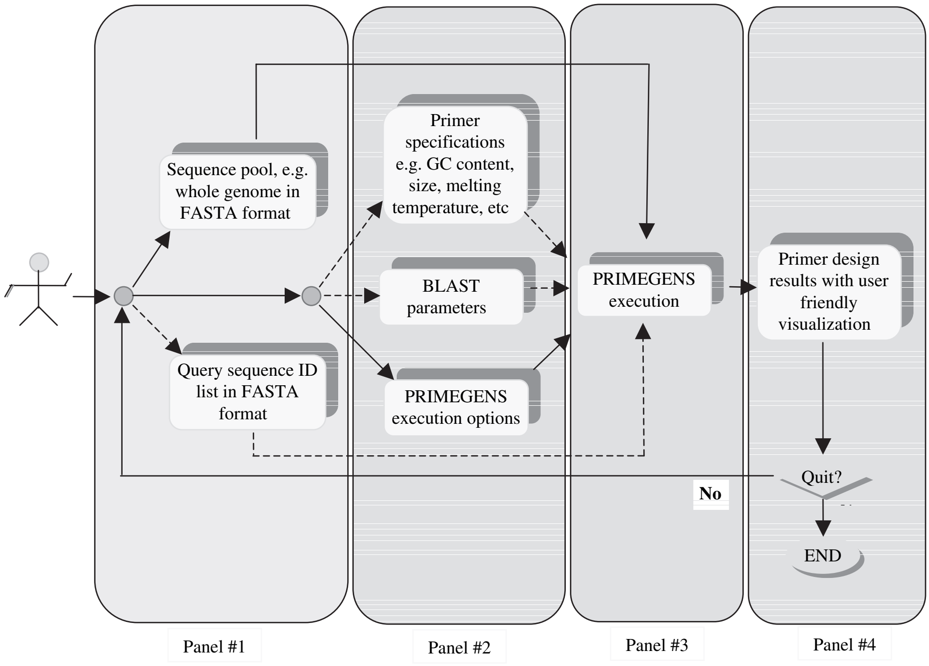

Sequence pool, e.g. whole genome in FASTA format
Query sequence ID list in FASTA format
Primer specifications e.g. GC content, size, melting temperature, etc
BLAST parameters
PRIMEGENS execution options
PRIMEGENS execution
Primer design results with user friendly visualization
Quit?
No
END
Panel #1
Panel #2
Panel #3
Panel #4

Fig. 5. Workflow for primer design. Database input (panel #1), design specification (panel #2), primer design process (panel #3), and results view (panel #4). Solid arrows represent required inputs, whereas dashed arrows show optional inputs.

More information about various options and command format is described in **Subheading 2.3**.

3.2. Running PRIMEGENS from GUI

To use the GUI version of PRIMEGENS, the database file also has to be prepared first. The user can run the software by double clicking on the executable. **Figure 5** provides a systematic workflow from the user perspective for primer design. The details for using the GUI version are explained in the following subheadings.

3.2.1. Database Input

Figure 6 shows the first panel, which will be visible when a user clicks to run PRIMEGENS. User should input the database, subset file (optional), and result storage location (optional). Once completed with input, the "next" button should be pressed for execution options.

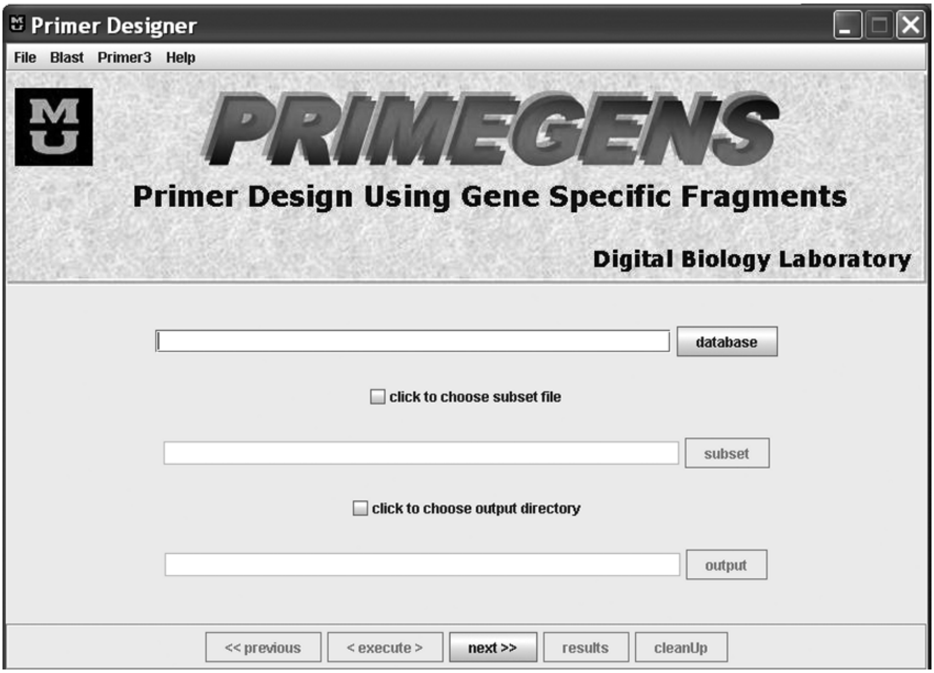

Primer Designer
File Blast Primer3 Help
MU
PRIMEGENS
Primer Design Using Gene Specific Fragments
Digital Biology Laboratory
database
☐ click to choose subset file
subset
☐ click to choose output directory
output
<< previous
< execute >
next >>
results
cleanUp

Fig. 6. PRIMEGENS graphical user interface (GUI).

3.2.2. Execution Option Window

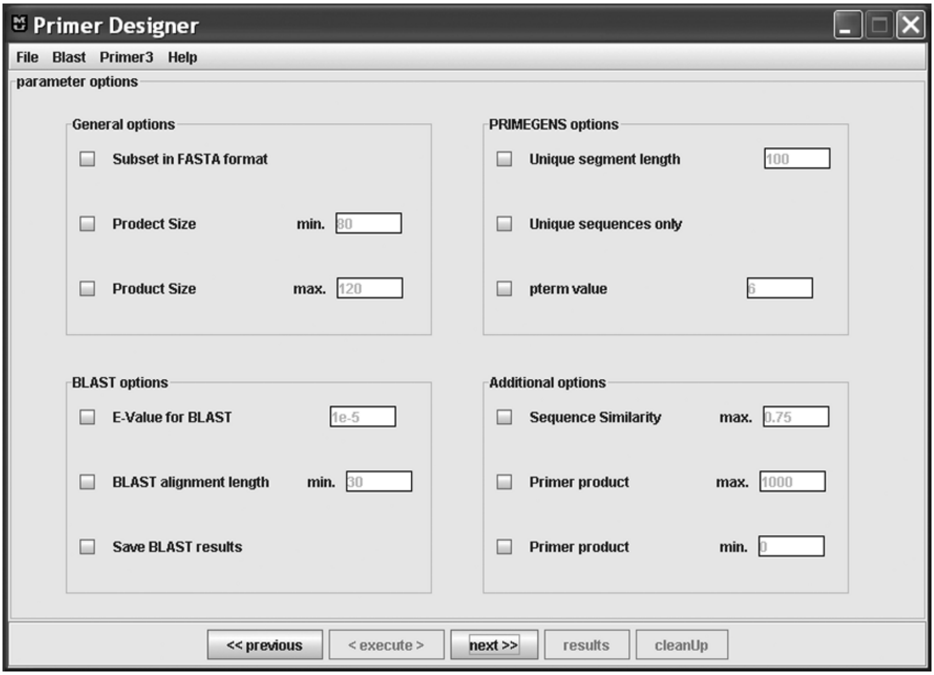
This panel contains various execution options as explained in **Subheading 2.3**. The execution features supported by the GUI map to all command-line features correspondingly. **Figure 7** shows the option panel. Once user selects any attribute, the optional attribute value field shows the default attribute value, which can be modified then.

3.2.3. Primer Pair Specification

In case a user has specific parameter requirements for primer pairs, he or she can specify those values in the primer-specification window. The user can click on the *Primer3* menu to modify primer specifications. These modifications correspond to changes in the *append.txt* file in the command-line version.

3.2.4. Execution-Display Window

After specifying inputs and options, the software allows a user to open the execution window. Once the primer design is completed, the *result* and the



parameter options

General options

□ Subset in FASTA format

□ Prodect Size min. 80

□ Product Size max. 120

PRIMEGENS options

□ Unique segment length 100

□ Unique sequences only

□ pterm value 6

BLAST options

□ E-Value for BLAST 1e-5

□ BLAST alignment length min. 30

□ Save BLAST results

Additional options

□ Sequence Similarity max. 0.75

□ Primer product max. 1000

□ Primer product min. 0

Fig. 7. Execution option window.

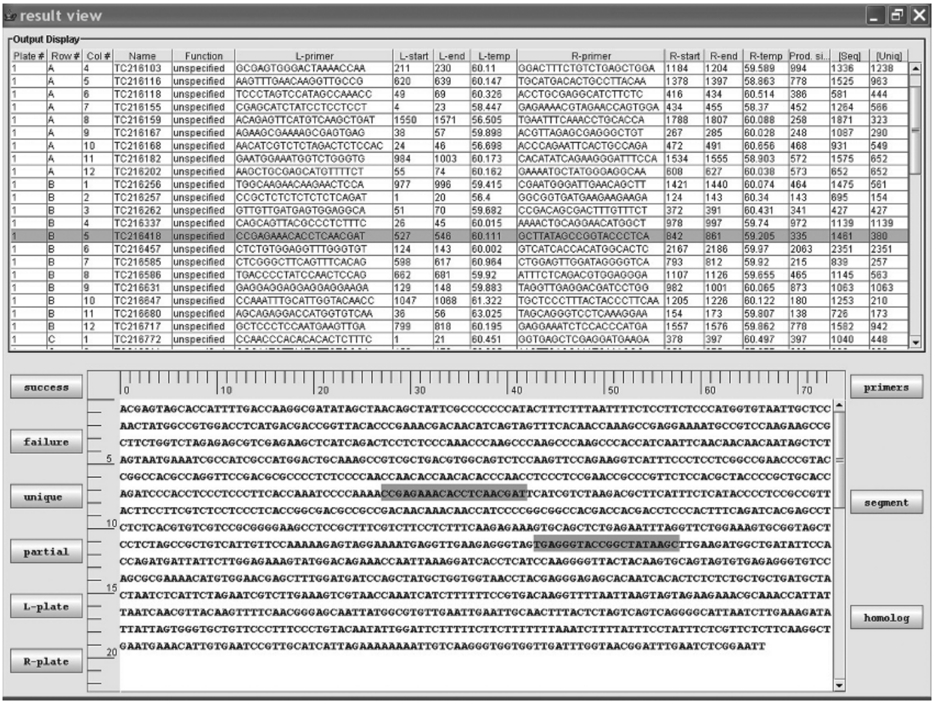


Plate #	Row #	Col #	Name	Function	L-primer	L-start	L-end	L-temp	R-primer	R-start	R-end	R-temp	Prod. si...	[Seq]	[Uniq]
1	A	4	TC216103	unspecified	GCGAGTGGGACTAAAACCAA	211	230	60.11	GGACTTTCTGTCTGAGCTGGA	1184	1204	59.589	994	1336	1238
1	A	5	TC216116	unspecified	AAGTTTGAACAAGGTTGCCG	620	639	60.147	TGCATGACACTGCCTTACAA	1378	1397	58.863	778	1525	963
1	A	6	TC216118	unspecified	TCCCTAGTCCATAGCCAAACC	49	69	60.326	ACCTGCGAGGCATCTTCTC	416	434	60.514	386	581	444
1	A	7	TC216155	unspecified	CGAGCATCTATCCTCCTCCT	4	23	58.447	GAGAAAACGTAGAACCAGTGGA	434	455	58.37	452	1264	566
1	A	8	TC216159	unspecified	ACAGAGTTCATGTCAAGCTGAT	1550	1571	56.505	TGAATTTCAAACCTGCACCA	1788	1807	60.088	258	1871	323
1	A	9	TC216167	unspecified	AGAAGCGAAAAGCGAGTGAG	38	57	59.898	ACGTTAGAGCGAGGGCTGT	267	285	60.028	248	1087	290
1	A	10	TC216168	unspecified	AACATCGTCTCTAGACTCTCCAC	24	46	56.698	ACCCAGAATTCACTGCCAGA	472	491	60.656	468	931	549
1	A	11	TC216182	unspecified	GAATGGAAATGGTCTGGGTG	984	1003	60.173	CACATATCAGAAGGGATTTCCA	1534	1555	58.903	572	1575	652
1	A	12	TC216202	unspecified	AAGCTGCGAGCATGTTTTCT	55	74	60.162	GAAAATGCTATGGGAGGCAA	608	627	60.038	573	652	652
1	B	1	TC216256	unspecified	TGGCAAGAACAAGAACTCCA	977	996	59.415	CGAATGGGATTGAACAGCTT	1421	1440	60.074	464	1475	561
1	B	2	TC216257	unspecified	CCGCTCTCTCTCTCTCAGAT	1	20	56.4	GGCGGTGATGAAGAAGAAGA	124	143	60.34	143	695	154
1	B	3	TC216262	unspecified	GTTGTTGATGAGTGGAGGCA	51	70	59.682	CCGACAGCGACTTTGTTTCT	372	391	60.431	341	427	427
1	B	4	TC216337	unspecified	CAGCAGTTACGCCCTCTTTC	26	45	60.015	AAAACTGCAGGAACATGGCT	978	997	59.74	972	1139	1139
1	B	5	TC216418	unspecified	CCGAGAAACACCTCAACGAT	527	546	60.111	GCTTATAGCCGGTACCCTCA	842	861	59.205	335	1461	380
1	B	6	TC216457	unspecified	CTCTGTGGAGGTTTGGGTGT	124	143	60.002	GTCATCACCACATGGCACTC	2167	2186	59.97	2063	2351	2351
1	B	7	TC216585	unspecified	CTCGGGCTTCAGTTTCACAG	598	617	60.964	CTGGAGTTGGATAGGGGTCA	793	812	59.92	215	839	257
1	B	8	TC216586	unspecified	TGACCCCTATCCAACTCCAG	662	681	59.92	ATTTCTCAGACGTGGAGGGA	1107	1126	59.655	465	1145	563
1	B	9	TC216631	unspecified	GAGGAGGAGGAGGAGGAAGA	129	148	59.883	TAGGTTGAGGACGATCCTGG	982	1001	60.065	873	1063	1063
1	B	10	TC216647	unspecified	CCAAATTTGCATTGGTACAACC	1047	1068	61.322	TGCTCCCTTTACTACCCTTCAA	1205	1226	60.122	180	1253	210
1	B	11	TC216680	unspecified	AGCAGAGGACCATGGTGTCAA	36	56	63.025	TAGCAGGGTCCTCAAAGGAA	154	173	59.807	138	726	173
1	B	12	TC216717	unspecified	GCTCCCTCCAATGAAGTTGA	799	818	60.195	GAGGAAATCTCCACCCATGA	1557	1576	59.862	778	1582	942
1	C	1	TC216772	unspecified	CCAACCCACACACACTCTTTC	1	21	60.451	GGTGAGCTCGAGGATGAAGA	378	397	60.497	397	1040	448

```
ACGAGTAGCACCATTTTGACCAAGGCGATATAGCTAACAGCTATTCGCCCCCCCATACTTTCTTTAATTTTCTCCTTCTCCCATGGTGTAATTGCTCC
AACTATGGCCGTGGACCTCATGACGACCGGTTACACCCGAAACGACAACATCAGTAGTTTCACAACCAAAGCCGAGGAAAATGCCGTCCAAGAAGCCG
CTTCTGGTCTAGAGAGCGTCGAGAAGCTCATCAGACTCCTCTCCCAAACCCAAGCCCAAGCCCAAGCCCACCATCAATTCAACAACAACAATAGCTCT
AGTAATGAAATCGCCATCGCCATGGACTGCAAAGCCGTCGCTGACGTGGCAGTCTCCAAGTTCCAGAAGGTCATTTCCCTCCTCGGCCGAACCCGTAC
CGGCCACGCCAGGTTCCGACGCGCCCCTCTCCCCAACCAACACCAACACACCCAACCTCCCTCCGAACCGCCCGTTCTCCACGCTACCCCGCTGCACC
AGATCCCACCTCCCTCCCTTCACCAAATCCCCAAAACCGAGAAACACCTCAACGATTCATCGTCTAAGACGCTTCATTTCTCATACCCCTCCGCCGTT
ACTTCCTTCGTCTCCTCCCTCACCGGCGACGCCGCCGACAACAAACAACCATCCCCGGCGGCCACGACCACGACCTCCCACTTTCAGATCACGAGCCT
CTCTCACGTGTCGTCCGCGGGGAAGCCTCCGCTTTCGTCTTCCTCTTTCAAGAGAAAGTGCAGCTCTGAGAATTTAGGTTCTGGAAAGTGCGGTAGCT
CCTCTAGCCGCTGTCATTGTTCCAAAAAGAGTAGGAAAATGAGGTTGAAGAGGGTAGTGAGGGTACCGGCTATAAGCTTGAAGATGGCTGATATTCCA
CCAGATGATTATTCTTGGAGAAAGTATGGACAGAAACCAATTAAAGGATCACCTCATCCAAGGGGTTACTACAAGTGCAGTAGTGTGAGAGGGTGTCC
AGCGCGAAAACATGTGGAACGAGCTTTGGATGATCCAGCTATGCTGGTGGTAACCTACGAGGGAGAGCACAATCACACTCTCTCTGCTGCTGATGCTA
CTAATCTCATTCTAGAATCGTCTTGAAAGTCGTAACCAAATCATCTTTTTTCCGTGACAAGGTTTTAATTAAGTAGTAGAAGAAACGCAAACCATTAT
TAATCAACGTTACAAGTTTTCAACGGGAGCAATTATGGCGTGTTGAATTGAATTGCAACTTTACTCTAGTCAGTCAGGGGCATTAATCTTGAAAGATA
TTATTAGTGGGTGCTGTTCCCTTTCCCTGTACAATATTGGATTCTTTTTCTTCTTTTTTTAAATCTTTTATTTCCTATTTCTCGTTCTCTTCAAGGCT
GAATGAAACATTGTGAATCCGTTGCATCATTAGAAAAAAAATTGTCAAGGGTGGTGGTTGATTTGGTAACGGATTTGAATCTCGGAATT
```

Fig. 8. Primer design result display window.

clean buttons are activated. User can click on the *result* button to see the results and *clean* to remove all the temporary results from buffer and reset the software to the first window.

3.2.5. Result Analysis and Visualization Window

This window displays the primer design results generated by PRIMEGENS for all the sequences whose primers are found, including the gene-specific fragment and the global alignment between the sequence and its BLAST hits. The various buttons are self-explanatory. **Figure 8** shows a sample result visualization window.

4. Application of PRIMEGENS in the Quantitative RT-PCR Primer Design for the Transcription Factors in the Soybean Genome

As an example of successful PRIMEGENS application, the following shows some details for an actual research project. The aim of this project is to use gene-specific primers to isolate transcriptional factors (TFs) in the soybean *Glycine max* through quantitative (q)RT-PCR. We have used several approaches toward identification of putative TFs in soybean. Data were acquired from multiple resources (as below) and combined to generate the non-redundant final list of 734 putative TFs.

1. Soybean sequences homologous to the ~ 2300 *Arabidopsis thaliana* TFs were identified by searching the National Center for Biotechnology Information (NCBI) *Glycine max* Unigene database at ftp://ftp.ncbi.nih.gov/repository/UniGene/Glycine_max. To assure a high homology between *Arabidopsis* and *Glycine max* sequences, an E-value ≤ 1e-12 was considered as a cutoff. These analyses yielded 238 putative TFs in *Glycine max*, which fulfill these requirements.
2. We searched the listing of several TF domains and families already registered at the TIGR Gene Indices database (http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=soybean). So far, 1321 tentative consensus (TC) sequences are available on this Web site.
3. To expand the above list, the expressed-sequence tag (EST) soybean database (dbEST, http://www.ncbi.nlm.nih.gov/dbEST) was also screened for TF sequences. An initial search identified 1043 ESTs.
4. All putative TFs identified from TC sequences and ESTs (from **steps 2** and **3** above) were combined to form a data set, and were compared against the *Glycine max* Unigene data set using BLAST with an E-value cutoff of ≤ 1e-12. This yielded 496 putative TFs in *Glycine max*, out of which 449 match the putative TF TC sequences and 47 match the putative TF ESTs.
5. The final list (from **steps 1** and **4**) contains 734 putative TFs in *Glycine max*. PRIMEGENS was used to design primers for qRT-PCR experiment for validation of the above 734 TFs. All sequences were compared against the *Glycine max* Unigene database to ensure that the designed primers are unique to the sequence and will give specific PCR amplification.

4.1. Input File

The NCBI unique unigenes database for soybean (15,047 unigenes at the time of the design), acquired from ftp://ftp.ncbi.nih.gov/repository/UniGene/Glycine_max, was used as the database, and 734 TFs identified in soybean were used as the subset file for primer design.

4.2. Execution Option

Product size range = 80–150.

The rest of the parameters were all set to the default values.

4.3. Primer Parameters (Append File)

```
PRIMER_EXPLAIN_FLAG=0
PRIMER_OPT_SIZE=20
PRIMER_MIN_SIZE=17
PRIMER_MAX_SIZE=25
PRIMER_MIN_TM=50.0
PRIMER_OPT_TM=60.0
PRIMER_MAX_TM=70.0
PRIMER_MAX_DIFF_TM=10
PRIMER_MAX_GC=65.0
PRIMER_MIN_GC=35.0
PRIMER_SALT_CONC=50.0
PRIMER_DNA_CONC=50.0
PRIMER_NUM_NS_ACCEPTED=0
PRIMER_SELF_ANY=8.00
PRIMER_SELF_END=3.00
PRIMER_FILE_FLAG=0
PRIMER_MAX_POLY_X=5
PRIMER_LIBERAL_BASE=0
PRIMER_FIRST_BASE_INDEX=1
```

4.4. Machine Configuration

Hardware: X86 64-bit processor.
Memory: 8 GB random access memory (RAM).
Platform: Linux release 2.6.12.

4.5. PRIMEGENS Result Statistics

Primer pair design targets = 734 genes.

Possible gene-specific fragments found = 680 genes [(680/734) = 92.64%].

Successfully designed primer pairs = 670 genes [(670/734) = 91.28%].

Primer design failure = 10 genes [(10/734) = 1.36%].

Unique query genes for the whole sequence = 182 genes [(182/680) = 26.8%].

Genes containing gene-specific fragments = 498 genes [(498/680) = 73.24%].

Total execution time = 4 h 44 min 58 s.

Average time per gene (sequence) = $[(4 \times 3600) + (44 \times 60) + 58]/734 = 23.29$ s].

Our collaborator Gary Stacey's laboratory at the University of Missouri-Columbia performed qRT-PCR for the first 96 designed primers. Among them, 89 (92.7%) primer pairs yield single bands, which means that unique PCR products are amplified as expected (manuscript in preparation). Given that the soybean whole genome has not been sequenced, that is, the 15,047 unigenes that we used do not cover all the genes; such a successful rate is satisfactory.

5. Conclusion

PRIMEGENS provides an easy-to-use tool for biologist to select gene-specific fragment and to design PCR primers. A biologist with little computer skill can easily use the GUI version, which is available for both Windows and Linux platforms. The primer amplification results of TFs in the soybean genome, results from other projects ***(5,7–9)***, and feedback from users all indicate that the software works successfully and efficiently. We are continuing to refine the software efficiency and develop more features. The new versions of the software together with documentations will be released at the PRIMEGENS Web site (http://digbio.missouri.edu/primegens/).

Acknowledgments

The development of PRIMEGENS was sponsored by the Office of Biological and Environmental Research, U.S. Department of Energy, under Contract DE-AC05-00OR22725, managed by UT-Battelle, LLC, and a startup fund for Dong Xu at the University of Missouri-Columbia. We thank Ying Xu, Jizhong Zhou, Gary Li, Liyou Wu, Trupti Joshi, and Gary Stacey for collaboration and assistance.

References

1. DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278, 680–686.
2. Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M. (1999) Expression profiling using cDNA microarrays. *Nat. Genet.*, 21, 10–14.
3. Rozen, S. and Skaletsky, H. J. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, 132, 365–386.
4. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25, 3389–3402.
5. Xu, D., Li, G., Wu, L., Zhou, J., and Xu, Y. (2002) PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis. *Bioinformatics*, 11, 1432–1437.

6. Smith, T. F. and Waterman, M. S. (1981) Comparison of biosequences. *Adv. Appl. Math.*, 2, 482–489.
7. Liu, Y., Zhou, J., Omelchenko, M. V., Beliaev, A. S., Venkateswaran, A., Stair, J., Wu, L., Thompson, D. K., Xu, D., Rogozin, I. B., Gaidamakova, E. K., Zhai, M., Makarova, K. S., Koonin, E. V., and Daly, M. J. (2003) Transcriptome dynamics of *Deinococcus radiodurans* recovering from ionizing radiation. *Proc. Natl. Acad. Sci. U. S. A.*, 100, 4191–4196.
8. David, J.-P., Strode, C., Vontas, J., Nikou, D., Vaughan, A., Pignatelli, P. M., Louis, C., Hemingway, J., and Ranson, H. (2005) The *Anopheles gambiae* detoxification chip: a highly specific microarray to study metabolic-based insecticide resistance in malaria vectors. *Proc. Natl. Acad. Sci. U. S. A.*, 102, 4080–4084.
9. Wu, L. Y., Thompson, D. K., Li, G. S., Hurt, R. A., Tiedje, J. M., and Zhou, J. Z. (2001) Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl. Environ. Microbiol.*, 67, 5780–5790.