GEMI

A program for designing PCR primers

Version 1.3.1

User's Guide

Haitham Sobhy
URMITE, Aix-Marseille University,
Marseille, France

12.2013

Table of Content

Introduction

Gemi is an automated and user-friendly tool to design primers for Polymerase Chain Reaction (PCR) test. The PCR test has increasing importance in molecular biology field to test the presence of special gene or to study its expression assay. Since, it mimics the transcription process in the cell; it depends on DNA polymerase (*Taq*-polymerase or ***Taq***, for short) that uses set of oligo-nucleotides as **primers**, which are complementary to the target DNA sequences, act as starting point for nucleic acid extension forming a new strand leading to amplification. Therefore, the primarily and basic step in PCR test is to design these primers.

Each conventional PCR test requires **forward** primer (**fwd**) and **reverse** primer (**rev**). In quantitative PCR (qRT-PCR), **probe** is also required. These oligos together form a **PCR system**. Generally, it is ranging from 100 to 150 in case of qRT-PCR and ranges between 250 and 600 for Sanger sequencing tests.

The tool is originally developed to help the experimental biologists and other researchers, with no programming skills to design primers quickly, easily and precisely. Therefore, the tool uses simple and basic commands with easy user interface.

Availability

The Gemi package is available on https://sourceforge.net/projects/gemi/ and it consists of three programs. The main module (program) is used for designing primers, while the others two modules are supplementary to the main one. The user manual and source codes are also found on the previous URL.

Specifications

The program composed of three modules.

1. First is the main window to design primers.

2. Convert from aln, gde or phy file formats to fasta one. Since, Gemi accepts only upper-cased fasta files, this module also can be used to convert small letters (low-case) fasta to the capitalized form (upper-case).

3. Reverse and/or complement module can be used to reserve or constructing complementary sequence of one or more motifs.

Requirements and compatibility

The program is developed by C#.NET 2005. The software runs and tested on different operating system (Mac OS, Ubuntu and Windows XP and 7).

In future, under certain circumstance a command-line script might be developed for the tool.

For Windows users, the software requires The Microsoft™ .NET (Dot Net) Framework version 2.0 (x86 or x64) redistributable package, which is freely available from Microsoft website. Please, consult your computer administrator for details on x86 or x64 types.

http://www.microsoft.com/download/en/details.aspx?id=19 (x86)

http://www.microsoft.com/download/en/details.aspx?id=6523 (x64)

For Windows 7, the .Net framework 2.0 is usually integrated Windows 7. So, you may NOT need to install it.

For the Linux, Ubuntu and Mac OS X users, please download Mono tool to run the software.

http://www.mono-project.com/Main_Page

http://monodevelop.com/

License

Please, cite to the article:

"Sobhy, H, Colson, P; Gemi: PCR primers prediction from multiple alignments, Comparative and Functional Genomics 2012:783138; doi: 10.1155/2012/783138; PubMed ID: 23316117".

What is new in version 1.3.1?

1. In the first version, the input file should be in 'fasta' file extension. In the newer version, 'txt' extension has been added to ease of use for non-bioinformatician users.

2. Calculation of melting temperature 'Tm' based on salt adjustment equations. For the accurate estimation, the following equations are used:

   For nucleotide length less than 14:

   Td = 2 * (number of A's or T's in primer) + 4 * (number of C's or G's in primer)

   For nucleotide length equal or larger than 14:

   Td = 64.9°C + (41°C * (number of G's and C's in primer – 16.4) / total length of the primer)

   For salt adjustment:

   Tm = 100.5 + (0.41 * GC percent) - (820 / total length of oligo) + 16.6 * log10 ([Na+])

   Usually, the concentration of Na+ used is 0.05M and the equation is as following:

   Tm = 100.5 + (0.41 * GC percent) - (820 / total length of oligo) + 16.6 * log10 (0.05)

   GC percent = 100 * (nG+nC) / (nA+nT+nG+nC); n is number of nucleotides

3. The options for selecting probes have been added independent from the primers.

4. The output file structure has been modified. The reverse primers will be written directly without mentioning the 3` to 5` sequence.

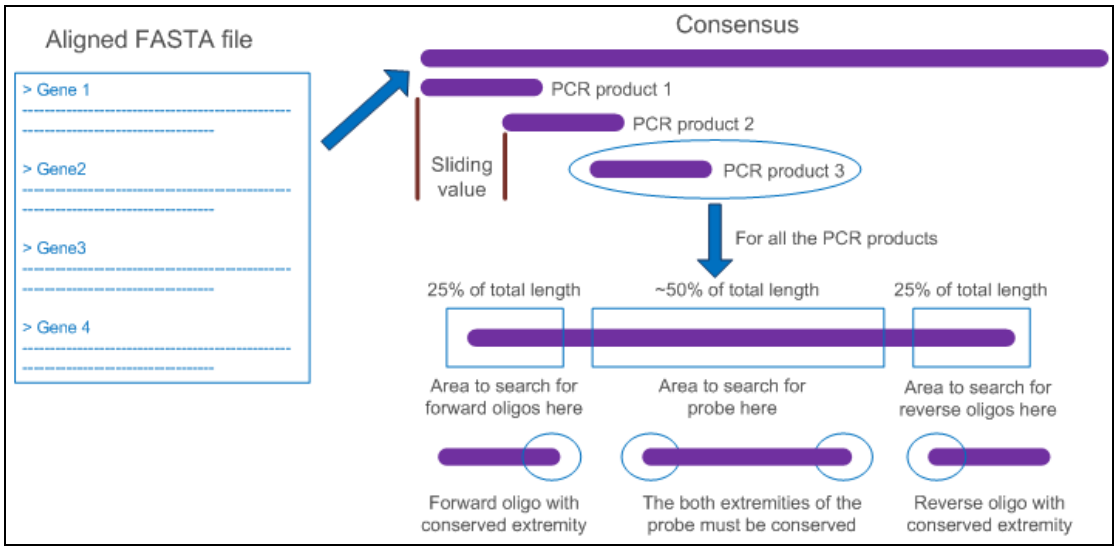
Method

* You may need to read the original article and supplementary file for details.

Once the input file (the multiple alignments in FASTA format) has been loaded into the program, the consensus sequence will be built. Gemi accepts the degenerated nucleotides (not As, Cs, Gs or Ts).

There are two approaches to search for oligos with the Gemi tool:

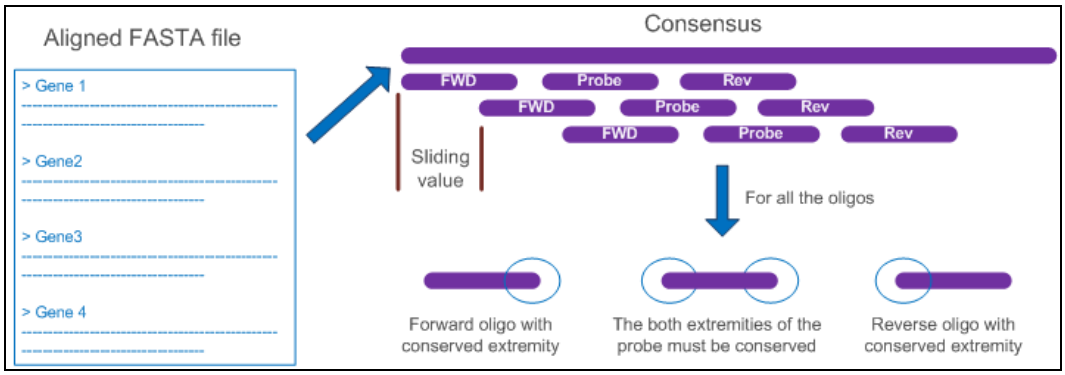
1. The First option is used to mine for full system, including PCR product with forward and reverse primers, and probe in case of real-time PCR. The user can control the length of the PCR product, the length of the oligos and their dissociation temperatures (Td) from the main window by editing the values on the specified text area. The final output file contains the full sequence of the PCR products on one hand, and the sequences of the primers and probes, their length, their Td and the number of degenerated bases on each of the oligos. Then, the tool moves along the consensus and search for other PCR product, as illustrated on figure.



2. The second approach is to find all the oligos in the consensus sequence. Here, the tool does not construct PCR systems. Instead, it deals with the whole consensus as a PCR product and search on it for oligos. Then, the user can choose for any combination of the proposed oligos.

Firstly, the tool searches for all possible forward primers. The search process starts from the start position entered by the user on the main screen. Then, Gemi moves along the consensus sequence from a number of nucleotides corresponding to the sliding value, and it searches for new primers until the end of the consensus.

After the tool has finished mining for forward primers, it searches for reverse, then for probe if needed, using the same strategy that it has been mentioned above (Figure).



The system moves over the consensus by distance equivalent to the sliding value entered by user on the main window. Therefore, each two searches are separated from each others by this sliding value or distance. For fussy search, it is recommended to use small sliding value. Small values are also recommended for sequences with a high variability.

The default criteria to choose primers and/or probes are containing three conserved bases at 3' end of the fwd, at 5' end of the rev and at both extremities in case of the probe.

On the other hand, the oligo usually contains no more than 3 degenerated bases. The user can determine the number of the degenerated bases in each oligo on the main window.

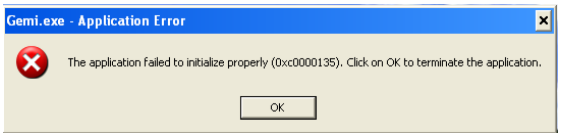
The temperature calculation obeys the equation: Td = 2 * (#A+#T) + 4 * (#C+#G), where, "#" refers to the number of As, Cs, Gs or Ts in the oligo. In case of degenerated bases, the tool calculates the minimum Td and maximum Td. If the bases harbored at a same position in the multiple alignment are complementary (A/T or C/G), the minimum and maximum Td are the same. Otherwise, the tool calculates the Td for A/T by multiplying the base by 2°C and reports it as minimum Td, whereas the maximum Td refers to C/G and is multiplied by 4°C.

For longer nucleotide sequences the Td follow this equation: 64.9°C + (41°C * (number of G's and C's in primer – 16.4) / total length of the primer). The Tm = 100.5 + (0.41 * GC percent) - (820 / total length of oligo) + 16.6 * log10 (0.05).

How to use Gemi?

Gemi is portable and ready-to-use program, no need to install it. You can run the particular program by double clicking on the binary files (.exe) on the package.

If you got the following message, it means that you need to install .Net framework 2.0 or later from Microsoft™ website, figure.



Input file:

Gemi accepts only FASTA format files, the figure. The file should be previously aligned by Multiple Sequence Alignment (MSA) software, like:

CLUSTAL (http://www.clustal.org/) program or

MUSCLE (http://www.drive5.com/muscle/) program

For other tools and more information, please visit EBI website: http://www.ebi.ac.uk/Tools/msa/

```
> Gene 1
GACTSTA---AACGCW ...

> Gene 2
GACTGTA---AGCGCW ...

> Gene 3
GACTCTA---AACGCW ...
...
...
```

The characters should be upper-cased. If you got sequences in low-case, you may use ConvertToFASTA module on the Gemi package.

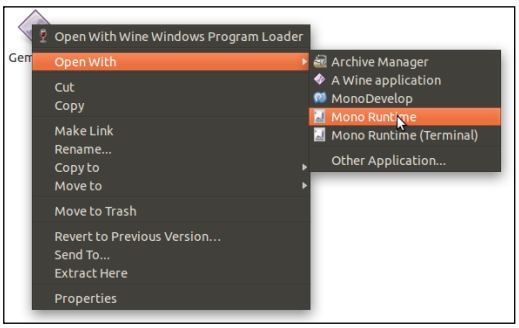
You may also use ConvertToFASTA module to convert the 'gde',' phy' or 'aln' file formats to upper-case FASTA one.

Basic Local Alignment Search Tool 'NCBI-BLAST' (http://blast.ncbi.nlm.nih.gov/Blast.cgi) is useful to group the orthologous genes with similar sequences together, which may ease the finding of oligos.

Running Gemi on Linux, Ubuntu or others:

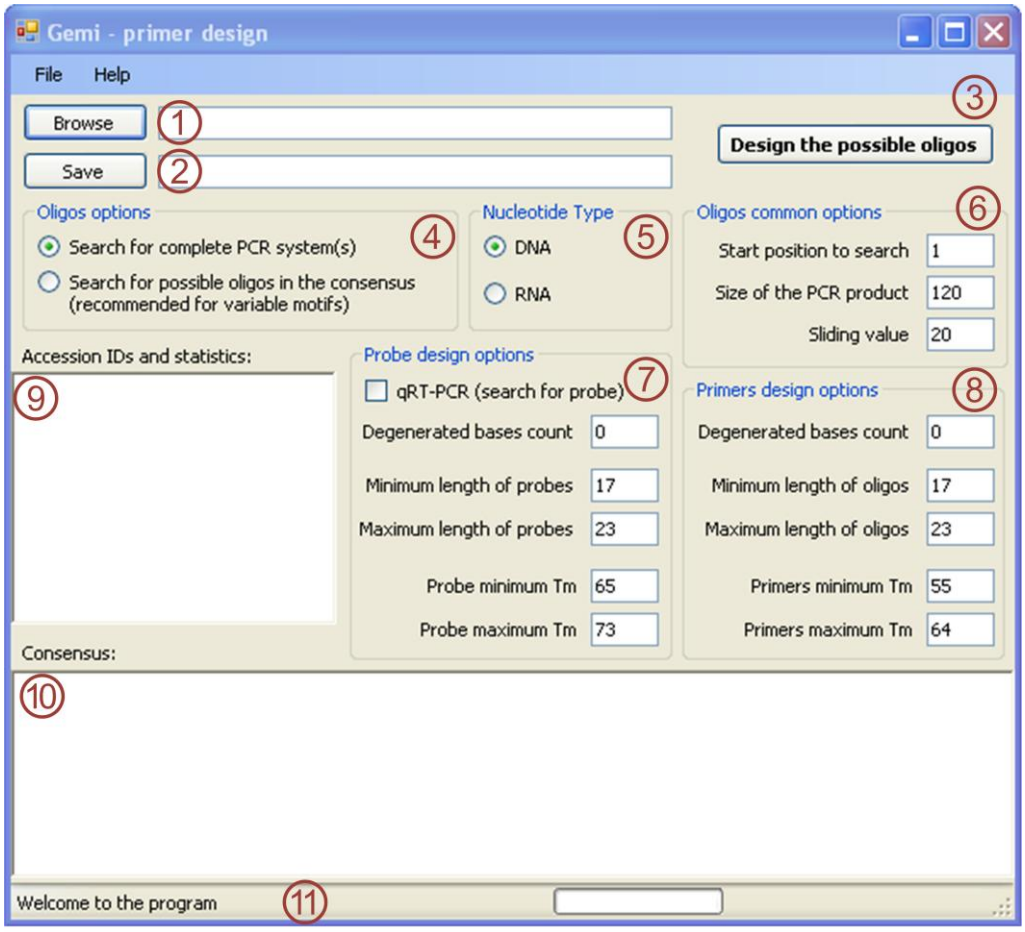
To run Gemi on operating system other than Windows, you need install **Mono-develop** software, as mentioned before. As in figure, **right click** on the application (.exe) file -> **Open With** and then choose **Mono Runtime** option from the menu. This will let the program works as in the windows atmosphere.

Please, note that the **Wine** software can not load this type of applications.



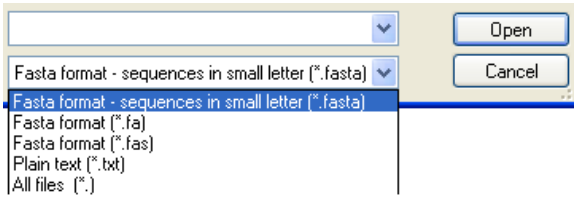
Primer design module:

The main window of the program, as shown in the figure, contains the initial search options or parameter settings, by which the user manages the resulted oligos, figure. The tool then uses these parameters as thresholds to design the primers.



1. By clicking '**Browse**' button, the user is able to explore the local drives and search for the input FASTA file. This may be also accessed from the main menu, select **File** -> **Open** or hit (**Ctrl+O**), (**Cmd+O** in Mac OS) on the keyboard to open the file. By clicking '**Browse**' button, the user is able to explore the local drives and search for the input FASTA file. This may be also accessed from the main menu, select **File** -> **Open** or hit (**Ctrl+O**), (**Cmd+O** in Mac OS) on the keyboard to open the file.

2. '**Save**' button can be used to place the final output file. The user chooses the destination of the file and then enters its name. Clicking on **File** -> **Save** or hit **Ctrl+S**.



3. Press the '**Design the possible oligos**' button to report the oligos into the output text file in the directory previously specified by clicking '**Save**' button.

4. The user may searches for the complete system with full PCR system and the oligos inside it by ticking on the '**Search for full PCR system(s)**' option. The tool will move upstream over the consensus by the sliding value constructing new PCR product and searches for oligos inside it. This method is compatible with low variability sequences or those contain few numbers of the degenerated bases. The other option is to tick '**Search for possible oligos in the consensus**'. Here, the system will list all the possible oligos and probes for qRT-PCR. The tool searches for fwd primer then moves for finding the second and so on, then searches for rev and probe. This method is useful for high variable sequences with several degenerated bases.

5. Check the nucleotide type (DNA or RNA) before choosing the input file and pass it to the tool. Switching between them can be easily done by clicking the corresponding check-button. It is important to check first the type of sequences. If the file is RNA and the user ticks DNA, an error will be fired. The tool accepts IUPAC-IUB symbols only.

6. The tool starts searching for oligos from a position on the consensus equal to the '**Start position to search**' value. '**Size of the PCR product'** is designed to control the length of the PCR system and the distance between fwd and rev primers. The tool searches for oligos within this product and report it with the olgios in the final file. This option is valid only for the '**Search for full PCR system(s)**' option. The '**Sliding value**' is the value by which the tool will step or move over the consensus and searches for a new oligo. For high variable sequences, with several degenerated bases, it is recommended to use small value.

7 & 8. IUPAC base is degenerated base (non-A/C/G/T base). '**Degenerated bases count**' controls the numbers degenerated base in the final oligos. Zero means no degenerated base at all, and oligos will contain A, C, G and T bases only.

The tool searches for oligos longer than the '**Minimum length of oligos**' or '**Minimum length of probes**' value, but shorter than the '**Maximum length of oligos**' or '**Maximum length of probes**' value. This may be decimal values.

The oligos temperatures are ranging between the values '**Primers minimum Tm**' and '**Primers maximum Tm**' for primers; and '**Probe minimum Tm**' and '**Probe maximum Tm**' values for probe [decimal values can be accepted here].

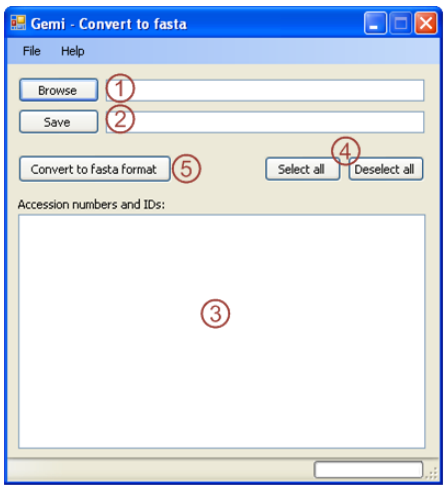
9. In textbox the statistics on the number of sequences, total length of consensus length, and number of conserved bases (A, C, G & T) will be written.

10. Once the input file has been selected, the consensus will be created automatically. The **statics** and the **consensus sequence** will be written into this text area.

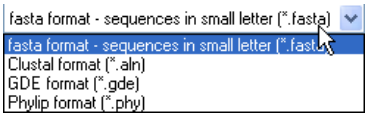
11. The status of the process and its progress will be written in this '**status bar**'.

ConvertToFasta module:

The main module of Gemi accepts only aligned multiple sequences in FASTA format andupper-cased sequences (capital letters). This module is a supplementary tool to convert FASTA files with low-case (small letters), Clustal format (.aln), Phylip format (.phy) and GDE one (.gde) to the upper-cased FASTA one.



1. By clicking '**Browse**' button, the user is able to explore the local drives and search for the input file. This may be also accessed from the main menu, select **File** -> **Open** or hit (**Ctrl+O**), (**Cmd+O** in Mac OS) on the keyboard to open the file. If the file does not appear, you may change it from the open file dialogue. As in figure and choose the file type from 'Files of types' box.



2. '**Save**' button can be used to place the final output file. The user chooses the destination of the file and then enters its name. Clicking on **File** -> **Save** or hit **Ctrl+S**.

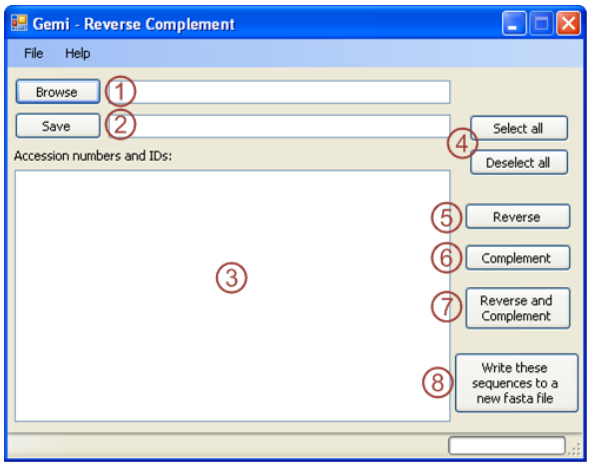
3. Once the input file has been selected, it automatically uploaded to the tool and the accession numbers will appear in the text area. The user may choose certain sequences to be converted from this area.

4. You may select specific sequences to be converted from the previous area. If you wish to select all of them or deselect all of them simple click one of these buttons.

5. The tool ready know to run and convert the sequences. Just click this button and the output will be saved to the specified directory.

ReverseComplement module:

In some cases, getting the reverse and complement counterpart of some sequences will help to find good universal primer. This module can help in this process. Moreover, extracting certain sequences from FASTA file can be done by this module.



1. By clicking '**Browse**' button, the user is able to explore the local drives and search for the input file. This may be also accessed from the main menu, select **File** -> **Open** or hit (**Ctrl+O**), (**Cmd+O** in Mac OS) on the keyboard to open the file. The formats allowed are (.fa), (.fas) or (.fasta). However, the aligned sequences are accepted but not recommended, using raw sequences file and run alignment tool will be good option.

2. '**Save**' button can be used to place the final output file. The user chooses the destination of the file and then enters its name. Clicking on **File** -> **Save** or hit **Ctrl+S**.

3. Once the input file has been selected, it automatically uploaded to the tool and the accession numbers will appear in the text area. The user may choose certain sequences to be converted from this area.

4. You may select specific sequences to be converted from the previous area. If you wish to select all of them or deselect all of them simple click one of these buttons.

5. Gets the reverse counterpart of selected sequences and writes to output file.

6. Gets the complement counterpart of selected sequences and writes to output file.

7. Gets the reverse/complement counterpart of selected sequences and writes to output file.

8. The button allows the user to write the selected sequences in the output file.

Output

The final output will be written in a tabulated text file format in the directory specified by the user at the beginning. The following table shows the headers of the columns in the final report with their representation.

Column header	Meaning
Position_of_oligo	The coordinate (start-end) of oligo on the consensus
Type_of_oligos	Forward (fwd), reverse (rev) or probe
Oligo	The sequence of the oligo
Oligo_Length	The length of the oligo
Degenerated_bases_count	Number of the IUPAC or degenerated bases (non-A/C/G/T)
Conservation	The percentage of the conserved bases (A/C/G/T), the value is '1' for all ACGT. Conservation = number of (ACGT) bases / the total length of the oligo.
CG	Number of the C and G bases on the oligo; S base always considered as C/G topology, while W as A/T one.
CG%	The percentage of the C and G bases on the oligo
CG_after_adding_ degenerated_bases	The number of the C and G bases after adding the degenerated symbols that represent C/G bases, for example, K, M, R, Y, B, D, H, V and N
Minimum_Td	The minimum temperature, on which the degenerated symbols are counted as A/T topology
Maximum_Td	The maximum temperature, on which the degenerated symbols are counted as C/G topology
Minimum_Tm	The minimum temperature after salt adjustment, on which the degenerated symbols are counted as A/T topology
Maximum_Tm	The maximum temperature after salt adjustment, on which the degenerated symbols are counted as C/G topology

Caveats and notes

1. The program is portable application. Therefore, there is no need to install it into your computer. You also can run it from your USB drive.

2. Some antivirus software blocks all (.exe) files. This may prevent the program from running.

3. The program constructs consensus sequence from multiple/aligned-sequences, therefore:

    a. Do NOT use format other than FASTA one.

    b. Do use ALIGNED sequences.

4. In case of single sequence, the tool deals with it as consensus and report oligos based on it. This allows the user may skip consensus constructing step.

5. Using BLAST tool will help to find the orthologous genes, while CLUSTAL or MUSCLE helps to align the sequences, these tools will help to find better matches of the oligos.

6. Avoid using the same names as gene identifiers or accession numbers.

7. The file extension can be viewed by changing the configuration of the local operating system, which is different from one to another.

8. The software writes a text file into your local desktop. You can choose the directory to save it from the main window. Therefore, please be sure that you have administrator privilege to write the file into output directory. In some computer, writing on "C:/" drive is not allowed, so please check with your computer administrator. It also may be saved to USB drive instead of local drives.

9. Entering numbers sometimes is tricky since some languages use comma (,) as a decimal numbers identifier, while others use (.) in a decimal numbers (e.g., 1.0 or 2.5). So, please check the language settings in your control panel or enter all integers (1 or 2).

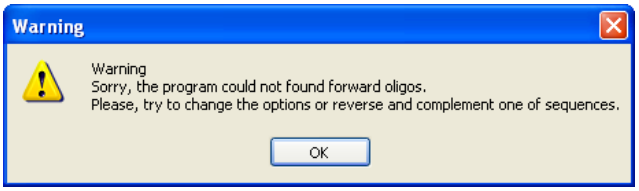
10. If you did not find oligos from first shot, just relax and also relax you search parameter. In case of the diversified sequences, you may need to change the sliding value, length of PCR product or oligos, number of degenerated bases or the temperatures. The small sliding window is designed for fussy search and for variable sequences. So, please use it freely.

11. It is important to notice that you may get more oligos by relaxing parameter settings or changing it. Also, switching between '**Search for full PCR system(s)**' and '**Search for possible oligos in the consensus**' options helps in finding more oligos.

12. Till this version Gemi does not resolve variability issue of oligos or graphical aids, which will be considered in the next versions together with further improvement in the method for searching the oligos.

If I did not find any oligo, what should I do?

The system uses the basic input parameters to search for the possible oligos on the consensus. If there is no complete PCR system or no one or more primers, an error message will be appeared indicating the absence of the oligos, like the following figure.



To overcome this problem you may do one or more of the following:

1. Try to relax your parameters, specially the number of the degenerated bases in the oligo. In general, it is accepted to use 3 bases per 20 bases length oligo.
2. Increasing the maximum length of the oligos could help to find them. The initial value in the tool is 23, but longer sequence will be also accepted for PCR test.
3. Using of '**Search for possible oligos in the consensus**' option is useful in finding oligos on high variable sequences.
4. The '**Sliding value**' is an important criterion for finding an oligo. The small value allows fussy searching through the sequence.
5. Changing the minimum and maximum temperature will affect the finding of the oligos.
6. In some cases, reverse (3' to 5') and/or complementary counterpart of one or more motifs may help to design oligos. For this process, Gemi package offers a module to retrieve the reverse, complement or reverse-complement counterpart of one or more sequences.

Citing to Gemi

Sobhy, H, Colson, P; Gemi: PCR primers prediction from multiple alignments, Comparative and Functional Genomics 2012:783138; doi: 10.1155/2012/783138; PubMed ID: 23316117.

Troubleshooting and contact

For troubleshooting, suggestions or questions please contact:

Haitham Sobhy, haithamsobhy@yahoo.com