EXPAR Template Sequence Analysis (ETSeq)

Documentation

Release 0.5.1
Last modified: February 12, 2012

This tool is designed to predict the performance of oligonucleotide templates used in the EXPonential Amplification Reaction (EXPAR).

EXPAR - General Assay Principle

EXPAR, an isothermal DNA amplification method, amplifies short oligonucleotides, called trigger X, at 55°C. The trigger oligonucleotides that initiate EXPAR can be enzymatically generated from specific sites within the targeted genomic DNA, and therefore represent the analyte. Exponential amplification of trigger X is facilitated by an EXPAR amplification template oligonucleotide provided in excess in the reaction. The EXPAR template contains two copies of the trigger reverse complement X', separated by the reverse complement of a nicking enzyme recognition site plus a required four base post-cut site spacer. Trigger X primes the template and is extended by a polymerase, which generates a double stranded 5'-GAGTC-3' on the top strand that is recognized by the nicking enzyme Nt.BstNBI. The nicking enzyme nicks the top strand four bases to the 3' end of its recognition sequence. This creates another copy of the oligonucleotide trigger X that melts off or is displaced from the amplification template. During EXPAR, templates with similar thermodynamic characteristics often exhibit very different trigger amplification rates, and differ in terms of propensity for non-specific background amplification.

Purpose of ETSeq

ETSeq enables the analysis of a user defined set of EXPAR template sequences to increase the percentage of likely good performers within a subset thereof to be screened experimentally.

For more information about EXPAR and ETSeq, please see:

<we will insert the reference to the NAR manuscript here>

For each entered sequence, the tool returns 7 parameters to the user. Three parameters, related to the thermodynamic behavior of the template, are calculated using the output of UNAFold.

For more information about Unafold:

Markham, N. R. & Zuker, M. (2008) UNAFold: software for nucleic acid folding and hybridization. In Keith, J. M., editor, *Bioinformatics, Volume II. Structure, Function and Applications*, number 453 in *Methods in Molecular Biology*, chapter 1, pages 3–31. Humana Press, Totowa, NJ. ISBN 978-1-60327-428-9.

The other four parameters relate to the sequence dependence of template performance. These four sequence dependence related parameters are generated by two machine learning methods: a position weight matrix based method, as described in <NAR ref>, and a naïve Bayes method, using the Orange Naïve Bayes module.

For more information about the Orange Naïve Bayes module:

Tomaž Curk, Janez Demšar, Qikai Xu, Gregor Leban, Uroš Petrovič, Ivan Bratko, Gad Shaulsky, Blaž Zupan. *Microarray data mining with visual programming*. Bioinformatics. 2005 Feb 1;21(3):396-8.

In addition, ETSeq suggests a set of possibly well performing EXPAR templates selected from the user provided set. If no sequence is returned, then the software determined that none of the provided sequences fall within the recommended parameter range.

Disclaimer: The classification of EXPAR template sequences using this tool is not perfect. This tool merely enhances the likelihood that a subset of chosen EXPAR template sequences will perform well, but this subset needs to be characterized experimentally to identify truly well-performing sequences.

Downloading and Installing:

ETSeq can be downloaded from: <URL goes here>

Operating system: This software tool has been designed for Windows 7 or XP having at least 512 Mb of RAM and 1Gb of hard drive space. Other operating systems have not yet been tested.

Other requirements: This tool integrates the UNAFold perl scripts to predict thermodynamic characteristics. Therefore, users need to install a perl compiler, such as Activeperl, on their computer before using this tool.

To download Activeperl:

http://www.activestate.com/activeperl/downloads

Installing the program: Unzip the file "ETseq.zip" to a designated folder. To check proper installation, double click the file "run.exe" in your unzipped designated folder. A user interface window will open, which means you have installed the tool successfully. An error message "Perl compiler required" indicates that a Perl compiler is either missing or not properly installed on the computer.

Running instruction:

1. Double click the file "Run.exe" in your ETSeq designated folder - a Graphical User Interface (GUI) window will open.
2. Copy and paste your template identifiers and sequences into the input area on the ETSeq GUI (see below for formatting requirements)
3. If desired, change the thermodynamic input parameters (or leave as default)
4. Click "Submit"

5. Run time can vary from a few seconds to minutes or longer, based on how many template sequences were submitted
6. Another dialog window will open when the analysis is completed, click “save”
7. Save the result as Excel spreadsheet (.xls format) to a chosen folder.

Input Format for EXPAR Template Sequences:

The program can accept an unrestricted number of EXPAR template input sequences at one time. However, run times may be long for large input files.

EXPAR template sequences must be entered in the 5’ to 3’ direction, and must consist of the following elements: (1) the reverse complement of the trigger sequence (at least 10 bases in length), (2) a 4-base post cut site, (3) GACTC, which is the reverse complement of the Nt.BstNBI nicking enzyme recognition site, and (4) a second copy of the reverse complement of the trigger sequence. The EXPAR template must have exactly one GACTC in its sequence. Only exact nucleotides (A, C, G, or T) are allowed.

Each EXPAR sequence requires a unique user defined sequence identifier (no blank spaces in user identifier). If multiple EXPAR template sequences have the same sequence identifier, then all but the first will be deleted by the program.

Enter the EXPAR template sequences to be analyzed into the input field. There are two acceptable input formats:

1. EXPAR template sequences plus sequence identifiers can be entered in fasta format (see http://blast.ncbi.nlm.nih.gov/blastcgihelp.shtml for a description of the FASTA format). Generate this file using another suitable computational tool, or using a simple text editor such as the Windows notepad accessory, then copy and paste into the input area. The input data would look as follows:

   >seq1

   CCTACGACTGAACAGACTCTCCTACGACTG

   >seq2

   CCTACGACTTAACAGACTCTCCTACGACTT

2. Generate an Excel file (or a tab delimited text file) with sequence identifiers in column 1, and the sequences in column 2. Copy and paste this data (without header rows) into the input area, resulting in a tab delimited format such as the following:

   seq1 CCTACGACTGAACAGACTCTCCTACGACTG

   seq2 CCTACGACTTAACAGACTCTCCTACGACTT

Other Input Parameters

Minimal trigger-template Tm

The Minimal melting temperature allowed for trigger- template duplex. Default: 40°C

Maximal trigger-template Tm

The Maximal melting temperature allowed for trigger- template duplex. Default: 60°C

Maximal template-template Tm
The Maximal melting temperature allowed for template- template duplex. Default: 25°C

Maximal template self-bonds
The Maximal number of hybridization bonds allowed for template- template hybridization. Default: 10

Output Format

An excel file containing three sheets will be generated by clicking the "save" button.

The sheet labeled "Good and desired thermo" contains all the templates that are predicted to be well performing by both PWM based method and naïve bayes method, and predicted to be have the desired thermodynamic characteristics as well.

"desired thermo" is for all the templates which are predicted to be have the desired thermodynamic characteristics.

"all submitted templates" is for all the templates which are submitted to the tool. The meaning of the header in the sheets are listed below:

Output Parameters

Each spreadsheet within the Excel output file, contains 9 columns with the following headers

name

The sequence ID

template_seq

The EXPAR template sequence

bayes_class

This is the template performance class predicted by the Naïve Bayes method. "Good" indicates fast amplification of trigger X, and decent temporal separation between specific amplification in the presence of trigger X, and non-specific amplification in the absence of trigger X. "Bad" indicates slow amplification and / or poor temporal separation between specific and non-specific amplification. The bayes_class is based on a Naïve Bayes classifier, obtained using the Orange Naïve Bayes module, that uses significant position motifs derived from previously characterized EXPAR template sequences. This Naïve Bayes classifier and the significant position motifs are embedded in the program.

pwm_class

This is the template performance class predicted by the position weight matrix based method. "Good" indicates fast amplification of trigger X, and decent temporal separation between specific amplification in the presence of trigger X, and non-specific amplification in the absence of trigger X. "Bad" indicates slow amplification and / or poor temporal separation between specific and non-specific amplification. The pwm_class is based on two PWMs and a support vector machine based classifier, which

are embedded in the program and have been derived from previously characterized EXPAR template sequences.

p90 score

This score is generated by the position weight matrix base method. A lower p90 score indicate faster amplification and vice versa. The P90 score is based on a PWM embedded in the program, which is derived from previously characterized EXPAR template sequences.

diff score

This score is generated by the position weight matrix base method. A lower diff score indicates decent temporal separation between specific and non-specific amplification. The diff score is based on a PWM embedded in the program, which is derived from previously characterized EXPAR template sequences.

tri-temp Tm

Trigger- template duplex melting temperature, assuming that the template oligonucleotide is present in excess in the reaction. This parameter is calculated from the UNAFold output values for ΔH and ΔS for trigger template hybridization.

temp-temp Tm

The template- template duplex melting temperature. In this case, the two strands are present in equal concentrations; therefore the output value for the template template melting temperature is used directly.

temp bonds

The number of template- template hybridization bonds.

Copyright Notice and Disclaimer

Citing ETSeq

We request that use of this software be cited in publications as:

<we will insert the reference to the NAR manuscript here>
Source code available at <we will insert the final url here>

Acknowledgments

The development of ETSeq was funded by the National Institutes of Health through award R01AI076247, plus an ARRA supplement to this award. This work was funded in part by the National Science Foundation's Frontiers in Integrative Biological Research grant FIBR-0527023, and NSF's BEACON Center for the Study of Evolution in Action under contract No. DBI-0939454.