TECHNISCHE
UNIVERSITÄT
WIEN
Vienna University of Technology

ivFAST

Contents

1 What is ivFAST?

ivFAST (*in vivo* Footprinting Analysis Software Tool) is a command line-based program to map peaks obtained from capillary gel electrophoresis to a DNA sequence and to make a pairwise comparison of the peak areas from different samples for each sequence position. The mapping and the results of the calculations are provided as text files. For easier visual analysis of the data generated a heatmap is created, comprising three color levels for protected and hypersensitive bases.

2 How does ivFAST work?

ivFAST conducts the following steps:

1. Data import
   ivFAST imports two types of files: a FASTA-file, containing the according DNA sequence, and the sequencing files, containing the measured peaks with retention time and area, given in plain text format. For each sequenced sample three sequencing files have to be provided: one containing all sample peaks sequenced (also the primer artefacts), one containing only the internal size standard peaks, and one containing only the sample peaks to be analyzed.

2. Sequence processing
   The input sequence is always given as the coding strand from 5' to 3'. Therefore it has to be transferred to the analyzed strand and direction, which is 3' to 5' on the non-coding strand for forward primed sequencing, and 3' to 5' on the coding strand for reverse primed sequencing.

3. Mapping
   The program starts the mapping by assigning the first peak to a sequence position specified by the user. Further positions are calculated from the pairwise distances of consecutive peaks, which are rounded to integers. If the calculated position matches a base in the sequence, listed in the configuration file under 'validBases', it is taken for further calculation. Valid bases are bases, at which DNA can be cut dependent on the methylation and cleavage method used. If the position matches a base not listed in 'validBases' it is treated as background noise and removed from further processing. The result of the mapping is outputted as plain text file. Peaks identified as background are indicated with an asterisk. Background peaks occur frequently at the beginning, but seldom at the end of the analyzed region. If it is likely, that a peak is falsely identified as background peak, an additional notification is displayed on the command line, so the user can make manual corrections in the according sequencing file containing the peaks to evaluate (the file containing all peaks does not need to be changed).

4. Calculation
   First the single peak areas of the peaks to be analyzed are normalized. The normalization factor is calculated for each sample in three parts. The first is the share of the sample peaks in all peaks (including standard

peaks), which accounts for slightly varying ratios of sample amount to standard amount. The second is the share of true sample peaks (without primer artifacts) to all sample peaks, which accounts for different reaction efficiencies in the previous PCR. The third is the sum of all areas (standard and sample), which accounts for differences in overall fluorescent signal due to varying CGE analysis. With x_a an area from the file containing all sample peaks, x_s an area from the file containing the standard peaks, and x_p an area from the file containing the sample peaks to be analyzed, the normalized peak area of x_p^i is defined as

$$\hat{x}_p = x_p \; / \left(\frac{\sum_{i=0}^{m} x_a^i}{\sum_{i=0}^{m} x_a^i + \sum_{i=0}^{m} x_s^i} \cdot \frac{\sum_{i=l}^{m} x_p^i}{\sum_{i=0}^{m} x_p^i} \cdot \left(\sum_{i=0}^{m} x_a^i + \sum_{i=0}^{m} x_s^i \right) \right), \quad (1)$$

with m the total number of peaks in this sample and l indicating the first peak with a size greater than the primer length (as defined in the configuration file).

After normalization the n replicates of a sample (belonging to the same condition) are grouped together and for each normalized peak $\hat{x}_p$ the sample mean

$$\overline{x}_p = \frac{1}{n} \sum_{i=1}^{n} \hat{x}_p^i, \quad (2)$$

the sample variance

$$s_p^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\hat{x}_p^i - \overline{x}_p)^2, \quad (3)$$

and the confidence interval for the true mean μ_p

$$\overline{x}_p - \frac{t \cdot s_p}{\sqrt{n}} \leq \mu_p \leq \overline{x}_p + \frac{t \cdot s_p}{\sqrt{n}} \quad (4)$$

are calculated, based on the Student's t-distribution, with t obtained from the tabularized values of the confidence interval $F_{n-1}(t) = 0.95$ (recommended value, but adjustable in the configuration file).

Now the program does a pairwise comparison between the sample means $\overline{x}_p$ of the user-defined pairs of sample (S) and reference (R) conditions where for each peak it is checked, whether the two sample means $\overline{x}_p(S)$ and $\overline{x}_p(R)$ can be said to be different. As criterion for this non-overlapping confidence intervals of the sample means

$$\left(\overline{x}_p(S) \pm \frac{t(S) \cdot s_p(S)}{\sqrt{n(S)}} \right) \cap \left(\overline{x}_p(R) \pm \frac{t(R) \cdot s_p(R)}{\sqrt{n(R)}} \right) = 0 \quad (5)$$

was defined. If this is fulfilled, then the quotient sample/reference $\frac{\overline{x}_p(S)}{\overline{x}_p(R)}$ is built and assigned a color in the heatmap.

5. Output of the result
   The values of the calculated quotients $\frac{\overline{x}_p(S)}{\overline{x}_p(R)}$ are written in a plain text file (one for each user-defined reference condition). Based on these a heatmap

is created (one for all user-defined pairs of conditions S/R) where protected bases (with S/R < 1) are depicted in shades of red, hypersensitive bases (with S/R > 1) in shades of blue (Fig. 1).

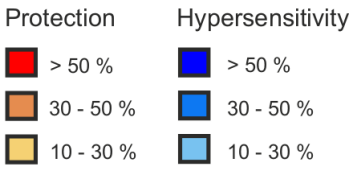

Figure 1: Legend heatmap

3 How to use ivFAST

3.1 Running the program

- Open the command line.
- Change to the directory where ivFAST.jar is located.
- Call the program with

  java -jar ivFAST.jar -d DIRECTORY [-s START] [-sd STARDIAGRAM] [-e END] [-p PAIRS]

  (arguments in brackets are optional)

  -d Directory that contains the input files.
  The program first performs a check to see if the file names comply with the specifications. If they do, it will try to read them. Make sure the directory contains only valid input files and only one sequence file, otherwise the program will produce a failure message and cancel the run.

  -s Sequence position where the program should start the mapping.
  The first peak of each data file will be assigned to this position. For files with direction argument 'f' (forward priming) a positive integer is expected, counting forward from the first character of the provided sequence. For files with direction argument 'r' (reverse priming) a negative integer is expected, counting backwards from the last character of the provided sequence.

  -sd Sequence position where the heatmap will start.
  The absolute value of STARDIAGRAM must be $\geq$ the absolute value of START. For files with direction argument 'f' (forward priming) a positive integer is expected, counting forward from the first character of the provided sequence. For files with direction argument 'r' (reverse priming) a negative integer is expected, counting backwards from the last character of the provided sequence.

-e Sequence position where the heatmap will end.
For files with direction argument ‘f‘ (forward priming) a positive integer is expected, counting forward from the first character of the provided sequence. For files with direction argument ‘r‘ (reverse priming) a negative integer is expected, counting backwards from the last character of the provided sequence.

-p Pairs of sample and references that should be calculated.
Syntax: Sample#/Reference#, Sample#/Reference#, ...
Samples are sorted alphabetically according to the content of the file name field ⟨CONDITION⟩ (see section 3.2.1), and numbered with integers, starting with 0.

- If all arguments have already been provided in the command line, the program will start immediately. Otherwise it will ask the user to provide the missing information. For the found conditions a numbered list will be provided in this case.
- The output of the calculation will be provided in a newly created subdirectory named with date and time, localized in the same directory as ivFAST.jar.

3.2 File specifications

3.2.1 Input files

All files must be plain text files and placed in one directory, which is given as parameter to the program. No other files are allowed in this directory.

Sequence file The user has to provide exactly one DNA sequence file, containing the DNA sequence to be analyzed, named

sequence_⟨NAME⟩.txt

There are no restrictions for the field ⟨NAME⟩. The content of this file must have a fasta-like format. This means that the first line is reserved for descriptions of any kind, whereas all following lines are expected to contain the DNA sequence. Line breaks within the sequence are ignored. Allowed characters are only ‘A‘, ‘C‘, ‘G‘ and ‘T‘ (not case-sensitive). General characters like ‘N‘, ‘Y‘, etc. are not accepted.

Data files The data files contain the measured fragment size and area from the capillary gel electrophoresis. Each peak has to be a separate line, where first the size and then, separated by a space or a tab, the area is given. As decimal separator ‘.‘ and ‘,‘ are both accepted. The file must not contain any headers.

The data files must be named

⟨ORGANISM⟩_⟨GENE⟩_⟨f|r⟩_⟨CONDITION⟩_⟨REPLICATE⟩_⟨a|p|s⟩.txt

The fields ⟨ORGANISM⟩ and ⟨GENE⟩ can contain any characters except ‘_‘, which is reserved as separator, but they must be identical for all data files in

the directory (case-sensitive). The content of these two fields will show up as title of the heatmap.

The field ⟨f|r⟩ has to contain either ‘f‘, if the forward primer was used for the analysis, or ‘f‘, if the reverse primer was used. In case it is ‘f‘, the program will start counting from the beginning of the provided sequence and will take the complement of it, to obtain the complementary strand to the primer. In case it is ‘r‘, the program will start counting from the end of the provided sequence and will take the reverse of it, to obtain the complementary strand to the primer.

The field ⟨CONDITION⟩ can contain any characters except ‘_‘ and indicates the different conditions that should be compared. The content of this field will be used to label the rows of the heatmap.

The field ⟨REPLICATE⟩ is used to distinguish replicates obtained from the same conditions. Since the program has to calculate the standard deviation for each condition, at least two files must be provided for each ⟨CONDITION⟩, whose names only differ in the field ⟨REPLICATE⟩. Any characters except ‘_‘ can be used in this field.

The field ⟨a|p|s⟩ is used to distinguish the files containing all sample peaks (‘a‘), only the sample peaks to be analysed (‘p‘), or only the standard peaks (‘s‘). For each replicate of a condition all these three fields have to be provided.

3.2.2 Output files

All output files are generated in a new directory named with date and time of the run.

For each given data file a mapping file is produced, which documents how the measured fragment sizes have been mapped to the provided sequence. Additionally it also contains the normalized areas for each peak. Asterisk indicate background peaks, i.e. peaks that occur at bases not listed in the field ‘validBases‘ in the configuration file, and therefore are neglected in the further calculation.

The result of the calculation is provided both as heatmap and as text files. There will be one heatmap for all compared pairs of sample and reference, but one text file for each reference. The given deviations represent the fraction of normalized sample area divided by normalized reference area. Additionally a file is generated that contains the sum over all areas for each sample, as well as the mean and standard deviation of this sum over all samples. A standard deviation of up to 15 % is acceptable to ensure proper normalization.

3.2.3 Configuration file

The configuration file config.properties is located in the folder config, which has to be placed in the same directory as ivFAST.jar. It contains calculation parameters, that can be adjusted by the user, which are: the type of bases at which the DNA can be cut, the probability for the t-distribution and the color ranges used in the heatmap. The default file content is:

validBases=AG
primerLength=40
probability=0.95
lowerRange=0.1
middleRange=0.3
upperRange=0.5

The field ‘validBases‘ specifies at which bases the DNA can be cut and is used by the program to find the valid positions for the mapping in the provided DNA sequence. Valid entries are $\in$ {A,C,G,T}, listed without deliminator. In the default specification it is assumed that the DNA can be cut at the bases ‘A‘ and ‘G‘, so each ‘A‘ and ‘G‘ in the DNA sequence is treated as a valid position for the mapping, whereas peaks mapped to a ‘C‘ or ‘T‘ are treated as background peaks.

The field ‘primerLength‘ specifies the length of the primers used for the PCR. It is used to define a cut-off size for small fragment artefacts in the sequencing, that do not correspond to meaningful peaks. Varying this parameter affects the normalization (see chapter 2, Calculation for details).

The field ‘probability‘ (α) specifies the width of the confidence interval (see eqn. 4) and is given as two-sided probability

$$\alpha = F_{n-1}(t) - F_{n-1}(-t) = 0.95 \tag{6}$$

(see also Fig. 2). It must hold a value $\in (0, 1)$. Varying this parameter affects the probability of two peaks being considered as different (see chapter 2, Calculation for details)

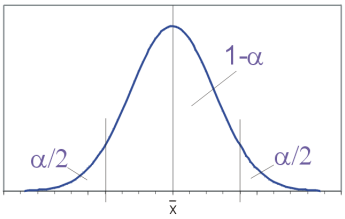

Figure 2: Confidence interval of $\overline{x}$ (sample mean) with probability α. 1-α and $\alpha/2$ denote the portion of the area included in the according region. (Source: Philipendula at the German language Wikipedia, GNU-FDL)

The fields ‘lowerRange‘, ‘middleRange‘ and ‘upperRange‘ are used to define the color ranges of the heatmap (see Tab. 1 and Fig. 1) and can take values $\in (0, 1)$, with the restraint lowerRange $<$ middleRange $<$ upperRange.

3.3 Manual adjustments

Background peaks occur frequently for small fragment sizes, but are unlikely for longer fragments. Normally, when a peak in the higher fragment size region is mapped to an invalid position (i.e. considered as background), this is a mapping failure. It occurs due to small inaccuracies in the size determination that can accumulate, when some bases are skipped between two peaks. The program will

Color	**RGB code**	**Range**
dark red	(255,0,0)	$S/R < 1 - upperRange$
middle red	(230,140,80)	$1 - upperRange \leq S/R < 1 - middleRange$
light red	(245,208,115)	$1 - middleRange \leq S/R < 1 - lowerRange$
light blue	(120,194,240)	$1 + lowerRange < S/R \leq 1 + middleRange$
middle blue	(14,121,242)	$1 + middleRange < S/R \leq 1 + upperRange$
dark blue	(0,0,255)	$1 + upperRange < S/R$

Table 1: Color definitions of the heatmap.

automatically detetect such possible mapping failures and produce a warning message on the comand-line. But the program is not able to correct these issues automatically, since it cannot judge wether it is a real failure or not. Therefore the user has to correct the according size values in the seqencing file manually (only in the file with the suffix 'p') and repeat the run.

Here is an example, how this is done:
Assuming we have an excerpt of a sequence, with the masses obtained from the sequencing file, as shown in Table 2. Given is the way how it should be mapped. But when calculating the difference between the two masses it is 3.29, which would be rounded to 3, so the calculated position for the second peak would not be the 'G' but the 'T', since the program counts the rounded mass difference between two peaks forward to determine the next position.

A	C	C	T	G
136,83				140,12

Table 2: Example for manual corrections.

In order to get the correct result, the user has to change the values of the size in the sequencing file manually, e.g. to 136.72 and 140.22. Now the difference is 3.50, which will be rounded to 4 and therefore will yield the correct mapping. But be careful not to alter the distances to the next surrounding peaks when changing the values!

After all changes have been done and saved, the program has to be run a second time in order to get the correct mapping results.

4 Credits

This program was written at the Gene Technology Group at the Institute of Chemical Engineering, Vienna University of Technology. For citation please refer to the original paper 'A highly sensitive *in vivo* footprinting technique for

condition-dependent identification of *cis* elements', written by Rita Gorsche, Birgit Jovanović, Loreta Gudynaite-Savitch, Robert L. Mach, and Astrid R. Mach-Aigner, submitted to Nucleic Acids Research.