Quick Start Guide for Revised Tfit

Michael Gohde

May 21, 2018

Abstract

Tfit is a tool that attempts to find useful parameters in sequencing datasets.

1 Building Tfit-revisions

In order to start working tiwh Tfit, it is first necessary to build it. Tfit depends on the following libraries to be built:

1. MPI (can be provided through OpenMPI)
2. GCC >=5.0
3. mpic++ (provided through OpenMPI)
4. (optional) docker

In a shared environment such as Fiji, all that is necessary to build and run the latest version of Tfit is to:

1. chdir into the src directory
2. run 'module load gcc/7.1.0'
3. run 'module load openmpi'
4. run 'make -j(some appropriate number of jobs)'

On a personal machine, these requirements can be satisfied through appropriate package installations. All that is necessary in that case is to follow the first and last steps specified above.

2 Invoking Tfit

Tfit has several modes of operation. Each mode is referred to as a module in all documentation, though they all share a number of input parameters.

Invoking Tfit without any parameters will result in the following:

```
Usage: TFit modulename [arguments]
Where modulename is one of the following:
       bidir - This module searches the genome for areas
          resembling bidirectional transcription
              by comparing a fixed template mixture model
                   to a noise model by a log-likelihood
                  ratio score.
       bidir_old - This module implements the
          functionality seen in versions of Tfit used in
           publications
       model - This module attempts to generate an
          optimal set of parameters per region instead
              of using a fixed set of parameters for the
                  entire genome


Where [arguments] is one or more of the following for the
     bidir module:
          -i     Forward bedgraph file
          -j     Reverse bedgraph file.
          -ij    Both forward and reverse bedgraph file.
             This parameter may be used in place of -i and
             -j if reads are in one bedgraph file.
          -N     Job name.
          -o     Output directory. If it does not exist, it
             will be created.
          -log_out      Log file output directory. If it
             does not exist, it will be created.

Additional (optional) parameters for the bidir module:
       -tss   Transcription model path. Models are
          provided for hg19 and mm10 in the annotations
          directory of this project.
       -chr   Run bidir only on the specified chromosome
```

```
          by name. The default is, "all"
       -bct   LLR threshold. Default=1

If -tss is not specified, the values normally specified
   by the model file can be specified explicitly using
   the following:
       -lambda This is the entry length parameter for the
            EMG density function. default=200 bp
       -sigma This is the variance parameter for the EMG
          density function. default=10bp
       -pi    This is the strand bias parameter for the
          EMG density function. default=0.5
       -w     This is the pausing probability parameter
          for the EMG denisty function. default=0.5
       -scores Some form of score output file. This
          parameter is presently undocumented.
       -r_mu  Some classification parameter. Default=0.
          This parameter is presently undocumented.
       -ms_pen Penalty term for model selection. Default
          =1.
       -max_noise    Maximum noise threshold. Default
          =0.05. This parameter is presently
          undocumented.
       -FDR   Generate a likelihood score distribution on
            the input data. This parameter has yet to be
          fully documented and tested.


Where [arguments] is one or more of the following for the
    model module:
       -i     Forward bedgraph file
       -j     Reverse bedgraph file
       -ij    Both forward and reverse bedgraph file.
          This parameter may be used in place of -i and
          -j if reads are in one bedgraph file.
       -k     Bedgraph file containing a set of regions
          of interest.
       -N     Job name.
       -o     Output directory. If it does not exist, it
          will be created.
```

```
       -log_out      Log file output directory. If it
          does not exist, it will be created.

Additional (optional) parameters for the model module:
       -mink  Minimum number of finite mixtures to
          consider. default=1
       -maxk  Maximum number of finite mixtures to
          consider. default=1
       -rounds Number of random seeds to use in the model
          . default=5
       -ct    Convergence threshold after which
          processing stops. default=0.0001
       -mi    Maximum number of model iterations after
          which processing stops. default=2000
The model module currently has experimental support for
   parameter inference. Ie. it can attempt to estimate
   lambda,
sigma, pi, and w via conjugate priors. These values are
   specified using the following parameters:
       -ALPHA_0      Prior parameter 1 for sigma.
          Recommended value=1
       -BETA_0 Prior parameter 2 for sigma. Recommended
          value=1
       -ALPHA_1      Prior parameter 1 for lambda.
       -BETA_1 Prior parameter 2 for lambda.
       -ALPHA_2      Symmetric prior on mixing weights.
          Higher values=stronger attempt to find
          components of equal mixing weights.
                     Recommended value=100
       -ALPHA_3      Symmetric prior on the strand bias.
          Higher values=stronger attempt to find
          bidirectional events with equal strand bias.
                     Recommended value=100
-elon    : (boolean integer) adjust support of elongation
    component, (default=0)
            useful only when fitting to FStitch[1] or
               groHMM[2] output intervals
Exiting...
```

The above is Tfit's *usage statement*. This presents a quick overview of all of Tfit's modules and their parameters. From it, we can deter-

mine that there are presently three available modules:

1. bidir – This module attempts to fit a distribution to existing data.
2. bidir_old – This module is the same as bidir, except it attempts to behave more similarly to early versions of Tfit.
3. model – This module attempts to model every given bidirectional in a dataset individually instead of applying precomputed parameters.

From the usage statement, we can begin to see how Tfit is to be invoked. It is first necessary to call Tfit, then to specify the name of a module. After this, various arguments for that module are specified.

2.1 Invoking the Bidir Module

The bidir module requires, at minimum, the following information to run:

1. An input bedgraph (this may be two bedgraphs if each covers only one strand).
2. An output directory (this will default to wherever Tfit is invoked from). This is specified with the '-o' parameter.

While it is possible to run Tfit with only the minimum number of parameters (technically just a bedgraph file specification), it may produce unexpected, difficult to find, or incorrect results. As such, it is advised to further specify the following:

1. A name for the Tfit job. This is used to name various output files. This is specified with the '-N' parameter.
2. A transcription model. Tfit uses this file to infer parameters about a given dataset. Models are provided for hg19 and mm10 in the annotations directory of the Tfit repository. This is specified with the '-tss' parameter.
3. A chromosome name. By default, Tfit will attempt to fit every chromosome in a given dataset, which tends to be extremely computationally intensive. This parameter matches chromosome specifications in a bedgraph (eg. 'chr1' etc.) This is specified with the '-chr' parameter.

4. Whether to run the model module on bidir's inferences. It is possible to have Tfit automatically run the model module after inferring information about a given dataset. This feature tends to be very useful when trying to find regions of transcription as quickly as possible. If this is to be enabled, then Tfit must be invoked with '-mle 1'.

Input bedgraphs are specified with the -i, -j, and -ij parameters. -i specifies a bedgraph with only forward strand data, while -j specifies a bedgraph with only reverse strand data. -ij can be used in place of -i or -j, as it specifies a bedgraph file with both forward and reverse strand data.

Given the above information, it is possible to construct an example Tfit invocation:

```
Tfit bidir -ij /path/to/my.bed -o /path/to/outdir -N
    myjob
     -tss hg19.tss -chr chr1 -mle 1
```

2.2 Invoking the Model Module

The model module requires, at minimum, the following information to run:

1. An input bedgraph file.
2. A bedgraph file specifying regions of interest. This is specified with the '-k' parameter.

While it is possible to run Tfit's model module with only the minimum parameters listed above, it may be helpful to specify some of the optional parameters listed below:

1. A name for the Tfit job. This is specified with the '-N' parameter.
2. A file output directory. This is specified with the '-o' parameter.
3. A maximum number of iterations to compute. This is useful when the model module doesn't converge immediately, and it is specified with the '-mi' parameter.
4. A convergence threshold. This is useful when the model module doesn't converge or when a tighter tolerance is necessary. This is specified with the '-ct' parameter.

There exist several other parameters, the behaviors of which have not been explored by the author of this document at present. Unfortunately, the bidir module has been the topic of significantly more development, though further documentation for the model module will be written later.

It further appears that the '-k' parameter may be able to use the results from an invocation of the bidir module, though further testing is n3ecessary to confirm that this is the case.

Based on the above listings, it is possible to construct an example Tfit invocation:

```
Tfit model -ij /path/to/my.bed -k /path/to/regions.bed
    -o outdir -N myjob
```