Documentation for Tfit

Michael Gohde

May 28, 2018

Abstract

Tfit is a tool that attempts to find useful parameters in sequencing datasets. This document describes the usage of Tfit in more detail than the quick start guide based on observed behavior.

Contents

1 Building and Invoking Tfit-revisions

Please see the quick start guide for documentation on building and invoking Tfit. This guide exists to detail various module behaviors discussed in the quick start guide.

2 The bidir_old module

During the course of Tfit's development history, several changes were made to how the bidir module operates. While some of these changes in behavior were desirable to the overall function of Tfit, others tended to lessen its utility in a broad context. As such, the bidir_old module

was introduced to implement behavior from a known good version of Tfit such that users could choose whether they wished for Tfit to behave more like it did on publication while retaining the user interface and other backend improvements made more recently.

It appears that bidir_old, when passed the same parameters as the current bidir module, will tend to generate a substantially higher number of calls in exchange for low overall confidence. As such, it seems that this module behaves more as a filter for the remainder of Tfit. This change in behavior is explained by the following list of changes:

1. The old code sets the values of various passed parameters to those obtained from its computed average model if and only if it is passed a training file. Otherwise, it retains the values passed. This should not affect how bedgraphs are loaded later in the module, as neither the *br* nor *ns* parameters are altered.

2. The old code does not have behavior to honor the *fdr* parameter. As such, it does not compute a likelihood score distribution. It appears that the new codebase alters the behavior of its template matching algorithm based on the results of the *fdr* parameter.

3. As noted in the previous item, the old code base does not have a means of computing a likelihood score distribution. As such, instead of doing so, it *should* simply use any parameters passed to Tfit. Further analysis is necessary.

3 The bidir module

The bidir module implements bidirectional scanning from the most recent Tfit code pushed to Github. As such, it tends to perform fairly well at choosing good bidirectional reference points, though it may tend to underfit the data provided. This is likely due to two of its changes relative to older versions of the codebase:

1. The new bidir module computes an average score distribution to potentially filter out "bad" calls, or those with low scores relative to the rest.

2. The new bidir module makes use of a modified backend to achieve the first point.

Unfortunately, little investigation into the exact backend changes made to achieve these changes has been conducted. As such, this is a topic of significant further review.

4   The model module