DEVELOPER-LEVEL DOCUMENTATION

NILS GUBELA

Contents

1. Introduction

The following document explains the code of pause_finding.py. For additional informations on the usage or the purpose of the program consult the readme or the project report.

2. Overview

For each sequence structure pair the following steps are performed:

(1) Call DrTransformer to obtain occupancy of structure without pausing.
(2) Create step set of possible pausing sites: remove sites at 5' end without alternative structures and remove last five nucleotides from 3' end.
(3) Perform adaptive walk or exhaustive search (only differ in step selection and break criterion):
  (a) Select step which is the next candidate for a pausing site
  (b) Calculate equilibrium distribution at this site to check if pausing can have an effect. Go back to step (a) if equilibrium is reached.
  (c) Call DrTransformer with candidate pausing site.
  (d) Read DrTransformer input and save new occupancy at transcription end.
  (e) If occupancy is increased, save step. Go back to (a)
(4) Prepare output: print out to command line and save results to csv file

3. Processing files

3.1. **DrTransformer logfiles.** The function *dr_list* takes a filename as input and returns two lists: one with the transcriptional information of the DrTransformer logfile and one with the distribution at the end. It removes all comments starting with # from the logfile. There are two components in the DrTransformer logfile that are divided by a comment line. The first contains all transcription steps and is saved to *liste*, the second contains the end distribution and is saved to the list *liste_end*. The input file is opened, and each line is processed. There is a keyword that indicates that the next line starts with the contents of the transcription and a second keyword that indicates that the end distribution starts in the next line. These keywords are used to control which lines are added to which list. Next, a cleaning step is performed which removes characters like "\n" from the strings.

3.2. **Input file.** The input file needs to be a text file that contains sequence structure pairs. In each line, we expect a sequence string followed by a space character and then the structure in dot bracket notation. The file is read in and saved to *meta*. For each line in *meta* the steps described in *Initialization* and *Search* are executed.

4. Initialization

DrTransformer is called to obtain the initial occupancy of the structure without pausing sites. Starting from the 5' end the steps that have no structural alternative are excluded from the step set. The last five nucleotides at the 3' end are also excluded. The step set contains all possible pausing sites. During the search it is queried via the index *step*.

5. Search

5.1. **Step selection and break criterion.** The selection of steps and the break criteria are dependent on the mode of search:

5.1.1. *Adaptive walk.* In the adaptive walk mode each step is selected at random from the step set. The control variable *step* is increased by 1 if the selected step is no improvement to the occupancy and it is set to zero if the step is an improvement. One of two criterion need to be met to end the search: either three pausing sites are found or *step* is equal to twice the length of the step set.

5.1.2. *Exhaustive search.* In the exhaustive search mode the steps are selected in ascending order. After the precessing of the pausing side the variable *step* is increased by one. If *step* is equal to the length of the step set the search ends.

5.2. **Pause length.** There are five different pause lengths: 5, 10, 20, 50, 100 seconds. There is a for-loop reaching over all five possibilities. Before DrTransformer is called, the equilibrium distribution of the selected site is compared to the occupancy calculated by DrTransformer. If the equilibrium is already reached, the next step is selected.

5.3. **DrTransformer call.** The call of DrTransformer is prepared by initializing the string *pause_call* that contains the pausing site, the character "=" and the length. This is processed by DrTransformer via the flag –pause-site.

5.4. **Evaluation of step.** The logfile from DrTransformer is processed by *dr_list*, and the end occupancy is searched in *liste_end*. A decision to save the step to list *pause_list* is made based on the new occupancy. The equilibrium distribution is calculated once again to check if additional time steps potentially yield new results.

5.5. **Structure alternatives.** If the structure can not be found in the DrTransformer logfile, the base pair distance of all structures present in the logfile to the structure is calculated. The structure with the lowest base pair distance is saved. The occupancy is also saved.

6. Output

A printout statement is created that contains the found pausing sites with lengths in the format that can be directly copied into the DrTransformer call. If no pausing site was found the message "no improvement found with pausing" is printed. The csv file is created. Its columns are sequence/structure/old occupancy/new occupancy/pauses/closest distance/closest structure/closest pause/closest occupancy.

7. Variables

Name	Description	Occurance
start_time	Saves the time at which the program is called. Used for measuring wall time of computations.	Main
parser	Contains arguments given via command line flags	Main
args	Parsed command line flags	Main
mode	Either "adaptive" or "exhaustive". Specified via command line flag. Determines if search is performed by adaptive walk or exhaustive search.	Main
file_name	Name of input file that contains sequence structure pairs. Must not have .txt ending. Specified via command line flag.	Main
output_name	Name of output file. Must not have .csv ending. Specified via command line flag.	Main
prune	Controls pruning parameter of DrTransformer. Specified via command line flag. See DrTransformer help for additional information.	Main
temp	Controls temperature parameter of DrTransformer. Specified via command line flag. See DrTransformer help for additional information.	Main
_RT	Boltzmann factor for calculating partition sum.	Main
f	Opened input file "file_name.txt"	Main
meta	List that contains two columns. Each row contains a sequence (column 0) and a structure (column 1)	Main
seq	Sequence from list meta that is currently used.	Main
struc	Structure from list meta that is currently used.	Main
n	Length of sequence seq that is currently used.	Main
liste	List that contains the transcriptional information form DrTransformer.	Main
liste_end	List that contains the distribution at the end from DrTransformer.	Main
new_ratio	Variable that is used to store the occupancy of structure struc in the DrTransformer output.	Main
highscore	Contains the highest observed value of new_ratio for each structure struc.	Main
old_highscore	Saves the occupancy of structure struc at transcription end without pausing site.	Main
no_alt	Contains the number of sites starting from the 5' end that have only one structure. If no_alt ==n, the program breaks. Otherwise it is used to discard steps at the 5' end that have no potential for a pausing site.	Main
clostest_dist	In case that the structure struc is not found in the DrTransformer output the clostest structure w.r.t base pair distance is saved. Variable contains current distance of closest selected structure. Initialized with n.	Main
closest_call	In case that the structure struc is not found in the DrTransformer output the clostest structure w.r.t base pair distance is saved. Variable is used for the command line print out.	Main
closest_pause	In case that the structure struc is not found in the DrTransformer output the clostest structure w.r.t base pair distance is saved. Contains the pausing site that increased the occupancy of the alternative structure.	Main
closest_ratio	In case that the structure struc is not found in the DrTransformer output the clostest structure w.r.t base pair distance is saved. Contains the occupancy after pausing of the alternative structure.	Main

closest_struc	In case that the structure struc is not found in the DrTransformer output the clostest structure w.r.t base pair distance is saved. Contains the dot bracket string of the alternative structure.	Main
closest_length	In case that the structure struc is not found in the DrTransformer output the clostest structure w.r.t base pair distance is saved. Contains the length of the pausing site of the alternative structure.	Main
pause_list	List that contains the position of found pausing sites.	Main
length_list	List that contains the length of found pausing sites.	Main
step	Variable that is used to control search. Adaptive walk: step increases every time a site is disregarded and is set to 0 if site is accepted. Search breaks when step == 2*n. Exhaustive search: step is the position of the pausing site.	Main
possible_length	List that contains all possible lengths of pausing sites. Choices are 5, 10, 20, 50, 100 seconds.	Main
step_set	List that contains all possible pausing sites. Sites at the 5' end are disregarded via variable no_alt, sites at the 3' end are disregarded by default (5 nucleotides).	Main
step_Z_list	List that contains the Boltzmann probability of all structures at a transcription step. Used to calculate the equilibrium distribution.	Main
step_occupancy_list	List that contains the occupancy of all structures at a transcription step. Used to compare distribution to calculated equilibrium distribution.	Main
Z	Partition function. Used to calculate equilibrium distribution.	Main
myp8	Contains equilibrium distribution at a transcription step.	Main
pause_len	Contains length of pause that is checked in current step.	Main
pause_call	String that specifies the --pause-site argument for DrTransformer call.	Main
f	Opened DrTransformer log-file.	dr_list
saveStep	Varibale that indicates if next line of f should be saved to output list "liste".	dr_list
saveEnd	Variable that indicates if next line of f should be saved to output list "liste_end".	dr_list
liste	List that contains the transcriptional information form DrTransformer.	dr_list
liste_end	List that contains the distribution at the end from DrTransformer.	dr_list

E-mail address: nils.gubela@univie.ac.at