BRAT – Bayesian Recombination Tracker

Manual v2.0(updated 3.11.2009)

Contents

Introduction 1
Installation 2
Input files 2
Analysing a data set 3
Interpreting the results 5
Creating summary of recombination intensity 6

Introduction

BRAT is a software tool designed for the detection of recombination events within a group of aligned multilocus DNA sequences. BRAT can be used to learn the optimal recombination model (i.e. the recombinant fragments and their origins) for a sequence and to estimate the uncertainty related to such a model by providing a marginal probability distribution over putative origins for each base in the sequence. BRAT v1.1 has been introduced in:

Marttinen, P., Baldwin, A., Hanage, W.P., Dowson, C., Mahenthiralingam, E. and Corander, J. (2008). Bayesian modeling of recombination events in bacterial populations. *BMC Bioinformatics*, 9:421, doi:10.1186/1471-2105-9-421.

This manual is intended for BRAT v2.0 which contains some upgraded features compared to the original version.

Figure 1 shows an overview of BRAT (details will be explained in later sections). BRAT is meant to be used in conjunction with BAPS program. BAPS is used to learn a clustering (i.e. an unsupervised classification) for the strains in the data set. The obtained clusters will be used in BRAT to define putative origins for the recombinant fragments of the considered strain. BAPS is freely available from:

http://web.abo.fi/fak/mnf//mate/jc/software/baps.html

BRAT can be used freely and comes with no warranty whatsoever. If you use BRAT, please cite the paper specified in the beginning of this document. Questions and feedback can be sent to

pekka.marttinen@hut.fi

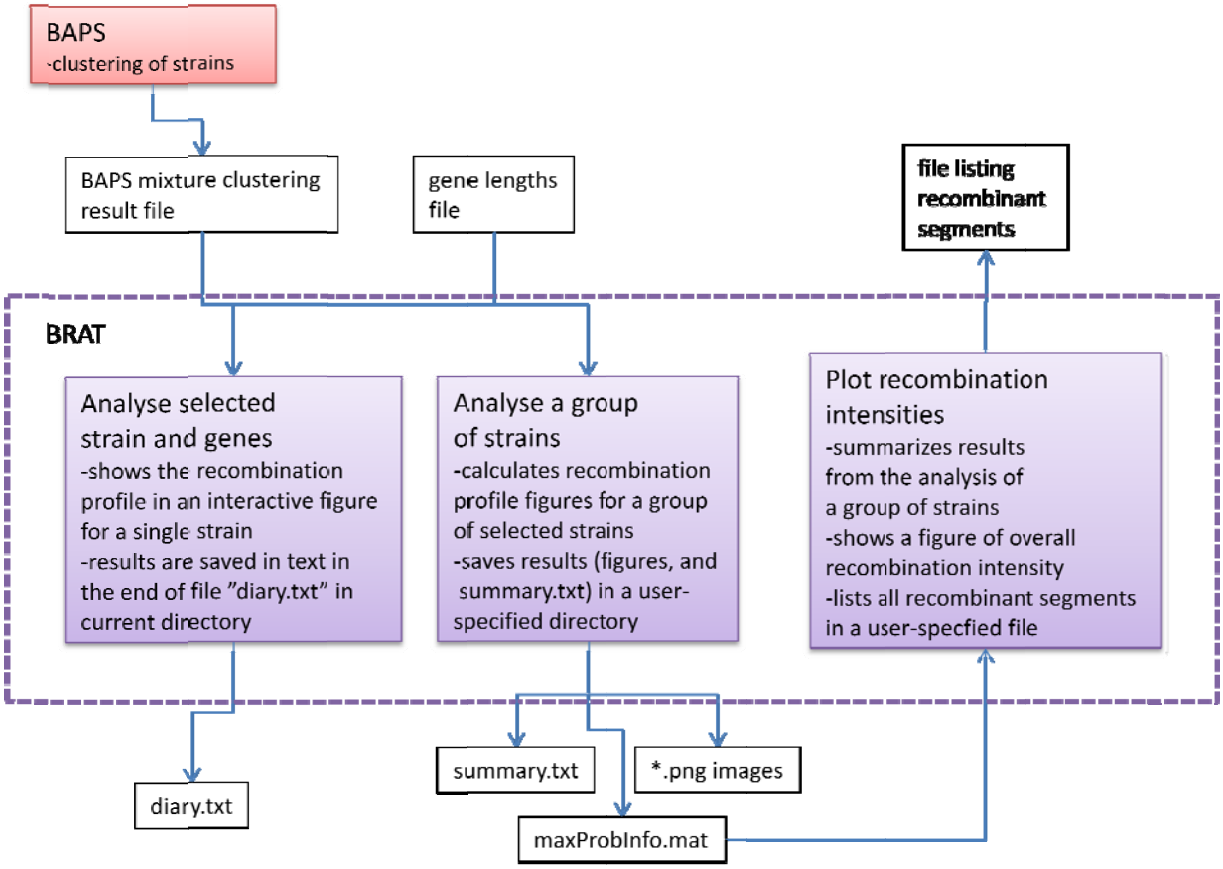

Figure 1: Overview of Brat. The components of Brat are colored in purple, input/output files are represented by white boxes.

Installation

BRAT was created with Matlab, and compiled with Matlab compiler. To run the compiled version of BRAT, you need to install Matlab Runtime Component (version 7.11, the same as for BAPS 5.3), which is freely downloadable with BAPS from http://web.abo.fi/fak/mnf//mate/jc/software/baps_xp.html.
The runtime component can be installed anywhere in the operating system, EXCEPT under the Matlab path, if Matlab is installed on your computer. After the installation of the runtime component, unzip the BRAT package in any folder, and the program is ready for use. (If you use 'Analyze a group of strains' you may need to change Display properties of your computer for a proper printing of the output figures, see the subsequent 'Analyze a group of strains' section.)

Input files

BRAT requires two input files: the mixture clustering result file obtained from BAPS and a file specifying the lengths of the different genes. It is suggested that the clustering analysis would be performed with "Clustering with linked loci" in BAPS. Also the option "Clustering of individuals" can be used. While the former is preferred, the clusterings obtained with the two models are usually very similar, if not exactly the same. For the

different data formats accepted by BAPS, see BAPS manual. The file specifying the gene lengths must be a plain text file with as many rows as there are genes in the data. Each row consists of a single number, the length of the corresponding gene. For example, if the data set comprises three genes with lengths 100, 150 and 250 bases, the gene length file would simply contain the following three rows:
100
150
250
It is also possible to include the names of the genes in another column after the lengths:
100 gene1
150 gene2
250 gene3
These names will be used in result figures. An example data is included in the BRAT package. The example consists of the following two files:
1) *test_result_baps.mat* (The clustering result file from BAPS for a data set of 60 simulated strains from three clusters, each of size 20 strains. The clustering detected by BAPS corresponds to the correct underlying clustering.)
2) *test_gene_lengths.txt* (The gene lengths for the data set.)

Analysing a data set

Once you run BRAT.exe, two windows open: one for the output and the other with two buttons, which you can use to start the analysis.

"**Analyze selected strain and genes**" performs the analysis for one strain and the genes specified by the user. After specifying the input files, the program requires the user to specify the index of the strain and genes to be analyzed. Several genes can be specified by using commas to separate them. The colon can be used to define an interval of values. For example the following input:
1:3,5,7:8
performs the analysis for the genes 1, 2, 3, 5, 7, and 8.

For text output you can specify three options 'None', 'Summary' (Prints the segments in the optimal profiles for each analyzed gene), and 'Detailed' (same as Summary, but, in addition, prints all the calculated marginal posterior probability distributions for the origins of each base).

After the results have been calculated, the program opens a new figure with the recombination profile. For the interpretation of the profile, see Section "Interpreting the results" below. Interesting areas in the profile can be investigated further by selecting an area with mouse. The first mouse click specifies the left end-point of the interesting interval, and the second mouse click specifies the right end-point of the interval. After the second mouse click, information of the selected area is written on the output window. The third mouse click resets the profile.

The text output will be written in file ‘diary.txt’ which you will find in the same directory as the program, after the analysis has been completed. One way to view the file properly is to use WordPad with option ‘No wrap’ (View->Options->Text->No wrap).

“**Analyze a group of strains**” calculates the profiles for all the genes of the strains specified by the user. Several strain indices can be specified in a similar way (using comma and colon) as several genes were specified in the analysis of a single strain. If a warning message, something like “Problems in UIW_SetUpGLPrinting” appears on the output window, the Figures will not be printed properly. To solve this, you can try changing the Color quality of your Display to 16bit (right click desktop->Properties->Settings->Color quality). (This worked for me.)

The results (figures of all the profiles and a text file summary.txt) will be saved in a directory specified by the user. Note that the figures will be ‘plain’ figures in png format, and thus do not have the same functionality as the figures obtained by analyzing individual strains. A convenient way to scan through the figures is to select View->Filmstrip in Windows Explorer in the directory in which the figures have been saved. Because the figures will be named in a similar way regardless of the data set (for example strain15.png for the fifteenth strain) it is important to select a different destination directory for the results from different data sets, so that the results would not get mixed up. Also the summary.txt file will be rewritten every time the program is executed, and the old file will be replaced. The running time for the analysis of a group of strains can be estimated straightforwardly by multiplying the processing time for one strain with the number of strains to be analyzed.

In addition to the figures of the profiles and the ‘summary.txt’ file, the analysis saves some results in binary format in file ‘maxProbInfo.mat’. This file is saved in the same directory with the other result files. The file is needed if you wish to calculate summaries of found recombinant segments over all analysed strains (see Creating summary of recombination intensity).

Interpreting the results

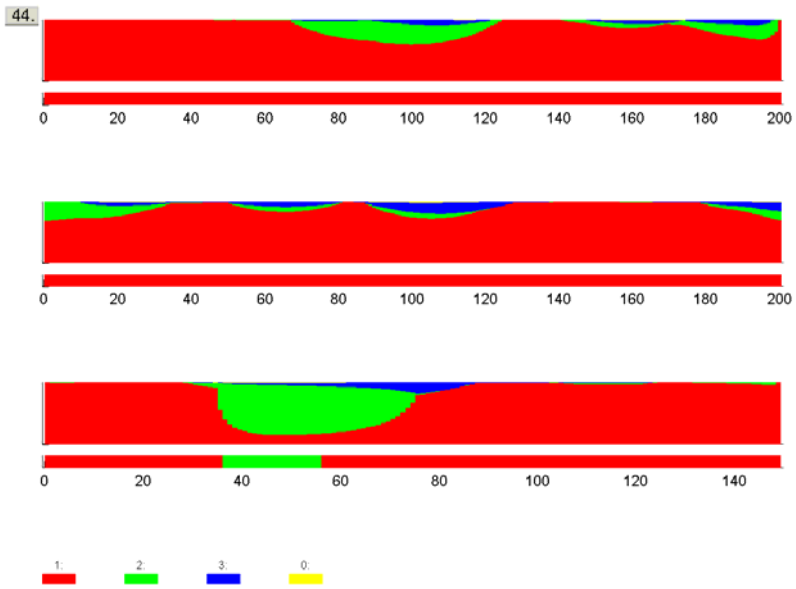

Figure 2: recombination profile for the strain #44 of the example data set

Figure 2 shows the recombination profile for the strain number 44 of the example data set included in the BRAT package. This figure can be obtained in the following way: 1) run Brat.exe, 2) select "Analyze selected strain and genes", 3) select *test_result_baps.mat* as the BAPS mixture clustering result file, 4) select *test_gene_legths.txt* as the gene length file, and 5) Specify strain index "44" and genes "1:3".

The following components can be found in Figure 2: two colored plots for each of the three genes of the strain: a narrow plot below a wider plot. The narrow plot specifies the optimal recombination model for the gene. The wider plot shows for each base the marginal posterior probability distribution over different origins, which are represented by different colors. The cluster labels corresponding to the different colors are specified at the bottom of the figure. The color corresponding to the cluster label 0 (yellow in Figure 2) refers to an unknown origin i.e. an origin, which is not represented by any of the clusters obtained in the mixture clustering analysis.

Thus, in Figure 2, the optimal recombination model assigns everything to the first cluster except for the short interval in the third gene around the 50^{th} base, which is assigned to the second cluster. Because the strain has been simulated, we know that there should not be any recombination, and the segment assigned to the second cluster should also be assigned to the first cluster. This strain was selected to serve as a cautioning example, which shows that one should not rely only on the optimal model when interpreting the results. While the optimal model assigns every base to the origin which is statistically the most plausible one, it does not say anything about how much better this origin is to some alternative. In general, the number of false recombination events detected by examining the optimal profiles is quite small (for the example data set, 6 out of 180 investigated genes have some segment in the optimal model assigned to an incorrect cluster, which

can be seen by running “Analyze a group of strains” with all the 60 strains in the data set). However, one can get rid of most of such false discoveries by examining the provided marginal posterior probability profile. For the third gene in Figure 2, the first cluster gets clearly non-zero values in the area assigned to the second cluster. Because the strain was assigned to the first cluster in the mixture clustering, the first choice for interpretation should be that all the genes of the strain originate from the population corresponding to the first cluster. Only if there is a segment of a reasonable length, say e.g. 50 bases, in some of the genes, in which the probability of the cluster to which the strain was assigned in the clustering phase is zero or very close to zero, only then one should consider an alternative origin for that part of the gene. As this is not the case with the third gene in Figure 2, we can conclude correctly that there is no conclusive statistical evidence that recombination has taken place.

Another way in which the evidence of a suggested recombination event can be evaluated is to use the mouse to select the interval of interest (For details of this, see “Analyze selected strain and genes” in the Section “Usage”). The program then outputs for the selected interval the average distances (measured as the number of differing bases) from the strain to the different clusters, and the average distances between the members of any cluster. For example, if one selects interval [37,56] in the third gene in Figure 2, the average distances to the different clusters are as follows: 1.11, 0.15, and 1.1. Thus, the fact that the optimal model favors the second cluster in this interval is caused by one point mutation, which by chance has changed the value of a base to the value common in the second cluster. For several examples on different data sets, see the paper by Marttinen et al.

Creating summary of recombination intensity

Sometimes the analysed data set may contain hundreds or thousands of strains and it is cumbersome to investigate the results by scanning through all the result figures. For this purpose, Brat has a feature which allows the user to summarize the found recombinant segments. To create the recombination intensity summary, click first the “**Plot recombination intensities**“ button. The user must specify the kinds of segments that he/she wishes to include in the summary. More specifically, the user must specify the ‘type’, minimum probability, and minimum length for a segment.

Type can be either 'liberal' or 'conservative', depending on the conditions which define the recombinant segments. If type is 'conservative', segments for which the probability of some foreign source population is higher than the given lower bound on a longer gene interval than the given minimum length, are considered as recombinant. Type 'liberal' is similar, except that the total probability of all foreign populations is compared with the given lower bound for probability, instead of the probability of the single foreign source population with the highest probability. Notice that all segments which are considered recombinant according to the ‘conservative’ definition will also be considered recombinant according to the ‘liberal’ definition, but the opposite is not true.

In addition to the conditions for recombination, also output file name and 'maxProbInfo.mat' file created by the Brat analysis of a group of strains must be specified. Brat lists all segments satisfying the given criteria to the specified output file. If one uses conditions: minimum length=10, minimum probability=0.5, type='conservative' (IMPORTANT: in practice with a real data set one should use conditions which are more strict. The values are here unreasonably low just for illustrational purposes. For example with minimum probability 0.95 and minimum length 50 no segments would be found from the example data set), there will be six segments satisfying in conditions, as shown in the corresponding output file:

```
LIST OF RECOMBINANT SEGMENTS:
                            Gene  Start  End    Origin
Strain 5 (Cluster 3):       3     86     99     0
Strain 15 (Cluster 3):      2     64     75     0
Strain 44 (Cluster 1):      3     37     70     2
Strain 45 (Cluster 1):      2     91     142    3
Strain 50 (Cluster 1):      1     186    200    0
Strain 52 (Cluster 1):      3     134    148    2
DONE
```

When the type is 'conservative' the last column specifies the foreign population which is assigned high probabilities in the segment (zero indicates an unknown source population outside of any of the sampled populations). However, if the type is 'liberal', then the last column of the table contains only zeros, because the segments are not considered to have originated from any population in particular (not even from the unknown 'outside population').

Brat also creates a graphical summary of the found segments, such that for each gene a profile of "recombinant segment intensity" i.e. a proportion of strains with recombinant dna (as specified by the user-given conditions) in a specific position along the gene is plotted. Note that the y-axis is scaled according to the maximum recombination intensity. The limits are shown in the figure.