SeedBin Manual

Yizhuang Zhou

Description

A program to automatically classify metagenomic sequences by using progressing composition and coverage (single or multiple samples), also similarity and features of single copy genes (SCGs). This program is a seed-based method.

Please Cite

If you use SeedBin in your publication, please cite:

Key features

- The state-of-the-art binning approaches combined only composition and coverage. It is the first time for us to combine four binning-useful information including composition, coverage, similarity and SCG features

- Our method was devised to select seeds from within metagenomic data without any extra effect such as primer walking. Additionally, one algorithm was devised to select small-scaffold seeds, which is useful when no large scaffold is obtained for some species. Validation shows that this algorithm is very effective.

- SCGs, which were previously used to assess binning performance, were found to carry two binning-useful features in our study and thus integrated into our method for binning. SCG features are indispensable to classify the scaffolds with indiscernible composition and coverage.

- Most previous approaches used static coverage and composition. In contrast, our method used progressing composition and coverage.
- Our method is a reference-independent algorithm without using phylogenetic affiliations for binning. Similarity is only applied to identify SCGs.

- Small scaffolds are binned by Naïve Bayesian classifier.

Installation

SeedBin is developed with Perl script, so there is no necessary to compile SeedBin. The file named cutoff.xls must be localized at the same directory of SeedBin.pl. it contains the cutoffs for seeding and merging. Also some Perl packages are needed to be installed in order to run SeedBin.pl. They can downloaded from http://www.cpan.org/. Please note that Statistics packages are included in the same

directory of SeedBin.pl. Packages include:

Getopt::Long
File::Basename
FindBin
lib
Statistics::Distributions
Cwd
File::Path
threads
threads::shared
POSIX

Usage

perl SeedBin.pl [options]* -fafile <file> -depthfile <file> -tablefile <file>
The following arguments must be provided:

-fafile <s>: fasta file containing scaffolds assembled from metagenomic reads
-depthfile <s>: the file containing depths for scaffolds
-tablefile <s>: the file containing SCG type and their Scaffolds

The following arguments are optional:

-seedfile <s>: the file containing seeds
-nb_process <i>: the number of threads to be used, the default is 20
-fold <f>: the fold for computing depth range, the default is 2.58
-minlen <i>: the minimal length of scaffold, not including Ns, the default is 0
-outdir <s>: the output directory
-LargeLen <i>: the minimal length of large scaffold derived seeds, the default is 200,000
-***SmallLen*** <i>: the minimal length of small scaffold derived seeds, the default is 100,000
-BinLen <i>: the minimal length of bins, the default is 500,000
-scgNum <i>: the minimal number of SCGs which can be considered as an bin, the default is 7
-scgNum4bin <i>: the minimal number of SCGs which can be considered as an final bin after merging, the default is 10
-help: show the help message

In the DNA sequences file (***-fafile***), the sequence should be in FASTA format. The data format is as follows:

```
>gi|494587256|gb|APMI01000001.1| Wastewater metagenome HPminus1.1, whole genome shotgun sequence
ATTTTTACCGTTGCATAGTAATTGCCCGAAGCATCTTTGCACTCAATAAAGTTGAAATTTACTGTCTTGC
TACTGGTTGGTTTATCGGTAACGATAGTGGCGTAATTACCCGATATGCCTTTATACAGCATAATATCACC
CGGTGCATAATCAAGTGAGATGCTTGCCTTGGTACTCTTAAGCGGGGCTACATATTTTGTTTCTTCAACA
CCCTGGAAAGAAATATCGGCCTTGCCAGTGTTATCAGAAACGCTGATAAGCTTTCCGGACACAGCGTATC
CGGCTCCCTGAACGCTGATAACATATGCCCCCGTGCGTAAGGAAACCTTGAAAGTATTGGTTCCTTTTTG
CAGATTTTGTGCATAAGCGGCAACCCTTCTGCCATCCATACTGTAAACATTGATGCTGGCATTGCCAGCT
TTTGCTGCATAAAATGAAACGTTTGAGCTTCCTGCAACTGAACTGGTAATACGCAAACCTTCATTCGCGG
```

All coverage across multiple samples are included in one file (***-depthfile***), separated by "#". Each line represents one scaffold, starting with the scaffold ID. Then for coverage from each sample, start with mean coverage across whole scaffolds and coverage values with window 500 bp, sliding 250 bp.

The table file (***-tablefile***) which generally named *SCGtype_SCGscaffoldDepth.xls* in our example data, provides SCG information. Each line represents information of one SCG type. The first column in this file is the HMM ID for SCG collected from TIGR or Pfam. The scaffolds which are annotated to carry the SCG type are ordered by their coverage of certain sample usually deep sequenced. The values in bracket are coverage and length for its scaffold, separeted by ",". All are separated by tab. Usually, The species with the highest abundance can be readily detected and thus selected as seeds for this species supplied by ***-seedfile***. NA in table means there is no SCG scaffold for this SCG type.

-seedfile argument is optional. When no known origins of some scaffolds, no seeds can be inputted. In this condition, we cannot provide seeds through ***-seedfile***. Each line presents one seed for one species. The seed may be one scaffold or scaffold collection concatenated by ";".

-nb_process set the number of threads. When classifying small scaffolds by Naïve Bayesian classifier. It take lots of time. So in this software, we parallel multiple threads to perform classification of small scaffolds.

-fold specifies the fold to calculate coverage intervals. Coverage intervals are approximated by abnormal distribution. The default for ***-fold*** is 2.58, which may expectedly cover 99% of coverage. You can set it as 1.99 (95%) or others.

-minlen specifies the sequence length threshold, scaffolds shorter than this value will not be included. The default is 0, suitable for simple community less than 10 species. Otherwise, we usually set it as 1000.

-outdir specifies the output directory, where all output files are localized at this directory. The default is the current working directory.

-LargeLen specifies the sequence length threshold of large-scaffold seeds. After classification, if the total length of assigned scaffolds of a large-scaffold seed is shorter than this value, this bin will be discarded before merging bins. Discarding small bins is to filter unreal bins. Whether a bin of large scaffold seed is to be discarded or remained is determined by ***-LargeLen*** **and** ***-scgNum4bin***.

-SmallLen specifies the sequence length threshold of small-scaffold seeds. After classification, if the total length of assigned scaffolds of a small-scaffold seed is shorter than this value, this bin will be discarded before merging bins. Discarding small bins

is to filter unreal bins. Although the default of ***-SmallLen*** is different from the default of ***-LargeLen***, our experience shows that setting ***-SmallLen*** as 200,000 is even better. Whether a bin of small scaffold seed is to be discarded or remained is determined by by ***-SmallLen*** **and** ***-scgNum4bin***.

-BinLen specifies the sequence length threshold of bins after merging. Our statistic analysis for all published genomes from NCBI shows that most species carry at least 1 Mb genomes. So we consider that a species short than half of the value, we discard this bin possibly generated by unreal seeds. Whether a bin after merging process is to be discarded or remained is determined by ***-BinLen*** **and** ***-scgNum4bin***.

-scgNum the minimal number of SCGs which can be considered as an bin, the default is 7, considering the complexity nature of metagenome compared to genome. This argument is useful for classification splitting.

-scgNum4bin specifies the minimal number of SCGs indicating that this bin is a real bin. In some condition, a rare species are not deep sequenced and thus not well assembled. Although this total length is short than the length threshold specified by ***-scgNum4bin***, it also a independent bin for a species. In this context, the bin with more than number of SCG specified by ***-scgNum4bin*** must be remained.

Coverage calculation

After assembly we map the reads of each sample back to the assembly using soap (http://soap.genomics.org.cn/). Of course, you can use other mapping software such as bowtie2. Then you can use soap.coverage (http://soap.genomics.org.cn/) to calculate base coverage. Finally, use Scripts/depth_mean_rawdepth.pl to calculate coverage file for ***-depthfile***. There is one example in Scripts/example/. The first parameter for Scripts/depth_mean_rawdepth.pl is a file with format that [file path] and [trimmed length] separated by tab. Please note the trimmed length for both end is very similar to the read length of Illimula read.

Generation of SCG table

Use Scripts/SCGtype_SCGscaffoldDepth.pl to generate SCG table for ***-tablefile***. To do this, we also need to install two extra softwares:

1) FragGeneScan, software for gene predication

It is available from http://omics.informatics.indiana.edu/FragGeneScan/.

2) hmmsearch, software for SCG predication

It is available from http://hmmer.janelia.org/.

And two files contains HMMs are singleCopygene_fromTIGR.HMM and singleCopygene_fromPfam.HMM. singleCopygene_samegene.xls is used to merge some SCGs. All three files are localized at directory Scripts/.

SeedBin output

All output files are in directory defined by ***-outdir***. There are 13 output files including:

1) largeScaffoldSeed_CovInterval.xls

This file contains large-scaffold seeds and their coverage intervals. Each line represents one large-scaffold seeds, starting with seeds, followed with minimal and maximal coverage of the coverage interval for one sample.

2) allscaffoldSeed_CovInterval.xls

After selecting small-scaffold seeds, all seeds including small- or large-scaffod seeds are recorded in this file. Also their coverage intervals are also included in this file. The format is similar to largeScaffoldSeed_CovInterval.xls.

3) allscaffold_RawClassified.xls

All scaffolds are classified by all seeds including small- or large-scaffold seeds and their assignments are recorded in this file. Each line represent one scaffold, starting with scaffold ID and their assigned bin named by its seeds.

4) SCGscaffold_RawClassified.xls

This file contains SCG information after classification of scaffolds by all seeds including small- or large-scaffold seeds. The format is similar to the input file specified by the argument ***-tablefile***. NA in table means there is no SCG scaffold for this SCG type.

5) allscaffoldSeed_ClassifiedLength_AfterSplit.xls

This file contains total length of assigned scaffolds for bins after splitting. You can check the total length of all bins, and reset ***-LargeLen*** and ***-SmallLen*** according to the information in this file. Each line represents assignments of all seeds, including seed, total length , the number of assigned scaffold and the assigned scaffold IDs (concatenated by ";") which are separated by tab.

6) SCGscaffold_AfterSplit.xls

This file contains SCG information after splitting. The format is similar to the input file specified by the argument ***-tablefile***. NA in table means there is no SCG scaffold for this SCG type.

7) SCGscaffold_AfterMerge_FirstTime.xls

This file contains SCG information after merging at the first time. The format is similar to the input file specified by the argument ***-tablefile***. NA in table means there is no SCG scaffold for this SCG type.

8) SCGscaffold_AfterMerge_SecondTime.xls

This file contains SCG information after final merging. The format is similar to the

input file specified by the argument ***-tablefile***. NA in table means there is no SCG scaffold for this SCG type.

9) allscaffoldSeed_ClassifiedLength_Final.xls

This file contains total length of assigned scaffolds for bins after final merging. You can trace the total length of all bins. Each line represents assignments of all seeds, including seed and total length which are separated by tab.

10) SCGscaffold_FinalClassified.xls

This file contains SCG information after reclassification of all SCG scaffolds. The format is similar to the input file specified by the argument ***-tablefile***. NA in table means there is no SCG scaffold for this SCG type.

11) FinalSeed_CovInterval.xls

After reclassification of all SCG scaffolds, all seeds including small- or large-scaffod seeds are recorded in this file. Also their coverage intervals are also included in this file. The format is similar to largeScaffoldSeed_CovInterval.xls.

12) classified_allscaffold.xls

The final classification of all scaffolds. Each line represents one scaffolds, starting with scaffold ID and then following the indexed bin number, scaffold length and mean coverage for each samples and finally the seed of its bin. They are separated by tab.

13) SeedBin.log

This file contains log information.

Performance evaluation

In this documentation, performances in terms of overall performance index (OPI) and individual performance index (IPI), together with precision, sensitivity and so on are evaluated by this determined origins. You can use performance_evaluation.pl, NBT_IPI.pl, and F1RT_performance_evaluation.pl to do it. All are localized at directory Scripts/.

Otherwise, If you want to evaluate performance with SCG. You can write a simple scripts to do it by using SCGscaffold_FinalClassified.xls.

Other scripts

There are some useful scripts which to repeat our study. All these scripts are localized at directory Scripts. They are included at least the following scripts.

1) genomealign_sortbyr.pl for performing alignments

2) filter_genomealign.pl and coords_filter.pl for alignment filtering

Examples

Six data set can also downloaded and their detailed analysis wrote in README of its directory are also available (https://github.com/Yizhuangzhou/SeedBin/) .

Support

If you are having issues, please email me via zhouyizhuang3@163.com .