FunctionSIM

FunctionSIM is a bioinformatics tool to generate microbial DNA sequences. With the help of GemSIM, a public domain next-generation sequencing simulator, it lets the user generate metagenomic read data set based on the functional roles of the genes. The FunctionSIM consists of two components: a GUI component and the SEED database.

Data

The FunctionSIM software uses the data from SEED database. The SEED database consists of annotated genomic data and is organized as a set of subsystems. The SEED data follows a hierarchical structure. The first level is the category and then subcategory, subsystem, and functional role. A subsystem consist of one or more functional roles. It is to be noted that a functional role may belong to multiple subsystems. Features or genes from several organism may have one or more functional role. The feature in SEED database is identified using a FIG Id. The FIG Id is of the form 'fig|<genome_id>.<locus_type>.<id_number>'.

The FunctionSIM bundles two database files, namely, a mapping file that contains the hierarchical data and a multifasta file that consists of the actual microbial DNA sequence data. Although the database is stored as a plain text file, there arises no reason for the user to directly deal with the database files; the database is purely intended to be read by the FunctionSIM software. The SEED data that has been bundled with the FunctionSIM was downloaded from 'PubSEED' using SEED API and SEED server scripts. More details of the SEED Project can be found on their website.

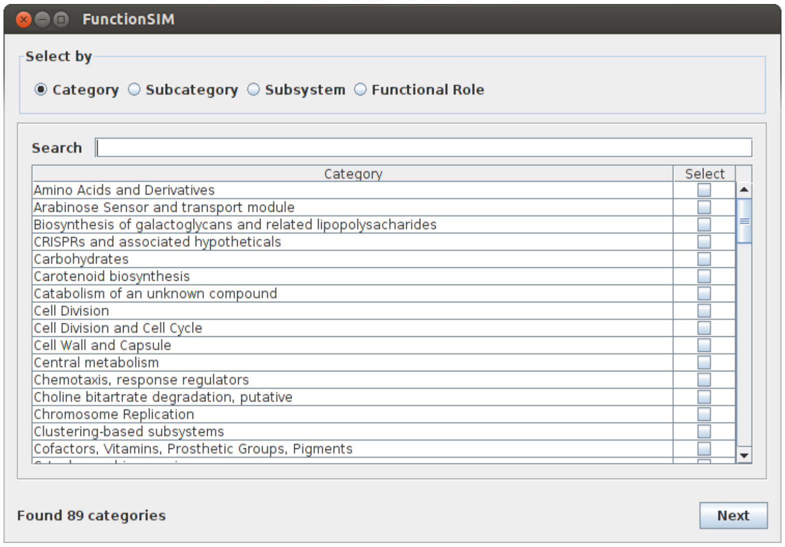
Software

FunctionSIM software has been written in Java, with GUI implemented using Swing API. The user is guided with a wizard like user interface to generate the required read data set. The software consists of mainly three screens:

- Data selection
- Review and Edit Abundance
- Simulator Parameters

Data Selection Screen

In the first step, the user begins by selecting the required functional roles in the final read data set. A functional role is basically the task performed by a feature in its host organism. A functional role may consist of several features from different species. The user selects a functional role by checking the corresponding checkboxes. The user can also select the functional roles based on its subsystem, subcategory or category. Selecting a higher hierarchy level will select all the functional roles contained within it. Once the selection has been made, the user moves onto the next screen by clicking on the 'Next' button on the bottom of the window.



Select by

● Category ○ Subcategory ○ Subsystem ○ Functional Role

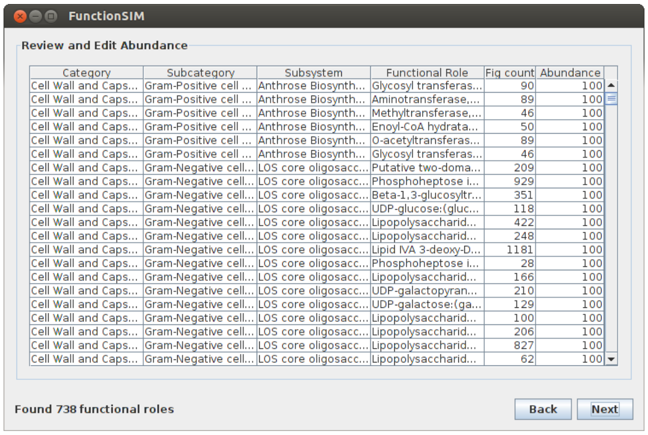
Search

Category	Select
Amino Acids and Derivatives	□
Arabinose Sensor and transport module	□
Biosynthesis of galactoglycans and related lipopolysacharides	□
CRISPRs and associated hypotheticals	□
Carbohydrates	□
Carotenoid biosynthesis	□
Catabolism of an unknown compound	□
Cell Division	□
Cell Division and Cell Cycle	□
Cell Wall and Capsule	□
Central metabolism	□
Chemotaxis, response regulators	□
Choline bitartrate degradation, putative	□
Chromosome Replication	□
Clustering-based subsystems	□
Cofactors, Vitamins, Prosthetic Groups, Pigments	□

Found 89 categories

Review and Edit Abundance Screen

In this screen, user can review their selection made in the previous screen. This screen lists out all the functional roles that have been selected, along with their parent hierarchy levels. It also shows how many features are present for each selected functional role. Each feature corresponds to one DNA sequence in the database. The 'Abundance' column specifies the relative abundance of features of that functional role. The default abundance value is 100, but that can be changed by the user by clicking on that field and entering a new value.



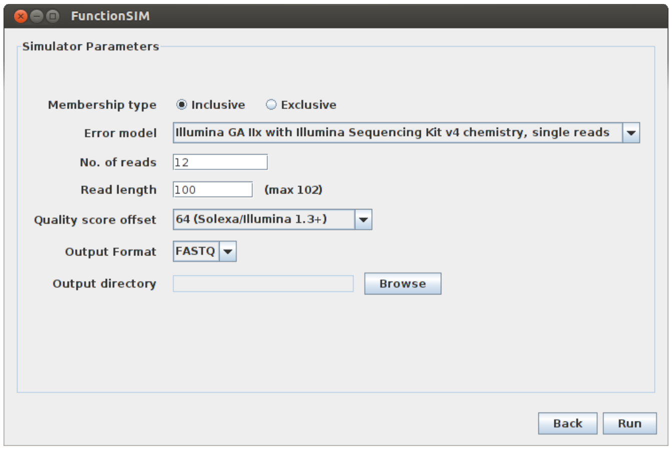
Review and Edit Abundance

Category	Subcategory	Subsystem	Functional Role	Fig count	Abundance
Cell Wall and Caps...	Gram-Positive cell ...	Anthrose Biosynth...	Glycosyl transferas...	90	100
Cell Wall and Caps...	Gram-Positive cell ...	Anthrose Biosynth...	Aminotransferase,...	89	100
Cell Wall and Caps...	Gram-Positive cell ...	Anthrose Biosynth...	Methyltransferase,...	46	100
Cell Wall and Caps...	Gram-Positive cell ...	Anthrose Biosynth...	Enoyl-CoA hydrata...	50	100
Cell Wall and Caps...	Gram-Positive cell ...	Anthrose Biosynth...	O-acetyltransferas...	89	100
Cell Wall and Caps...	Gram-Positive cell ...	Anthrose Biosynth...	Glycosyl transferas...	46	100
Cell Wall and Caps...	Gram-Negative cell...	LOS core oligosacc...	Putative two-doma...	209	100
Cell Wall and Caps...	Gram-Negative cell...	LOS core oligosacc...	Phosphoheptose i...	929	100
Cell Wall and Caps...	Gram-Negative cell...	LOS core oligosacc...	Beta-1,3-glucosyltr...	351	100
Cell Wall and Caps...	Gram-Negative cell...	LOS core oligosacc...	UDP-glucose:(gluc...	118	100
Cell Wall and Caps...	Gram-Negative cell...	LOS core oligosacc...	Lipopolysaccharid...	422	100
Cell Wall and Caps...	Gram-Negative cell...	LOS core oligosacc...	Lipopolysaccharid...	248	100
Cell Wall and Caps...	Gram-Negative cell...	LOS core oligosacc...	Lipid IVA 3-deoxy-D...	1181	100
Cell Wall and Caps...	Gram-Negative cell...	LOS core oligosacc...	Phosphoheptose i...	28	100
Cell Wall and Caps...	Gram-Negative cell...	LOS core oligosacc...	Lipopolysaccharid...	166	100
Cell Wall and Caps...	Gram-Negative cell...	LOS core oligosacc...	UDP-galactopyran...	210	100
Cell Wall and Caps...	Gram-Negative cell...	LOS core oligosacc...	UDP-galactose:(ga...	129	100
Cell Wall and Caps...	Gram-Negative cell...	LOS core oligosacc...	Lipopolysaccharid...	100	100
Cell Wall and Caps...	Gram-Negative cell...	LOS core oligosacc...	Lipopolysaccharid...	206	100
Cell Wall and Caps...	Gram-Negative cell...	LOS core oligosacc...	Lipopolysaccharid...	827	100
Cell Wall and Caps...	Gram-Negative cell...	LOS core oligosacc...	Lipopolysaccharid...	62	100

Found 738 functional roles

Simulator Parameters Screen

In this final screen, user can specify the simulator parameters such as read length, no. of reads, error model etc that are stipulated by the GemSIM simulator. An important configuration parameter is the membership type. It determines how the features implementing multiple functional roles are selected. There are two types of membership type: inclusive and exclusive. To explain the concept, consider three hypothetical functional roles: A, B and C. Assume the user has selected functional role A and B in the first data selection screen. In the inclusive mode, all the features that belong to functional role A and/or B will be chosen, even if the feature have an additional functional role C. In the exclusive mode, only those features that has functional role A and/or B will be chosen; if a feature has functional role C along with functional role A and/or B, it will be discarded.



Membership type ● Inclusive ○ Exclusive

Error model: Illumina GA IIx with Illumina Sequencing Kit v4 chemistry, single reads

No. of reads: 12

Read length: 100 (max 102)

Quality score offset: 64 (Solexa/Illumina 1.3+)

Output Format: FASTQ

Output directory:

Once the required parameters are set, the user can now run the simulator by pressing the ‘Run’ button. At this time, the FunctionSIM will read through the SEED database and extract all the sequences that will be provided as the input for the GemSIM simulator. The extracted sequences will be placed under a directory within the output directory specified. Once the sequences have been extracted, the simulator will be invoked with the specified abundance values and the sequences to generate the final read data set. The final output is a FASTA or FASTQ file (depending on the user selection) and will be placed in a specified output directory.

Note: If you experience an error like “GemSIM/GemReads.py:444: RuntimeWarning: invalid value encountered in divide Mprobs=mx[d0][d1][d2][d3][d4][d5]/tot” it is due to a bug in GemSIM simulator. You may select different functional roles.

References

[1] http://www.theseed.org/wiki/Home_of_the_SEED

[2] http://www.nmpdr.org/FIG/wiki/view.cgi/FIG/Subsystem

[3] McElroy, Kerensa E., Fabio Luciani, and Torsten Thomas. "GemSIM: general, error-model based simulator of next-generation sequencing data." *BMC genomics* 13.1 (2012): 74.

[4] http://pubseed.theseed.org/

[5] http://pubseed.theseed.org/sapling/server.cgi?pod=SAP

[6] http://pubseed.theseed.org/sapling/server.cgi?pod=ServerScripts