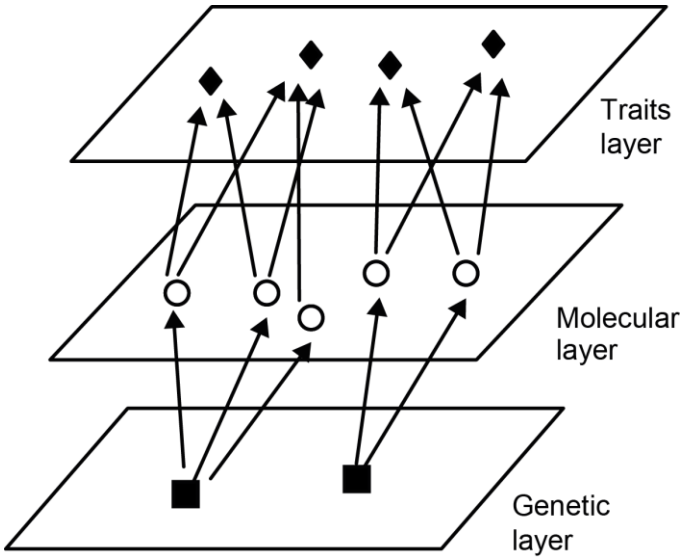
GEMOT algorithm
Quick start manual



GEnetic & MOlecular mechanisms of Trait

Yael Oren, Aharon Nachshon, Amit Frishberg, Roni Wilentzik, Irit Gat-Viks

Quick start

The following manual describes how to use the GEMOT algorithm.

GEMOT's executable JAR and source code can be downloaded from http://csgi2.tau.ac.il/gemot/ using different browsers including FireFox version 22 (22.0 and higher), Chrome version 28 (28.0.1500.72 and higher), Safari version 6 (6.0.2 and higher) and Explorer version 10 (10.0.9200 and higher).

General Overview:

The GEMOT algorithm takes as input molecular data (e.g. gene expression data), traits (phenotypes) data, and genotyping data, and suggests groups of related phenotypes as well as the genetic and molecular mechanisms behind them. These are provided in the form of *modules*, each containing a set of traits, a genomic interval, and a set of molecular components suggested as the intermediate layer between the genomic interval and the traits.

GEMOT takes the following inputs:

1. A molecular data (e.g., gene expression) file.
2. A trait measurements file.
3. A genotyping file.
4. An index file for molecular components (e.g., genes).
5. An index file for traits.
6. An index file for genetic variants.
7. A run parameters file.
8. Association scores file (optional for inbred strains).

All these input files should be uploaded by the user. Example files are available at: http://csgi2.tau.ac.il/gemot/SampleDataFiles (see details below) and can be downloaded through the http://csgi2.tau.ac.il/gemot/help.html. Detailed explanations of the input files structure is provided below.

GEMOT provides the following outputs:

1. A module summary file.
2. Modules' traits file.
3. Modules' genes file.

The output files names are defined in the run parameters file (input #7). Additional temporary files are created for the internal use of the GEMOT algorithm.

Running GEMOT

To run the GEMOT algorithm, proceed as follows:

1. Prepare all input files and update the run-parameters file accordingly. The names for the output file names should also be specified in the run-parameters file. The run-parameters file should be called run-parameters.txt

2. Put all the input files, including the run-parameters file in a certain directory (e.g. "my_dir").

3. Use the following three commands sequentially:

```
> java -jar -Xmx4096m GEMOT.jar 0 my_dir > log.txt
```

Input files - details

For simplicity, we refer to gene expression as the molecular data and genes as the molecular components throughout the following description. However, any other high-throughput data and any other molecular component can be used instead.

Input files build on several main identifiers: the unique *gene identifier* (any string can be used); the *genetic background identifier* (any string can be used); the *trait identifier* (sequential integer numbers, starting at 0); and the *genetic variant identifier* (sequential numbers, starting at 0).

1. A molecular measurements (here, gene expression) file

A tab-delimited full matrix, where each entry is the gene expression data of a certain gene (specified in rows) in a certain individual (specified in columns).

- The first column should be a unique *identifier of a gene* (the same number of genes in the same order as in file #4)
- The first row should be a *genetic background identifier* of an individual (the same number of genetic backgrounds in the same order as in file #3).
- Each entry provides the log gene expression data, normalized for mean 0 and standard deviation 1. Missing data should be marked by the constant -999.

Sample file: http://csgi2.tau.ac.il/gemot/SampleDataFiles/molecular-measurements.tab

2. A trait measurements file

A tab-delimited full matrix, where each entry is the trait measurements data of a certain trait (specified in columns) for a certain individual (specified in rows).

- The first column should be the *trait identifier* (sequential numbers, starting at 0; in the same number and order as in file #5).
- The first row is the *genetic background identifier* for an individual (the same number of genetic backgrounds in the same order as in file #3).
- Each entry should provide the trait measurement, normalized for mean 0 and standard deviation 1. Missing data should be marked by the constant -999.
- Sample file: http://csgi2.tau.ac.il/gemot/SampleDataFiles/trait-measurements.tab

3. The genotyping file

A tab-delimited full matrix, where each entry is the genotyping data of a certain individual (specified in columns) for a certain genetic variant (specified in rows).

- The first columns should be a sequential *genetic variant identifier* (sequential numbers, starting at 0; in the same number and order as in file #6)
- The first row should be the *genetic background identifier* of an individual. In the case of outbred population, the identifier is unique since each individual attains a different genetic background. For inbred population, the data may include several individuals of the same strain; in such case all individuals of the same strain should have the same genetic background identifier and the same genotyping data.
- The GEMOT algorithm assumes only two possible alleles in each variant. In accordance, the genotyping data is either 0 or 1 (for homozygous cases), or -1 for the heterozygous case. Sample file: http://csgi2.tau.ac.il/gemot/SampleDataFiles/genotyping.tab

4. An index file for molecular components (here genes).

A tab-delimited full matrix, where each row is a gene and each column provides additional information about the genes, such as their genomic position and entrez ID.

- The columns should contain:
    - Column 1: A unique *gene identifier*. The gene identifier can be any string.
    - Column 2: The gene's chromosome (an integer)
    - Column 3: Start genomic position of the gene
    - Column 4: End genomic position of the gene
- The first row should contain the titles for the additional information provided.
- Any information can be added in column 5 and on.
- Missing data entries should include a non-empty string (e.g., a single space).
- Sample file: http://csgi2.tau.ac.il/gemot/SampleDataFiles/info-molecular-measurements.tab

5. An index file for the traits

A tab-delimited full matrix, where each row is a trait and each column provides additional information about the trait, such as their description and reference to the original study (*e.g.*, a PubMed ID).

- The first column should include the *trait identifiers*. Trait identifiers are sequential and start at zero. Any information can be added in column 2 and on.
- The first row should contain the titles for each column.
- Missing data entries should include a non-empty string (e.g., a single space).
- Sample file: http://csgi2.tau.ac.il/gemot/SampleDataFiles/info-trait-measurements.tab

6. An index file for genetic variants

A tab-delimited full matrix, where each row is a genetic variant and each column provides additional information about the variant, such as genomic positions and title.

- The columns should contain:
    - Column 1: the genetic variant identifiers. The identifiers should be sequential integer numbers, starting at 0.
    - Column 2: The variant's chromosome (an integer)
    - Column 3: genomic position of the variant
- The first row should contain the titles for the additional information provided.
- Missing data entries should include a non-empty string (e.g., a single space).
- Additional information can be added in column 4 and on.
- Sample file: http://csgi2.tau.ac.il/gemot/SampleDataFiles/info-genotyping.tab

7. A run parameters file

A tab-delimited full matrix:

- Each row contains the name of a run-time parameter (column 1) and its value (column 2).
- The names of the run-time parameters are case sensitive.
- Sample file: http://csgi2.tau.ac.il/gemot/SampleDataFiles/run-parameters.txt

The following table provides detailed explanations about the run-time parameters.

Parameter name	Description
MOLECULAR_MEASUREMENTS_FILE_NAME	The name of the molecular data input file #1 (e.g., gene expression data).
NUMBER_OF_MOLECULAR_MEASUREMENTS	The number of data rows in the molecular data file #1. Should be the exact number, excluding the title line.
TRAITS_FILE_NAME	The name of traits data input file #2.
NUMBER_OF_TRAITS	The number of data rows in the traits data file #2. Should be the exact number, excluding the title line.
VARIANTS_FILE_NAME	The name of genotype input file #3.
NUMBER_OF_VARIANTS	The number of data rows in the genotype input file #3. Should be the exact number, excluding the title line.
MOLECULES_INFO_FILE_NAME	The name of the index file for the molecular components (input file #4).
MAX_MOLECULES_INFO_COLUMNS	The exact number of data columns in the index file #4, not including the ID column.
TRAITS_INFO_FILE_NAME	The name of index file for the traits (input file #5).
MAX_TRAITS_INFO_COLUMNS	The exact number of data columns in the index file #5, excluding the ID column.
VARIANTS_INFO_FILE_NAME	The name of index file for the genetic variants (input file #6).
MAX_VARIANTS_INFO_COLUMNS	The exact number of data columns in the index input file #6, excluding the ID column.
ASSOCIATION_INPUT	TRUE / FALSE - is the association score input file #8 given as input.
ASSCIATION_FILE_NAME	The name of the input association file #8. This parameter is optional, should be included only if ASSOCIATION_INPUT=TRUE.
NUMBER_OF_INDIVIDUALS	The exact number of genetic backgrounds (should match the number of unique 'genetic background identifiers' in all input files).
INITIATION_CUTOFF	The initialization cutoff of the REL software package.
EXTENSION_CUTOFF	The improvement cutoff of the REL software package.
LINK_POTENTIAL	The link potential cutoff for including a gene in the module.
MAX_NUMBER_OF_MODULES	The maximal number of output modules.
MODULES_SUMMARY_TABLE_FILE_NAME	The name of the modules summary output file #1.
TRAITS_SUMMARY_TABLE_FILE_NAME	The name of the traits summary output file #2.
MOLECULES_SUMMARY_TABLE_FILE_NAME	The name of the genes summary output file #3.

8. Association scores file (optional for inbred strains).

A tab-delimited full matrix, including the association score for each molecular component (gene) and each variant. For the case of outbred strains, this file is mandatory. For the case of inbred strains, if this file is omitted, GEMOT using a t-test to calculate the association scores.

- The first column should be a *genetic variant identifier* (exactly the same number and order of variants as appears in file #6).
- The first row should be the *gene identifier* (with the same number and order of genes as in input file #4).
- Each cell contains the assigned association score P-value for the variant (row) and the molecular component (such as expressed gene, column).
- Sample file: http://csgi2.tau.ac.il/gemot/SampleDataFiles/association.tab

Output files - details

1. A module summary file

A tab delimited text file.

- This file contains one row for each of the predicted GeMOT modules. Each row provides the following information:
    - Module number (column 1)
    - Module score(column 2)
    - The genomic position of the module's genomic interval (column 3)
    - The number of traits (phenotypes) in the module (column 4)
    - The number of genes in the module (column 5)
- Sample file: http://csgi2.tau.ac.il/gemot/SampleDataFiles/GEMOT-summary-modules.tab

2. A modules' traits file

A tab delimited text file.

- The file contains one row for each of the traits in the final GEMOT modules. Each row provides the following information:
    - Module number (column 1)
    - Trait identifier (column 2; as specified in the traits-measurements input file #5).
    - All of the trait's additional information, as given in the traits-measurements input file #5 (column 3 and on).
- Sample file: http://csgi2.tau.ac.il/gemot/SampleDataFiles/GEMOT-summary-phenotypes.tab

3. A modules' genes file

A tab delimited text file.

- The file contains one row for each of the gene in the final GEMOT modules. Each row provides the following information:
    - Module number (column 1)
    - CIS/TRANS association (column 2)
    - Gene identifier (column 3; as specified in the molecular-measurements input file #4).
    - Gene position (column 4)
    - All of the gene's additional information, as given in input file #4 (columns 5 and on).
- Sample file: http://csgi2.tau.ac.il/gemot/SampleDataFiles/GEMOT-summary-genes.tab