# Anet v1.0: software for generation of association networks

Byung H. Park and Tatiana V. Karpinets
April, 2012

This is a brief description of two programs (anet and t2t.pl) used for generation of association networks.

*anet* (abbreviated from Association NETworks) is a C++ program that extracts associations between annotaions in the transaction-like records used as the input. The result can be used as a network of associations for a cluster analysis or for visualization using Cytoscape or other tools.

*t2t.pl* (abbreviated from Table TO Transactions), given a tabular data with fixed rows and columns, checks whether the data is in the correct format and then produces a transaction-like record data. The output file of *t2t.pl* is of the suitable format for both conventional association rule learning program such as *apriori* and *anet* (association network generation).

## Installation

Copy the file anet.tar.gz into a directory of your choice and run the following command

*tar xzf anet.tar.gz*

This will create directory *anet* with subdirectories *src* and *bin*.
Run the following commands to generate the *anet*.

*cd src*
*make*

This will create the executable *anet* under *bin* directory.
Copy it into a desired directory.

*t2t.pl*:
This is placed under "bin" directory. Just copy it into a desired directory.

## Converting a table to the type-value formatted transaction-like table using *t2t.pl*

### Input Format
**==========**
The Perl program *t2t.pl* transforms data examines whether the input file conforms to the requirements and convert it into type-value formatted transaction records. The requirements for the input file are as follows.

1. The table is of tab delimited format. Also the number of columns is larger than two.
2. The first row of the data is the header, and the first column is titled "ID".
3. “ID” column is unique for each row, but the rest of the columns need not to be.
4. The rest columns of the header are classification types, e.g., controlled vocabulary that denotes biological, physiological and other traits of the object.
5. Colon “:” is not allowed anywhere in the table.
6. No duplicate columns are allowed in the table.

### *t2t.p*l Usage
**==========**
***perl t2t.pl -f <input file> -o <output file>***

**- input file**
input file name (or full path): a table with annotations
The file format must meet the requirements as aforementioned.

**- output file**
output file name (or full path): transaction records with annotations augmented by their types (column names)
The output file of the program can be used as an input file for the anet program.

**Example**
**=======**

*perl t2t.pl -f GB.data -o GB.data.tr*

Input file GB.data is a sample table included in this package. It is a genome annotation data and of tab-delimited format: one row of annotations per genome. Output file GB.data.tr is type-value formatted transaction-like records.

## Generating association network using anet

***anet* usage**
**=========**

*anet --file=<input file> --method=<correlation type> --output=<output file> --by=<filtering method> --count=<filtering threshold> --threshold=<output threshold>*

- *input file:* file name (or full path) of a transaction-like records; each record is a list of annotations separated by blank space. An output file of *t2t.pl* can be used as the input.

*-- correlation type*
spearman (default), cosine, or pearson

*- output file*
output file name (or full path)

*- filtering method*
*by_cooccur_count or by_item_count*

This option is used to filter out insignificant annotations that are rarely found in records (*by_item_count*) or co-occur with very few annotations (*by_cooccur_count*). For each case, the threshold is set by another option, *--count* (see below).

*- count*
This option is used to set the threshold, the number of occurrences or co-occurrences for the filtering option. The default value is one, i.e. no annotation will be filtered out.

*- threshold*
This is a p-value threshold for similarity of support vectors for a pair of annotations. An input to this option is 1.0. We recommend the level of p-value threshold for large dataset be less than 0.05.

**Output File of *anet***
**==========**

There are 9 columns in the output file of *anet*:

1. Annotation 1 – also referred as Anno1

2. Annotation 2 - also referred as Anno2
3. Correlation - Correlation coefficient value
4. p-value - p-value for the coefficient
5. #Records(Anno1,Anno2) - Number of records where Anno1 and Anno2 co-occur
6. #Records(Anno1) - Number of records with Anno1
7. #Records(Anno2 ) - Number of records with Anno2
8. #CoAnnos(Anno1) - Number of annotations co-occurred with Anno1
9. #CoAnnos(Anno2) - Number of annotations co-occurred with Anno2

In addition, *anet* generates two additional auxiliary files:
<output file>.map – a list of unique annotations in the transaction records with their IDs.
<output file>.2num – transaction records with annotation IDs instead of names

**Example:**
anet --file=GB.data.tr --output= GB.data.tr.out --by=by_item_count --count=6 --threshold=0.05

In this example *anet* is used to process GB.data.tr, generated by the program *t2t.pl* in the previous example. *by_item_count* (option --*by*) excludes all annotations, that are found in less than 6 genomes (option --*count*). Threshold for the p-value is set to 0.05.