AIRnet User Manual

by David Oviatt

08/18/09

Table of Contents

1 Introduction

Using microarray data, AIRnet creates network graphs representing the correlations between genes. AIRnet also, when given the proper data, will compare the networks of two different genotypes, providing the user with a third graph to easily view differences in gene correlation between the genotypes.

AIRnet works best when files intended for reading, such as microarray files, are stored in the data folder. The graphs and other files AIRnet produces are all placed in the output folder.

The various calls to other scripts and programs made by AIRnet is outlined in Illustration 1.

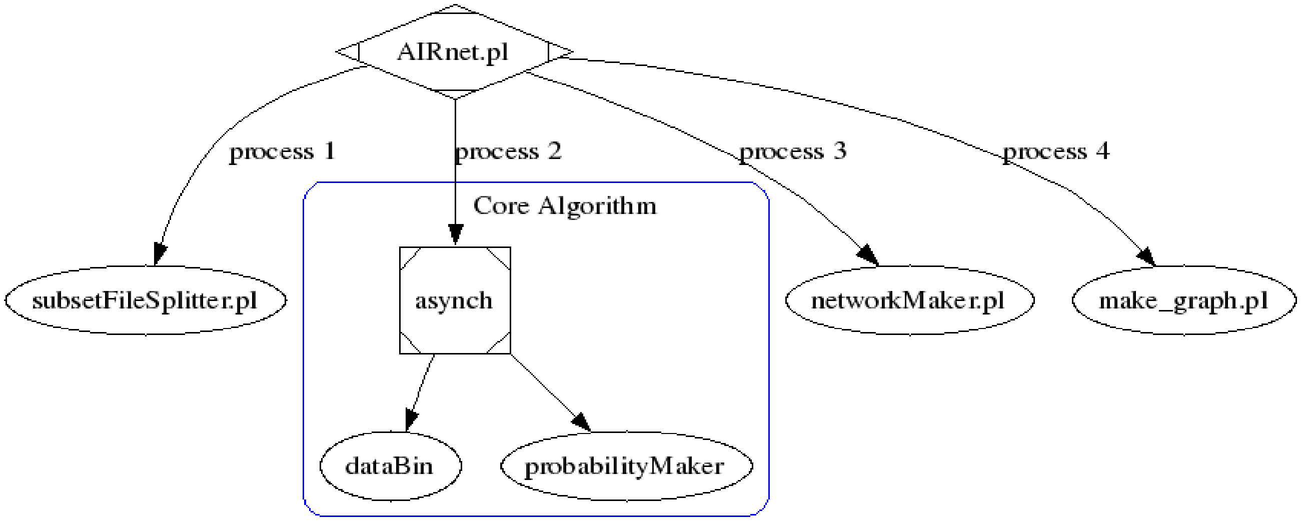

Illustration 1: Calling tree for AIRnet

2 Requirements

AIRnet uses OpenMP, PERL, graphviz, and the g++ compiler. More information about each of these, and their respective installation instructions can be found at their respective websites, listed in Table 1: Additional resources needed to run AIRnet.

AIRnet uses 3.5 GB of RAM when running full-genome analysis (20,000+ genes) and uses approximately 7 GB of disk space for a full-genome analysis. Analysis of smaller sets of genes require much less RAM and disk space. An analysis of 100 genes will use only a minimal amount of RAM, and less than 1 MB of disk space.

Tool:	Where to get it:
OpenMP	http://openmp.org
PERL	http://www.perl.org
graphviz	http://www.graphviz.org
g++ compiler	http://gcc.gnu.org

Table 1: Additional resources needed to run AIRnet

Runtime for AIRnet varies based on the number of genes being analyzed, number of microarray samples, and computer hardware. As AIRnet is designed to take advantage of multi-core architecture, it is recommended to use a multi-core machine, particularly when running a full-genome analysis, or any other similarly large dataset.

3 Installation

The following steps illustrate all that is necessary to install AIRnet:

1. Download AIRnet.tar.gz from http://dna.cs.byu.edu/airnet and save it in your home directory.
2. Extract files from the AIRnet in a terminal (a) or using the file manager (b):
   a) Extracting files in the terminal
      i) Open a terminal.
      ii) Type '`tar zxf AIRnet.tar.gz`'
   b) Extracting files using the file manager
      i) open AIRnet.tar.gz with Archive Manager.
      ii) Click 'extract' on the tool bar.
      iii) Select 'All Files' under 'Files' heading.
      iv) Check the 'Re-create folders' box under the 'Actions' heading
      v) Click 'Extract' in the lower right corner.
3. Open a terminal and change directories to where AIRnet was extracted

   Ex: cd AIRnet
4. Type '`which perl`' to ensure PERL is installed on your system. If you receive an error

message in steps 4-6, refer to Table 1 for information on where to obtain the appropriate tool.

5. Type '`which g++`' to ensure the g++ compiler is installed on your system.
6. Type '`which neato`' to ensure graphviz is installed on your system.
7. Type './AIRnet.pl -compile_only'

3.1 Troubleshooting

If the AIRnet script does not run, refer to step 4. If PERL is properly installed, try running the command: `chmod 755 AIRnet.pl` and re-try step 7.

If g++ is not installed on your machine, you receive the following message after step 7:

```
Can't exec "g++": No such file or directory at ./AIRnet.pl line 185.
```

You will need to install g++. More information is available for this process at the website listed in Table 1.

Finally, if you receive an error message similar to:

```
which: no neato in (/usr/lib/bin:/usr/local/bin)
```

You need to install graphviz. More information is available for this process at the website listed in Table 1.

4 Usage

Basic usage of AIRnet on the command line is:

```
./AIRnet.pl {options} [threshold] [difference] [bins] [microarray file]
[split index] [genes of interest file]+
```

Type "./AIRnet.pl --help" to access the help menu. The help menu contains further information about AIRnet's required arguments and options. Examples and explanations follow below.

```
./AIRnet.pl .50 .40 F2 data/GDS681.soft 6 data/Ts1Cje_genes.txt
```

This is a very basic example. No options are specified, .50 is the minimum threshold for including an edge in the final graph, and scores must differ by at least .40 in order to count as being different. The clustering of the data will be done using 2 clusters, and the Full Set of gene data for each genome will be used in clustering, rather than dividing the set up by gene before clustering. The microarray file, GDS681.soft (found in the data folder) will be split into 2 files, with the first 6 columns put into a file labeled 'normal', the following columns placed in a file labeled 'diseased'. Both the normal and diseased have data for genes listed in either the Ts1Cje_genes.txt or the jaw_development_genes.txt files, both of which are in the data folder.

```
./AIRnet.pl .50 .40 F2 data/GSE11472_family.soft “euploid” “Ts1Cje” 6
data/Ts1Cje_genes.txt
```

This example is almost exactly the same as the previous example. The microarray file has been changed to GSE11472_family.soft, which is in a different format than the previous file. The split index now means nothing, as this format of microarray file is not displayed in tab-separated columns, and can be omitted, as in the following example.

```
./AIRnet.pl .50 .40 F2 data/GSE11472_family.soft “euploid” “Ts1Cje”
data/Ts1Cje_genes.txt
```

Omitting the split index in a tab-separated column file sets the split index value to 0, and has a similar effect to using the --single option.

5 Microarray File Formats

Microarray files should be saved in the data folder. Currently, AIRnet will read GSE family.soft formatted files as well as files with the relevant data in tab-separated columns, both of which can be found at the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/).

5.1 GSE Files

When reading the GSE family.soft type files, unique genotype labels must be supplied for the two genotypes contained in the file. Please see the section on Genotype Labels (page) for more information on how to select proper labels.

5.2 GDS and Tab-separated Files

GDS and other tab-separated files are much less stringent about the genotype labels. Any label can be used, as long as the two labels are different.

Examples of tab-separated files AIRnet is able to read are shown below:

```
ID_REF     IDENTIFIER  sample1  sample2  sample3  sample4  sample5

100001_at  Irf2bp1     170.700  110.200  161.800  119.300  141.000

100002_at  Cd3g        1.600    1.600    1.600    2.400    1.200

100003_at  Traf4       168.200  93.200   170.900  138.700  168.400

100004_at  Sox2        130.800  99.500   213.000  119.300  154.300

100005_at  Shh         15.400   7.800    15.100   7.900    6.500
```

This is a standard .soft format file.

```
IDENTIFIER sample1     sample2  sample3  sample4  sample5
```

```
Irf2bp1     170.700     110.200    161.800   119.300   141.000
Cd3g        1.600       1.600      1.600     2.400     1.200
Traf4       168.200     93.200     170.900   138.700   168.400
Sox2        130.800     99.500     213.000   119.300   154.300
Shh         15.400      7.800      15.100    7.900     6.500
```

Similar to the .soft format file, but without the first column. In the first two examples, IDENTIFIER may be substituted with TargetID.

```
IDENTIFIER  Irf2bp1     Cd3g       Traf4     Sox2      Shh
sample1     170.7       1.6        168.2     130.8     15.4
sample2     110.2       1.6        93.2      99.5      7.8
sample3     161.8       1.6        170.9     213       15.1
sample4     119.3       2.4        138.7     119.3     7.9
sample5     141         1.2        168.4     154.3     6.5
```

Files in this format must be run with the --transpose option. In all three examples, the word IDENTIFIER or TargetID can be substituted with anything IF the -single option is used.

6 Threshold

The threshold value is the minimum allowed likelihood for any pair of genes, that will still result in the inclusion of the interaction in the resulting graphs.

The threshold can be any floating point number between 0.0 and 1.0. Numbers closer to 1.0 tend to produce smaller, more precise networks, while lower values, those close to 0.0, tend to produce larger networks, though with lower accuracy.

```
./AIRnet.pl .50 .40 G2 data/GDS681.soft 6 data/Ts1Cje_genes.txt
```

In this example, the threshold value is .50. This means that all of the connections in the inferred network will have scores >= .50, or <= -.50.

7 Difference

In datasets where two genotypes are present, AIRnet will compare the likelihood of each pair of genes interacting for the two genotypes. The difference value is the threshold used to determine if the gene pairs are functioning the same across genotypes or differently.

For example, if the likelihood of gene A interacting with gene B is 0.8 in genotype 1 and 0.3 in genotype 2, a difference value <= 0.5 will decide the two genotypes behave differently, while a difference value > 0.5 will decide they behave the same. Difference values should be between 0.0 and 2.0. Smaller values will show greater differentiation between genotypes.

```
./AIRnet.pl .50 .40 G2 data/GDS681.soft 6 data/Ts1Cje_genes.txt
```

In this example, the difference value is .40. This means that if the scores of any pair of genes for each genotype are within .40 of each other, the interaction between the two genes will be considered to be the same type of interaction.

8 Bins

AIRnet divides the gene expression data into different bins using a k-means clustering algorithm. The bins variable is used to specify the number of bins the user would like to divide the data into. Additionally, the data can be clustered using data for each gene as a dataset, or using all data as a dataset.

```
./AIRnet.pl .50 .40 G2 data/GDS681.soft 6 data/Ts1Cje_genes.txt
```

In this example, the bins variable is G2. Because the first character is 'G', the clustering algorithm will treat the data for a single gene as its entire dataset. The clustering algorithm will be run once for each gene in the dataset.

```
./AIRnet.pl .50 .40 F3 data/GDS681.soft 6 data/Ts1Cje_genes.txt
```

In this example, the bins variable is F3. Because the first character is 'F', the clustering algorithm will use all the data as a single dataset. In this case, the clustering algorithm will only be run once, but on a much larger dataset.

All bin values must begin with either 'F' or 'G', and be immediately followed by a number.

9 Genotype Labels

9.1 General Guidelines

Genotype labels are used to identify AIRnet's output files. Labels of more than one word must be enclosed in double quotes (“). Additionally, genotype labels **may not** contain the following characters:

```
 ( ) [ ] { } ″ ' . ? * + = $ ^ \ / | &
```

Using labels that contain only alpha-numeric characters, spaces, -, and _ whenever possible will eliminate errors produced by invalid characters, and produce more desirable results.

The default values for the genotype labels are normal and diseased. While these defaults will work when processing the tab-separated column files, they will not function properly for the GSE family.soft formatted files. In order to more accurately label the output files, it is highly recommended that the genotype labels are always provided.

9.2 GSE Labels

In the GSE files, the genotype labels are also used to divide the microarray data between different genotypes. Because of this, it is nearly guaranteed that the default labels will not perform correctly. A suitable label can normally be found in the headings before each sample's data, particularly in the ”!Sample_characteristics” headings. Below is an example of using AIRnet, and the headers used to identify the genotype labels.

```
./AIRnet.pl .50 .40 G2 data/GSE11472_family.soft “control littermates”
“Down syndrome mouse model” data/Ts1Cje_genes.txt
```

Headings from the `data/GSE11472_family.soft` microarray file:

```
!Sample_characteristics_ch1 = Strain: B6EiC3Sn, control littermates for
Ts65Dn mice (B6EiC3Sn-T(1716)65Dn) Age: postnatal day 2 (P2) Tissue:
cerebellum
```

```
!Sample_characteristics_ch1 = Strain: B6EiC3Sn-T(1716)65Dn (Down syndrome
mouse model)
```

The two phrases, 'control littermates' and 'Down syndrome mouse model', are only used in these lines above and do not contain anything but letters and whitespace. The lines are repeated in the headers for each of the samples, thus the two phrases can be used to accurately distinguish each sample as its particular genotype. “B6EiC3Sn-T” could not be used to identify the Down Syndrome mouse model data, because it also appears in the control samples' characteristics description as well.

A second example is shown below:

```
./AIRnet.pl .50 .40 G2 data/GSE11472_family.soft “euploid” “Ts1Cje”
data/Ts1Cje_genes.txt
```

```
!Sample_characteristics_ch1 = Genotype : Ts1Cje
```

```
!Sample_characteristics_ch1 = Genotype : euploid
```

9.3 GDS and Tab-separated Labels

In GDS and other tab-separated files, the default genotype labels may be used, however it is still recommended that the user provides more descriptive labels. See the “Split Index” section for more information on properly dividing GDS and tab-separated files by genotype.

10 Split Index

The split index is used to tell AIRnet how many microarray experiments belong in the file labeled with the first genotype label. Because the split index is only used in GDS and other tab-separated files, it is unnecessary when using GSE formatted files. Additionally, when using the '-single' option, the split index must be omitted. In all cases, the original file containing all of the microarray data will remain intact.

When dividing the files into two genotype-labeled files, AIRnet puts the number of data columns defined by the split index into a file labeled with genotype label 1, starting with the left-most data column. The remainder of the data columns are placed into a file labeled with genotype label 2. An example follows, with the split index in bold:

```
./AIRnet.pl .50 .40 G2 data/GDS681.soft 2 data/Ts1Cje_genes.txt
```

A sample of the microarray file, `GDS681.soft`, found in the data folder:

```
                      |    Genotype 1  | |        Genotype 2       |
ID_REF     IDENTIFIER  sample1   sample2  sample3  sample4   sample5
100001_at  Irf2bp1     170.700   110.200  161.800  119.300   141.000
100002_at  Cd3g        1.600     1.600    1.600    2.400     1.200
100003_at  Traf4       168.200   93.200   170.900  138.700   168.400
100004_at  Sox2        130.800   99.500   213.000  119.300   154.300
100005_at  Shh         15.400    7.800    15.100   7.900     6.500
```

becomes two files (created with the default genotype labels), `GDS681_normal.soft`:

```
ID_REF     IDENTIFIER  sample1   sample2
100001_at  Irf2bp1     170.700   110.200
100002_at  Cd3g        1.600     1.600
100003_at  Traf4       168.200   93.200
100004_at  Sox2        130.800   99.500
100005_at  Shh         15.400    7.800
```

and `GDS681_diseased.soft`:

```
ID_REF     IDENTIFIER  sample3   sample4   sample5
100001_at  Irf2bp1     161.800   119.300   141.000
100002_at  Cd3g        1.600     2.400     1.200
100003_at  Traf4       170.900   138.700   168.400
```

```
100004_at  Sox2        213.000  119.300  154.300
100005_at  Shh         15.100   7.900    6.500
```

11 Genes of Interest File

The genes of interest file should be the name of a file, which contains a tab-separated list of genes. A small sample of such a file is shown below. AIRnet will run using a subset of the microarray data, including all of the data for the genes listed in the genes of interest file. More than one file may be specified, as well as either or both of the predefined keywords.

```
GART       SON        C21orf60   CRYZL1
ITSN       ATP5O      BB856722   MRPS6
SLC5A3     KCNE2      C21orf51   AK006730
AK016199   KCNE1      BB644499   DSCR1
```

In the resulting network graphs, AIRnet will color the nodes according to what file the gene they represent was found in and a key will be created at the bottom of each graph, labeling the colors with the corresponding file names. A few examples of the genes of interest files are given below.

```
./AIRnet.pl .50 .40 G2 data/GDS681.soft 6 data/Ts1Cje_genes.txt
```

```
./AIRnet.pl .50 .40 G2 data/GDS681.soft 6 data/Ts1Cje_genes.txt
data/wnt_pathway_genes.txt
```

11.1 all

The keyword 'all' uses all of the genes in the microarray file to construct the network, not just the genes specified in the genes of interest files. Although all of the genes are used when creating the network, only the connections in the network that directly involve one or more of the genes found in the genes of interest files will be printed in the graph.

```
./AIRnet.pl .50 .40 G2 data/GDS681.soft 6 data/Ts1Cje_genes.txt all
```

11.2 non-key

When 'non-key' is used in addition to 'all', AIRnet will show the entire inferred network, including connections between genes that are not listed in the genes of interest files.

```
./AIRnet.pl .50 .40 G2 data/GDS681.soft 6 data/Ts1Cje_genes.txt all non-key
```

12 Options

Each option must be immediately preceded by '-', as in -help. The description of each option is followed by an example of using it and further explanation, if needed. Excluding the help and compile_only options, multiple options

12.1 help

Opens a help menu with quick reference to most topics covered in the manual.

```
./AIRnet -help
```

12.2 compile_only

Compiles AIRnet for use with the local machine and exits.

```
./AIRnet -compile_only
```

12.3 compile

Compiles AIRnet for use with the local machine and then runs if the appropriate arguments have been provided.

```
./AIRnet.pl -compile .50 .40 G2 data/GDS681.soft 6 data/Ts1Cje_genes.txt
```

12.4 clean

Removes non-essential files after run-time completion.

```
./AIRnet.pl -clean .50 .40 G2 data/GDS681.soft 6 data/Ts1Cje_genes.txt
```

12.5 directed

Creates a directed gene regulatory network. Requires one wild-type experiment and one knockout/knockdown experiment for each gene in the dataset.

```
./AIRnet.pl -directed .50 .40 G2 data/GDS681.soft 6 data/Ts1Cje_genes.txt
```

12.6 fast

Logs into a remote machine. Must be followed by the login name and the name of the machine you are logging in to. In order for this option to work, AIRnet should be installed the home directory, and your account must be mounting the same home directory, regardless of what machine you are physically sitting in front of. **If any of this does not make sense, do not use the fast option.**

```
./AIRnet.pl -fast doviatt shakespeare .50 .40 G2 data/GDS681.soft 6
data/Ts1Cje_genes.txt
```

This example uses the user name `doviatt` and the computer `shakespeare`, which is where the core algorithms for AIRnet will be run. After running this command, AIRnet will test the connection to shakespeare, and let me know if there is a problem. If there are no problems, I will be asked to give my password in order to log into shakespeare, and AIRnet will continue execution like normal.

12.7 transpose

Transpose the gene expression matrix. Use ONLY IF the genes are listed in columns with the samples in rows.

12.8 single

Creates one network from the microarray data, rather compare two networks. Negates the need for 'difference' and 'split index' arguments.

13 Output Files

AIRnet will produce several files in the output folder, most notable are the gif files and the html files. The gif is a picture representation of the network. The html is a web page with the picture of the network in it, with the addition of the nodes in the graph linking to an NCBI search for that gene.

In most cases, AIRnet will produce three sets of files, one for each genotype and one to show the differences between the genotypes. The two genotype file groups will be labeled according to the genotype labels provided, with defaults of normal and diseased. The file group showing the differences between the genotypes are labeled with the original microarray file name. All files are also labeled with the threshold used, the number of clusters and the type of grouping of data used prior to clustering.

14 Additional Files

*.dot - used to create the html and gif files. One is produced by the asynch program, which is used by the networkMaker.pl script to produce other .dot files and the html and gif files.

*.mtx - shows the microarray data after clustering.

*_CTS.csv - this file contains the correlation matrix. It is used by the make_graph.pl script and by the Dataset_Comparer program.

*.lst - this contains a list of all the edges in a particular network. It is used by the make_graph.pl script.

15 Add ons

Currently, there is one add on for AIRnet, the Dataset_Comparer program. This program is used to compare two datasets that AIRnet has already been run on. Each dataset must contain two genotypes. If you can consider the genotype comparisons that AIRnet runs to be the first derivative of the microarray data, then the comparison done by the Dataset_Comparer program is like the second derivative. It will compare the differences and similarities of the different genotypes and data sources, comparing four networks instead of two.

In order to use this add on, you must take the following steps:

1. Have AIRnet installed, according to the steps outlined in section 3.
2. obtain two datasets that are compatible with AIRnet
3. run AIRnet on each dataset
4. from the AIRnet directory, fun the command: `g++ -fopenmp addons/Dataset_Comparer.cpp addons/Score_Table.cpp -o addons/Dataset_Comparer` (this need only be done once per installation)
5. finally, run the command `addons/Dataset_Comparer [Microarray data file 1] [genotype label 1a] [genotype label 1b] [Microarray data file 2] [genotype label 2a] [genotype label 2b] [edge threshold] [difference threshold] [genes of interest file]+`

As an example, if I had run AIRnet with the microarray file `data/GDS681.soft`, using the genotype labels "Normal" and "Ts1Cje", and the microarray file `GDS682.soft`, using the genotype labels "euploid" and "trisomic", my commands to run AIRnet would have looked something like this:

```
./AIRnet.pl .5 .4 G2 data/GDS681.soft “Normal” “Ts1Cje” 6
data/Ts1Cje_genes.txt
```

```
./AIRnet.pl .5 .4 G2 data/GDS682.soft “euploid” “trisomic” 6
data/Ts1Cje_genes.txt
```

The command to run Dataset_Comparer would look like this:

```
addons/Dataset_Comparer data/GDS681.soft “Normal” “Ts1Cje” data/GDS682.soft
“euploid” “trisomic” .5 .4 data/Ts1Cje_genes.txt
```

The result of this is a dot file, which can be used with graphviz to create a picture highlighting the differences in the network. In this example, the dot file would be `output/GDS681_GDS682_0.5–0.4.dot`. The command to create a picture from this file is:

```
neato -Tgif output/GDS681_GDS682_0.5--0.4.dot -ooutput/GDS681_GDS682_0.5—
0.4.gif
```

The extra o in -ooutput/ GDS681_GDS682_0.5—0.4.gif is not a mistake. The first o tells neato that this is the name of the file to put the picture into.