MDR-PDT User's Guide

MDR-PDT Version 1.0beta1

Compile Date: 10/20/06

About this document.

This document is an introduction to the use of the MDR-PDT analysis program described in Martin, et al (2006). The analysis program described is currently approaching completion, however as testers give us input, it is possible that things will change. Be sure that the version of this document matches the version of the application being used.

Because the details of the method are described in Martin, et al (2006), this document will be focused on the use of the application rather than how the analysis works.

Datasets Appropriate for MDR-PDT

While MDR (Ritchie, et al., 2001) was designed to discover locus-disease associations in case-control data, MDR-PDT was designed to discover locus-disease associations in pedigree data using the genotype-PDT statistic (Martin et al., 2003). This version of MDR-PDT is capable of analyzing input in standard pedigree format, which allows for a mix of trio data and discordant sibships. In both types of family structures, the informative element is the transfer of genotypes to an affected child as compared with the genotypes transferred to an unaffected child. In the trio data, a virtual unaffected sibling is created who is the genotypic complement of the affected child as derived from the original affected child and the two parents.

Each locus must contain only two alleles and must denote missing data with a zero for each allele that is missing.

Pedigree Format

All data files intended for use with MDR-PDT should be written with one of the following delimiters: space, tab or commas. The application ignores empty fields, so two or more spaces side beside are treated as a single space.

Trios

Trio families are represented by two parents and a single affected child. During the preparation of a triad, a virtual unaffected sibling is produced. The production of virtual siblings is documented in the pedigree report. These entries include the description of the parents, the affected sibling and the newly created virtual sibling. The new children will be given new IDs. It should be noted that missing data from any parent will be propagated to both siblings.

Discordant Sibships

A discordant sibship is a sibship in which there one ore more affected individuals and at least one unaffected.

Table 1: A description of the data in pedigree format.

Column Number	Description
1	Pedigree number (must be an integer)
2	Individual ID number (must be an integer)
3	ID of father (0 if this individual is a founder)
4	ID of mother (0 if this individual is a founder)
5	First offspring ID
6	Next paternal sibling ID
7	Next paternal sibling ID
8	Sex (1=male, 2=female, 0=unknown)
9	Proband status (1=proband, all others have a 0 in this field.)
10	Disease status (2=affected, 1=unaffected, 0=unknown)
11	The 1rst allele pair (must be a pair of integers)
12	
...	...
n-1	The n/2th allele pair
n	

Application Usage

There are currently 3 modes of operation: configuration generation, analysis and model exploration.

Configuration Generation

To simplify the generation of configuration files, the application provides a way to display an example configuration file. Users can capture this display to a file and use it for analyses. The example configuration produced depends on the Analysis Style selected as the 1rst argument. Currently, there is only one option, PDT.

Optional arguments may follow in order to specify important runtime options. While all other arguments are optional, they must be present in the order on the command line to be interpreted correctly (i.e. If you wish to change the Max Model Size, you have to set Data Filename and Min Model Size as well even if you are happy with the defaults).

Argument #	Purpose	Possible Values
1	**Analysis type**	**PDT**
(2)	Data Filename	Name of your pedigree formatted input file
(3)	Min Model Size	1 to Max Model Size
(4)	Max Model Size	Min Model Size or more

Distributed with the application is a simulated data set named ExamplePedigree.ped. In order to create configuration file for running a 1 & 2 SNP search with permutation tests on the example pedigree file, type the following:

```
./mdr-pdt PDT ExamplePedigree.ped 1 2  > Example.pdt.2
```

The resulting file, Example.pdt.2 should be ready to use. At this point, changes can be made to Example.pdt.2 according to the user's needs.

The following is the result of running the line above with the the beta release version of the application.

```
bash-3.1$ more Example.pdt.2
/******************* Basic Settings
 * Describe the type of models of interest. Models consist on 1 or more SNPs.
 * The minimum number of SNPs in a model to be investigated. Valid settings: 1..COMBO_END
COMBO_START          1
```

```
 *  The maximum number of SNPs to be considered. Valid settings: COMBO_START...
COMBO_END            2
 * You can exclude loci from analyses by adding them to the following variable
 * The application does not actually keep the data associated with these loci, but they
 * do retain the positions, so model 13x22 with none excluded would be the same as
 * 13x22 with 14, 15 and 16 excluded
EXCLUDE_LOCUS
* The following would exclude loci 1, 4 and 31 from analyses
*EXCLUDE_LOCUS 1 31 4
 *  The value used to indicate that an individual is affected
AFFECTED_VALUE        2
 *  The value used to indicate that an individual is unaffected
UNAFFECTED_VALUE      1
 *  All other individuals will be considered to be of unknown status and will not contribute to the calculations



/****************** Input format
 * PEDIGREE   - For PDT anayses there is only one format supported
INPUTFORMAT          PEDIGREE
 *  The name of the input file where your SNP data is to be found. This file must be space delimited
INPUTFILE            ExamplePedigree.ped
 *  There is only one Pedigree analysis currently available
 *  You can exclude certain pedigrees from analysis. Buy removing the asterisk from the line below
 *  will tell the application to exclude pedigrees 13, 21 and 1003
*PEDIGREE_EXCLUSIONS 13 21 1003
ANALYSISSTYLE        PDT
 *  Turn On/Off Triad expansion
EXPAND_TRIOS         Yes
 *  Matched Odds Verbose On/Off. Turning this on generates the table of pair contributions
VERBOSE_MATCHED_ODDS_RATIO Off
 *  Evaluation Verbose Yes/No
VERBOSE_EVALUATION    No



/****************** Reporting
 * Reports are set up in the form: REPORT_NAME.EXT Each type or report has it's own different extension.
 * Any report whose extension is STDOUT will be redirected to standard out instead of written to a file.
 * Any report whose extension is NOLOG will be skipped altogether.
 * By default, REPORT_NAME is just the name of the configuration file
 *REPORT_NAME MyReportName
EXT_DISTRIBUTION      pdist     //This is where the distribution of the p-tests is written
EXT_REPORT           STDOUT    //This is where the final results are written
EXT_PEDIGREE         pedigree    //Pedigree related errors and notes are logged here
```

```
/****************** Permutation Tests
 * Number of permutation runs to be executed. 1000 is recommended.
PTEST_COUNT          1000
 * The seed associated with the tests. Each test gets a new seed
PTEST_SEED           1371
 * Reporting can be done based on a p-value threshold. Any model whose significance exceeds this value will be reported.
PVAL_THRESHOLD       0.05
```

Analysis

To perform analysis, simply create a valid configuration file and execute the application with the configuration file as the sole parameter.

To run the analysis setup with the example data file using the configuration file, Example.pdt.2, type the following:

Example:

```
./mdr-pdt Example.pdt.2
```

The following is an excerpt from a run based on the example dataset included with the application.

```
mdr-pdt - Multifactor Dimensionality Reduction - Pedigree Disequilibrium Test: 1.0 (0)
Vanderbilt University
Center for Human Genetics Research
Marylyn Ritchie & Eric Torstenson
Please forward any comments or errors to: mdr-pdt@chgr.mc.vanderbilt.edu

Configuration:
                                Data Source: [ExamplePedigree.ped]
                            Input File Type: Pedigree Format
                                             * Trios were expanded to valid
                                               DSPs where parental data was present
                    Value Denoting Affected: 2
                  Value Denoting Unaffected: 1
                     Total Individuals Seen: 600
   Total individuals to be used in analysis: 800
                            Affected Individuals: 400
                          Unaffected Individuals: 400
              Number of SNPS to be analyzed: 10
                                Results Log: Standard Out
                           Distribution Log: ExamplePedigree.mdrpdt.pdist
                               Pedigree Log: ExamplePedigree.mdrpdt.pedigree
                                 Model Size: 1-2
                          Permutation Tests: 1000
                      Permutation Test Seed: 1371
                             Analysis Style: Pedigree Analysis using PDT statistic
  Searching
```

This is just an overview of the analyses that have been run.

Application output continued:

Below are the search results. If EXT_REPORT were set to something other than STD_OUT, the details on the next few pages would be part of the output named *.report. Each model is treated with the subsections: Model Overview, Classification Details, Matched Odds and Statistic Summary.

The first few lines represents an overview of the findings:

```
              Model      PDT T     Pred.
                 ID       Stat     Error
    ------------------------------------------------
                  1     2.8846     44.75%
               5x10     9.7468     26.25%
```

Followed by Details for the specific models that were selected with the highest t-statistic for each model size.

```
Model Details [ 1 ]
           Affected:  50.00%         Unaffected:  50.00%
       Missing Data:   0.00%

Genotype Individuals
                  1        A     U   TOT   Ratio
    ------------------------------------------------
                 1/1      84   104   188    0.81
                 1/2 *   226   184   410    1.23
                 2/2      90   112   202    0.80



Model Details [ 5x10 ]
           Affected:  50.00%         Unaffected:  50.00%
       Missing Data:   0.00%

Genotype Individuals
             5   10        A     U   TOT   Ratio
    ------------------------------------------------
           1/1   1/1       0    32    32    0.00
           1/1   1/2 *    91    52   143    1.75
           1/1   2/2       0    20    20    0.00
           1/2   1/1 *    98    44   142    2.23
           1/2   1/2       0    68    68    0.00
           1/2   2/2 *    98    54   152    1.81
           2/2   1/1       0    46    46    0.00
           2/2   1/2 *   113    60   173    1.88
           2/2   2/2       0    24    24    0.00
               *Indicates models that are determined to be High Risk

                Heterozygote alleles are not necessarily ordered the
                same as they were found in the original data

Pedigree Search of 55 : 55 completed in 0.06 seconds
```

Each permutation is logged in order to give the user a sense of how much time is required before completion.

```
Performing 100 Permutation Tests:
Test #  Load Time (s)             Execution Time(s)    Models Seen
     0          0.000                         0.060             55
     1          0.010                         0.070             55
     2          0.000                         0.060             55
     3          0.000                         0.070             55
     4          0.000                         0.070             55
     5          0.000                         0.070             55
                              (Lines Removed)
   999          0.000                         0.050             55
```

Based on the various model's estimated p-values, the final part of the report contains all “statistically” interesting models. Interesting models are those whose p-values fall below the threshold set by the configuration parameter: PVAL_THRESHOLD. The distribution required for the t-statistics is an N-Distribution. As a result, models are reported separately according to their size.

```
Distribution Report 1 SNP(s) per model
        * Only models with a p-value < 0.001 are reported.
           T         Statistical
Model   Statistic    Significance
   1       2.885       p < 0.001
  10       2.073       p < 0.001
   7       1.925       p < 0.001
   5       1.812       p < 0.001
   4       1.793       p < 0.001
   6       1.508       p < 0.001
   2       1.134       p < 0.001
   3       1.114       p < 0.001
   8       1.086       p < 0.001


Distribution Report 2 SNP(s) per model
```

```
     * Only models with a p-value < 0.001 are reported.
            T         Statistical
Model   Statistic    Significance
 5x10       9.747       p < 0.001
  1x7       3.757       p < 0.001
  4x8       3.389       p < 0.001
 7x10       3.345       p < 0.001
  1x3       3.312       p < 0.001
  1x8       3.206       p < 0.001
 1x10       3.190       p < 0.001
  5x8       3.170       p < 0.001
  1x9       3.151       p < 0.001
  1x5       3.130       p < 0.001
  1x4       3.101       p < 0.001
 2x10       3.087       p < 0.001
  4x6       2.991       p < 0.001
  1x6       2.885       p < 0.001
        (Lines Removed)
```

The last part of the output is a summary of the top models.

```
Top Model of the PDT Search:
Model
Size           Model  T Statstic     p Value
1                  1      2.8846     p < 0.0010
2               5x10      9.7468     p < 0.0010
bash-3.1$
```

Model Exploration

During regular analyses, only the best model for each model size is described in high detail. However, should users wish to explore other models, mdr-pdt provides a way to interactively explore any model they wish.

Exploration is easy to do, but it requires a few pieces of information. First, the configuration file specifying details about the data (format and location) as well as the analysis style desired. One or more models should be given on the command line separated by space. The models are specified using the SNPs position in the repository separated by "x"s.

Example:

If in a previous analysis two other significant models were reported as 1x10 and 5x10, we could type the following:

```
./mdr-pdt  sampledata.mdrpdt.3L 1x10 5x10
```

Below is an example of the exploration of a single model, 2x10

```
bash-3.1$ ./mdr-pdt Example.pdt.2 2x10
mdr-pdt - Multifactor Dimensionality Reduction - Pedigree Disequilibrium Test: 1.0 (0)
Vanderbilt University
Center for Human Genetics Research
Marylyn Ritchie & Eric Torstenson
Please forward any comments or errors to: mdr-pdt@chgr.mc.vanderbilt.edu

Configuration:
                                Data Source: [ExamplePedigree.ped]
                            Input File Type: Pedigree Format
                                             * Trios were expanded to valid
                                               DSPs where parental data was present
                    Value Denoting Affected: 2
                  Value Denoting Unaffected: 1
                     Total Individuals Seen: 600
   Total individuals to be used in analysis: 800
                                 Affected Individuals: 400
                               Unaffected Individuals: 400
              Number of SNPS to be analyzed: 10
                                Results Log: Standard Out
                           Distribution Log: ExamplePedigree.mdrpdt.pdist
                               Pedigree Log: ExamplePedigree.mdrpdt.pedigree
                                 Model Size: 1-2
```

```
                             Risk Assesment: Based on each model's local threshold
                          Permutation Tests: 1000
                      Permutation Test Seed: 1371
                             Analysis Style: Pedigree Analysis using PDT statistic


Model Details
        Model ID: 2x10
        Total Affected: 400 (100.00%)    Total Unaffected: 400 (100.00%)

          Genotype                 Individual Counts
           2   10       Affected  Unaffected       Total       Ratio
   --------------------------------------------------------------------
         1/1   1/1           19          28          47      0.6786
         1/1   1/2 *         54          42          96      1.2857
         1/1   2/2           20          38          58      0.5263
         1/2   1/1           61          64         125      0.9531
         1/2   1/2           98         102         200      0.9608
         1/2   2/2 *         44          34          78      1.2941
         2/2   1/1           18          30          48      0.6000
         2/2   1/2 *         52          36          88      1.4444
         2/2   2/2 *         34          26          60      1.3077
                *Indicates models that are determined to be High Risk

                 Heterozygote alleles are not necessarily ordered the
                 same as they were found in the original data



Classification Details:
                     Correctly       Incorrectly
                    Classified        Classified
    Affected   184 ('High-Risk')   216 ('Low-Risk')
  Unaffected   262 ('Low-Risk')    138 ('High-Risk')
              55.75%              44.25%           of 800

Summary:
Model ID:2x10
                 Sum(D): 46
               Sum(D*D): 222
            T-Statistic: 3.0873
```

Pedigree Report

During the parsing of a pedigree file, and throughout the generation of virtual sibs and discordant sib pairs, mdr-pdt logs important details.

Pedigree Contributions

Below is the first few lines from the pedigree report generated when analyzing the example data distributed with the application.

```
bash-3.1$ more Example.pdt.3.pedigree
Total Families Identified: 200
The following represents the contributions of each pedigree to the DSPs used in the analyses.
Column 1 is the pedigree ID. 2 represents the total number of individuals contributed to the
analysis.
   Pedidgree   Total  Unique  Unique   Discordant sibship breakdown
      ID             Cont.    Aff.   Unaff      FxM [ sibs ] ...
  Pedigree# 0         4       2       1 :       2x3 [ *1  4 *5 ]
  Pedigree# 1         4       2       1 :       2x3 [ *1 *4  5 ]
  Pedigree# 10        4       2       1 :       2x3 [ *1 *4  5 ]
  Pedigree# 100       4       2       1 :       2x3 [ *1 *4  5 ]
  Pedigree# 101       4       2       1 :       2x3 [ *1  4 *5 ]
  Pedigree# 102       4       2       1 :       2x3 [ *1  4 *5 ]
  Pedigree# 103       4       2       1 :       2x3 [ *1 *4  5 ]
  Pedigree# 104       4       2       1 :       2x3 [ *1 *4  5 ]
  Pedigree# 105       4       2       1 :       2x3 [ *1  4 *5 ]
  Pedigree# 106       4       2       1 :       2x3 [ *1  4 *5 ]
  Pedigree# 107       4       2       1 :       2x3 [ *1 *4  5 ]
  Pedigree# 108       4       2       1 :       2x3 [ *1 *4  5 ]
                      (Continued)
```

Because this is simulated data, the report is very consistent. Each pedigree is contributing 4 members (2 DSPs) to the analysis. The last part of each line attempts to describe the various discordant sibships associated with each pedigree.

The following entries were recorded during the processing of a real dataset that had various types of sibships including some that lacked meaningful data altogether.

```
  Pedigree# 1036       8       1       5 :      11x2 [ ?5 ]  1x2 [  3 ]  4x3 [ *10  6  7  8  9 ]
  Pedigree# 1038       8       2       3 :      1x2 [  3 ]   4x3 [  5 *6  7 *8 ]
  Pedigree# 1040       4       2       1 :      1x2 [ *3  4 *5 ]
```

```
Pedigree# 1045       0       0         :      1x2 [  6  7 ]   4x5 [ *2 *3 ]
Pedigree# 1062       0       0         :      1x2 [ ?4 *5 ]   3x4 [ 6 ]
```

In the example above, when looking at the pedigree, 1036, the sibship whose parents are 11 and 2 had only a single child whose status was undetermined. The sibship with parents 1 and 2 consisted only of a single unaffected individual. Therefore, only the sibship that contributed anything to the analyses were the children whose parents were 4 and 3. Notice that the contribution was 8. This is because there were 4 DSPs produced with the single affected child, 10, being replicated for each of the pairs.

For pedigrees 1045 and 1062, we find that of the two available sibships, neither were discordant and therefore contributed nothing to the analyses.

Trio generation

If a valid trio is encountered and the configuration setting, EXPAND_TRIOS is *On,* details about the trio and the generation of the virtual control are reported. The following is an example of a trio expansion report.

```
Trio found: Family ID:1336
Father:         1336x2  [0:0] Unaffected : (2 2  1 2  1 1  2 2  1 1 )
Mother:         1336x1  [0:0] Unaffected : (1 2  1 2  1 1  1 1  1 1 )
Child:          1336x3  [2:1] Affected   : (2 2  2 2  1 1  1 2  1 1 )
New Child:      1336x18 [2:1] Unaffected : (1 2  1 1  1 1  1 2  1 1 )
```

In the example above, the child and it's two parents are shown along with the virtual sibling. The numbers in the braces is just the father:mother ids.

If the parents are not present in the data file, the following will be reported:

```
Trio: 1339       Unable to create virtual sibling due to lack of parental genotypes
```

Distribution Report

The contents of the distribution report is simply the values observed from the highest scoring model for each of the N runs.

```
1 SNP models
    Test   Model
  Number     ID    Score
       0       2   3.703
       1      10   3.637
       2      10   3.402
       3      10   3.320
       4      10   3.077
       5       2   3.038
       6      10   3.000
       7       2   2.994
       8      10   2.942
       9      10   2.914
      10      10   2.899
      (Lines Removed)
```

```
2 SNP models
    Test   Model
  Number     ID    Score
       0     5x9   4.091
       1    2x10   3.939
       2     1x2   3.855
       3    1x10   3.769
       4    2x10   3.732
       5    2x10   3.657
       6    2x10   3.613
       7    1x10   3.585
       8    1x10   3.553
       9    9x10   3.543
      10    5x10   3.478
      11    1x10   3.448
      12    2x10   3.378
      13    1x10   3.354
      (Lines Removed)
```

Compiling MDR-PDT

MDR-PDT is built using the gcc (gnu c compiler) version 4.1.1 with compatibility mode set for gcc 3.2 to be compatible with most redhat distrubtions. However, if the precompiled version doesn't run properly on a particular platform, most users should no difficulty in compiling a version that will run on their machine.

Prerequisites

- **STL** – The Standard Template Library is assumed to be available on all machines with gcc 3.2 or higher installed. Under most circumstances, if it isn't installed, the administrator of the system can add the package on using the package manager associated with the distribution installed on the machine.
- **Boost** 1.33.1– Boost is an open source extension to the STL and offers a few classes used heavily by MDR-PDT. It is required to have this library available before building the application. It is generally assumed that users can install or have boost installed.

  Boost can be downloaded at: http://www.boost.org
- **Blitz** 0.9 – Blitz is an open source extension to the STL specifically designed to bring performance comparable to Fortran 77/90.

  Blitz can be downloaded at: http://www.oonumerics.org/blitz/

Compilation

Extract the files to a place in your home directory and change to that directory. In order to ensure that the old version is removed, type:

```
make clean; make
```

This will delete the version that came with the distribution and begin the compilation. If successful, the application can be found at:

```
/bin/apps/mdr-pdt
```

There is no “install” script associated with this build. Users are free to move the executable to a place of their own choosing.

Issues

There are no known issues with the application at this time. However, in the event that the program aborts unexpectedly, a debug version can be built using the following command:

```
make clean DEBUG=1; make debug
```

Configuration Parameters

Search Parameters

COMBO_START	Integer	Specifies the minimum size of models to be evaluated *Default Value:* 1
COMBO_STOP	Integer	Specifies the maximum size of models to be evaluated *Default Value:* 1
EXCLUDE_LOCUS	List of Integers	Follow the command with 1 or more loci that should not be considered during analyses. These do not shift the columnar id of subsequent snp numbers (i.e. model 10x11 is the same whether or not you excluded locus 9)
AFFECTED_VALUE	Integer	Specify the value used to represent an affected individual *Default Value:* 2
UNAFFECTED_VALUE	Integer	Specify the value used to represent an unaffected individual *Default Value:* 1

Input Format

Two formats are supported, Pedigree and DSP. However, both are analyzed the same, so reformatting data to another format will not provide any benefit. All data files should be space delimited.

INPUTFORMAT		Indicates which format the data can be found in
	PEDIGREE	This is standard pedigree format
INPUTFILE	String	Specifies the name of the file to be analyzed
PEDIGREE_EXCLUSIONS	List of Integers	Follow the command with one or more pedigree ids that should not participate in analyses. These pedigree ids are listed on overview report at the very beginning of analysis.
EXPAND_TRIOS	On/Off	When on (default), the application will attempt to create virtual unaffected siblings.
ANALYSISSTYLE		Specify which analysis method is to be applied to the dataset.
	PDT	Instructs the application to use the PDT method.
VERBOSE_EVALUATION	On/Off	When "On" the application displays each T Statistic as it is calculated during the regular run.

Reporting

Report files are created using the following naming scheme: REPORTNAME.EXT where EXT is the appropriate extension for the given report. In general there is not much need to change these settings. However, if any of the following extensions are set to STDOUT, that report will be written to standard out.

REPORT_NAME	String	Set the name of the reports that will be created during analysis. *Default Value:* Name of the configuration file.
EXT_DISTRIBUTION	String	Set the extension used for the distribution(s) used in calculating a model's significance. *Default Value:* pdist
EXT_REPORT	String	Set the extension used for the detailed report. *Default Value: STDOUT*
EXT_PEDIGREE	String	Set the extension used for the pedigree report. *Default Value: pedigree*

Permutation Tests

Permutation testing is performed to build an N Test distribution. The sibling's status is shuffled according to their immediate parental groups allowing for complex multi-generation families to be used. During a shuffle, the number of affected and unaffected siblings will not change from parental group to parental group, so the number of DSPs generated will be identical to the original.

PTEST_COUNT	Integer	The number of tests to use in the distribution. *Default Value:* 1000
PTEST_SEED	Integer	The initial seed used for seeding the status shuffle *Default Value: 1371*
PVAL_THRESHOLD	Float	In the final report, any model whose p-value falls below the PVAL_THRESHOLD will be reported in addition to a detailed description of the best model. **Note:** It is possible that the minimum value can not be reached if there weren't enough tests run. A value of 0.01 will report nothing as being significant if only 10 tests were run. *Default Value:* 0.05

Related Bibliography

Martin, E.R., Ritchie, M.D., Hahn, L.W., Kang, S., Moore, J.H. (2006) A Novel Method to Identify Gene-Gene Effects in Nuclear Families: The MDR-PDT. Genetic Epidemiology 30:111-123.

Related Web Sites

MDR-PDT at Marylyn Ritchie’s lab – http://chgr.mc.vanderbilt.edu/ritchielab/MDRPDT.html

PDT from Eden Martin at Duke – http://www.chg.duke.edu/software/index.html

MDR at Jason Moore’s lab – http://www.epistasis.org/mdr.html