Documentation of

P E D I G R E E – E X P L O R E R

(version 2.0)

P R O G R A M D E S C R I P T I O N

PEDIGREE-EXPLORER is a program package with a focus on genetic-epidemiological questions as they typically occur prior to linkage analysis. Especially, the program has been developed and optimized to

(1) analyze thoroughly pedigree structures,

(2) check genotypic marker data for consistency,

(3) create consistent marker sets, and

(4) calculate single marker likelihoods for pedigrees.

In doing so, PEDIGREE-EXPLORER can handle arbitrarily large and complex pedigrees limited just by the amount of computer memory and computation time.

The COMPLETENESS CHECK tests for accurate encoding, completeness and integrity of the pedigree- and marker data and should be run before the actual analysis. At the beginning, the input is printed out once more for control.

The PLAUSIBILITY CHECK analyses the pedigrees with respect to connectivity and coherence, particularly the existence of loops and cycles. All loops will be classified either into inbreeding or marriage loops. The PLAUSIBILITY CHECK is based only on the pedigree structure and does not use any marker data.

The GENOTYPE CHECK determines for each bi- or multi-allelic marker whether it is consistent with the Mendelian laws of inheritance or not. If any inconsistency exists, a detailed report of the most likely causes, the involved nuclear families and the critical genotypes is given. Autosomal and X-chromosomal markers are dealt with as well as pedigrees, with multiple occurrences of errors.

The LIKELIHOOD module can be used to calculate single marker likelihoods for pedigrees. In doing so, the calculation is not restricted with respect to size, i.e. how many individuals a pedigree has, or complexity, i.e. how many loops it has.

The STATISTIC module provides an overview on the included pedigrees in the data set, particularly affected trios and affected sib pairs (ASPs). Depending on the researcher's preferences - either focusing on the existence of pedigree members or on the availability of DNA - it summarizes and counts all affected units (i.e. ASPs and trios) and extracts and writes out the corresponding pedigree identifiers for further investigations.

PROGRAM SCOPE

PEDIGREE-EXPLORER is a check and test routine to verify data completeness, plausibility and consistency in the run-up of linkage analysis. It neither interprets nor judges the results.

Whether the data confront us with the incidence of incest or inbreeding (the analogues of loops), or lead us to assume man's "playing God" (through self- production, viz. self-loops) - so that one is led to assume that a new era of mankind may have started (characterized by the incidence of reincarnation cycles) - it is all the more important that the user of the data will know how to differentiate between these explanations, or whether he will merely limit himself to taking exception at possible data errors.

However, no abnormalities, excesses, or data errors will stay unrevealed and unraveled to PEDIGREE-EXPLORER, and all secrets and discrepancies in the pedigrees will be exposed.

Regardless of the size and complexity of the pedigree PEDIGREE-EXPLORER

- decomposes a data set into its connected components (i.e. pedigrees and families) and provides an overview about the affected units (i.e. trios and affected sib pairs (ASPs)) in each component,
- determines the maximal number of generations in a pedigree,
- finds the longest path from a founder to an offspring,
- finds all loops in a complex pedigree and classifies them as inbreeding and marriage loops,
- determines an approximately optimal loopbreaker set,

- verifies if a pedigree is inconsistent with the Mendelian laws of inheritance,
- generates a detailed error report specifying the most likely causes of errors,
- determines the critical genotypes and suggests consistent ones,
- automatically generates an error free marker set,
- reports the maximal possible genotype set for each pedigree member,

- calculates the maximal number of compatible genotype combinations,
- calculates single-marker likelihoods for complex pedigrees,

- deals with autosomal and X-chromosomal markers,
- deals with multiple-occurrences of errors in a single pedigree,

and last but not least

summarizes all of this, hopefully useful for somebody still swearing by, and imperturbably believing in his data (i.e. "the true-blue hero" of old).

I N P U T F O R M A T

PEDIGREE-EXPLORER assumes the input records to decode one or more pedigrees.

File format of each line in the input file (the header is optional):

(i) (ii) (iii)

Header: PED PID FID MID SEX AFF <<DNA xxx>|<M1 M2 ...>|<M1 M1 M2 M2 ... ...>>

PED : pedigree identifier --> obligatory (alphanumeric)
PID : person identifier --> obligatory (alphanumeric)
FID : father identifier --> obligatory (alphanumeric)
MID : mother identifier --> obligatory (alphanumeric)
SEX : 1 = male, 2 = female --> obligatory
AFF : 1 = unaffected, 2 = affected
--> obligatory for the extraction of affected units (trios and ASPs)

(i) Either one single column signed over with 'DNA'
DNA : 1 = available, 0 = not available
--> obligatory for the extraction of affected units (trios and ASPs),
if the extraction modus 'DNA AVAILABILITY' is chosen
xxx : further information that will be ignored

(ii) or a sequence of columns signed with a marker name containing genotypes
M1 : one genotype for marker M1, alleles separated by '/'

(iii) or a sequence of column pairs, one column pair for one marker specifying the two alleles of a genotype, both columns are signed with the same marker name

The first five fields have to be specified for each person. No blanks are allowed as input values. Missing values should be set to zero (i.e. parents of founders).
All entries have to be coded as integers except for the pedigree and person identifiers. Lines with leading characters '#', '*', or '>' will be ignored. Men's genotypes are expected to be coded as homozygote for markers on the X-chromosome.

An optional marker file can be read in to specify on which chromosome each marker is located. The first column is expected to contain the marker name and the second column the chromosome.
If the number of markers in the marker file corresponds to the number of genotypes in the PED-file the marker information will be imported in chronological order. Otherwise, the marker file is used as look-up table for each marker given in the header of the PED-file (if no header is specified the markers will be referred to by numbers in increasing order).

O U T P U T F O R M A T

By default, if no output file is specified, the whole output is written into a LOG-File in the current working directory. All clues regarding the program flow are omitted in the output file, so that it can be parsed easily.

++++++++++++++++++++ EXAMPLE 1 ++++++++++++++++++++

The output of the PLAUSIBILITY CHECK looks similar to the following lines:

```
---------- Pedigree 0001 consists of 1 connected component(s)
Component 1: PIDs --> 1 2 3 4 5 6 7 8 9 10 11 12 13 14
Component 1: Maximal number of generations: 8
Component 1: Example of longest path --> 2 > 5 > 1 > 4 > 7 > 11 > 12 > 13
Component 1: Number of self-loops: 1 --> 10
Component 1: Number of cycles: 4 --> 10 < 10, 11 < 7 < 5 < 13 < 12 < 11, ...
Component 1: Number of loops: 7 --> 2 > 4 > 7 < 5 < 2, 2 > 4 < 1 < 5 < 2, ...

Pedigree 1, Component 1 includes 2 marker(s).
```

++

++++++++++++++++++++ EXAMPLE 2 ++++++++++++++++++++

The output of the GENOTYPE CHECK looks similar to the following lines:

```
Marker M1 on chromosome 1 is inconsistent with Mendelian laws of inheritance.
> Inconsistent nuclear family:
  (U) FID 1: 0/0
  (U) MID 2: 0/0
  (T) CID 3: 2/1
  (T) CID 4: 4/3
  (T) CID 5: 5/1
  Sibship 3, 4, 5 has more than four alleles.
> Inconsistent nuclear family:
  (T) FID 6: 4/9
  (T) MID 3: 2/1
  (T) CID 7: 4/3
  PID 7 has no compatible alleles with MID 3.
> Inconsistent nuclear family:
  (U) FID 8: 0/0
  (T) MID 5: 5/1
  (T) CID 9: 8/8
  PID 9 has no compatible alleles with MID 5.
> Critical genotypes belong to PID(s) --> 3, 5; 3, 9
```

++

In case of the option '-s' a summary statistic of the affected units (ASPs and trios) is created and a bunch of files is being written out listing the corresponding pedigree identifiers dependent of the chosen extraction modus. The summary file with the ending '_STATISTIC.out' provides an overview of the affected units in all pedigrees.

O P T I O N S

-c : Test for completeness and correctness of the input data
- check for missing data and if all entries are numbers
- check for duplicated individuals
- check for individuals (non-founders) with missing parents
- check for individuals (non-founders) with twice the same parents
- check for individuals with parents of the wrong gender
- check for individuals with missing affection status or DNA indication
- check for falsely coded and missing genotypes
- check for the same number of markers for each person in a pedigree

-p : Test for plausibility of the pedigree
- check for individuals if they are their own parents (self-loops)
- check for individuals if they are their own ancestor (cycles)
- check for inbreeding and incest (loops)
- check for marriage loops
- check for the maximum number of generations in the pedigree
- check for the longest path in the pedigree
- check for the pedigrees entireness (connected components)

-g : Test for consistency with the Mendelian laws of inheritance
- check for inconsistencies between the alleles of a child and its parents
- check for inconsistencies between the alleles of all children
- check for inconsistencies between all alleles in a nuclear family
- check for inconsistencies within pedigrees without loops
- check for inconsistencies within pedigrees with loops

-li: Likelihood calculation of the pedigree for a single marker
By default the likelihood will be calculated using peeling of nuclear families. For test purposes the likelihood can additionally be calculated in a simpler way using a brute-force approach if the option -li_bf is on.

-s : Summary and statistic
- statistic about the affected units in the file depending on the chosen extraction modus
- listing and extraction of PED-IDs of the affected units

-m : Marker file specification

By default all markers are assumed to be located on autosomes unless it is specified otherwise in the marker file.

-e : Extraction modus

- EXISTENCE -->

  All members of an affected trio or ASP have to exist in the file, otherwise no unit is counted.

- DNA <DNA AVAILABILITY> -->

  An affected unit is only counted in case of DNA availability for all unit members.

- DNA_EPAR <DNA AVAILABILITY ECEPT FOR PARENTS> -->

  Units are sufficiently specified by DNA availability for affected children, but not necessarily for the parents.

- INFORMATIVE -->

  Extracts from a pedigree all uninformative persons, i.e. persons in-law without any relationship to an affected pedigree member. The input file is written new shortened by uninformative persons.

-cr: Report of the marker call rate based upon all input data, whereby it is not assumed that all pedigrees are typed by the same number of markers.

-cg: Determines the minimal sets of critical genotypes.

These are the genotypes, which, if are deleted, result in a consistent pedigree. Depending on how many genotypes have to be deleted the running time can grow up extremely.

-ex: Exhaustive graph analysis

The graph analysis does not assume anymore that all individuals descend from founders. Multi-cyclic graphs as well as graphs consisting exclusively of cycles can be handled.

-mg: Prints for each marker and pedigree the maximal possible genotype set of a person that is consistent with the Mendelian laws of inheritance.

-ped: Optional input option to pass PEDIGREE-EXPLORER a list of pedigrees to be included into the analysis. By default all pedigrees will be analyzed. The pedigree identifiers are expected to follow the option name as multiple arguments without any separator string.

-noped: Optional input option to pass PEDIGREE-EXPLORER a list of pedigrees, which have to be excluded from the analysis.

-pre: Transforms the input file into ‘PRE-file- format.

--force: Special option to enforce an interpretation of a data set even if the input data are incomplete or show discrepancies (This option may be reasonable to get a first rough overview of the data, but the results cannot be expected to be perfectly correct in any term).

-z : Creates consistent linkage files of various types. The genotypes of the markers leading to an inconsistent pedigree are blanked according to the chosen option value:
1: Blanks the genotypes for *all* persons in an inconsistent pedigree.
2: Blanks the genotypes for *all critical* persons in the pedigree.
3: Blanks the genotypes for *the most critical* persons in the pedigree.
4: Replaces the genotypes of the most critical persons with the most probable genotypes calculated by likelihood estimation.

A L G O R I T H M S

In general, a pedigree can be described by a directed graph, whereby the pedigree members are symbolized by nodes and the relationship among the individuals by edges. Thus, primarily standard algorithms from graph theory have been applied in this program. To derive the connected components the union-find technique has been applied. Topological sorting can ease tracing the shortest path in a directed acyclic graph. Equivalently the longest path can be obtained from a topological sort assuming negative weighted edges.

In directed graphs closed paths can either be loops or cycles. The difference is that in a cycle one always arrives at the starting point by following the direction of its edges, while in a loop this is impossible. Therefore it does not make sense to derive the maximal path length of cycles, because it is infinite.

Classical breadth-first search with backtracking has been used to traverse the graph and find out all loops. PEDIGREE-EXPLORER distinguishes between inbreeding and marriage loops. Whereas inbreeding loops can be detected easily in a directed graph, marriage loops can be found by treating the pedigree as an undirected graph.

The main algorithm used to find Mendelian inconsistencies is an extension of the Lange-Goradia algorithm, which is correct only for pedigrees without loops. With the help of a loopbreaker set a cyclic pedigree can be broken down into a acyclic one. It is convenient to imagine that every loopbreaker node is replaced by a proxy with the same fixed genotype for each split loop. Thus, summing up all possible genotype combinations of a loopbreaker set in a broken pedigree ensures the correctness of the Lange-Goradia algorithm for pedigrees with loops.

Finding an optimal loopbreaker set in arbitrary graphs is NP-hard. However, an approximately optimal loopbreaker set can be obtained by the adjacent nodes of the edges from the difference set between the marriage graph and its minimum spanning tree. In this case, the edge weights should be weighted proportional to the degree of the nodes and inverse proportional to the number of genotypes of the adjacent nodes.

Likelihood calculation has been implemented using the strategy of peeling nuclear families according to the Elston-Steward algorithm. Again, with the help of a loopbreaker set it is possible to deploy this strategy even in pedigrees with loops.

For questions, annotations or bug reports, please contact
steffens@imbie.meb.uni-bonn.de