Documentation for the software package "PARENTE"

Alain CERCUEIL[1,2], Eva BELLEMAIN[2] and Stéphanie MANEL[2]

[1] Laboratoire de modélisation et Calcul. IMAG, BP 53, 38041 Grenoble cedex 9
[2] Laboratoire de Biologie des Populations d'altitude, UMR CNRS 5553, Université Joseph Fourier BP53 X, 38041Grenoble Cedex09 France.

Correspondance :
Alain Cercueil
Laboratoire de modélisation et Calcul. IMAG, BP 53, 38041 Grenoble cedex 9
Tel: 33 4 76 63 58 29; Fax : (33) 4 76 63 12 63; e-mail : **alain.cercueil@imag.fr**

Contents

I. INTRODUCTION

PARENTE reads the data from a single text file containing the genotype of each individual in the sampled population (from codominant markers) and, if available, other individual characteristics (birth and death dates, sex…). For each individual, the software determines the set of potential mothers, potential fathers and potential pairs {mothers; fathers). To determine the set of potential mothers for a given individual, the software first checks for each female whether birth and death dates allow this maternity. Then, it checks for the genetic data compatibility between the individual under consideration and the "age compatible" females. Females (or individuals whose sex is unknown) that satisfy both compatibility conditions are added to the set of potential mothers. The same principle holds for potential fathers. For the pairs {mothers; fathers}, PARENTE checks the genetic and age compatibilities for all triplets {individual; potential mother, potential father}.
In each case, the program calculates the parentage probability (see part III).

II. INSTALLATION

The software package (parente.exe), a file required during a run (param.txt) and an example file, are available at the following address: http://www2.ujf-grenoble.fr/leca/membres/manel.html. You should save them on your computer in the same directory. The parameters currently defined in the file param.txt allow to run the program with the input file "example.txt".

III. DESCRIPTION

1- Compatibility of birth and death dates

An individual is considered as a potential parent of another one if :

- He/She is old enough to reproduce at the young's birth date (age of sexual maturity);
- He/She is still alive at the young's birth date (or some time before);
- He/She is old enough in comparison to the age of the young.

These differences between birth and death dates are set up in the parameter file and may differ between males and females depending on the biology of the species under consideration.

2- Genetic compatibility

- **Between a single parent and the individual:** for each locus, the individual should inherit one of his alleles from one parent. The software also checks that, for each locus, the individual and the candidate parent have at least one common allele. A locus for which the individual and the candidate parent have no common alleles is considered as an incompatibility. The number of incompatibilities between a single potential parent and the individual are counted by the software.
- **Between both parents and the individual:** for each locus, the individual should inherit one of its alleles from one parent and the other allele from the other parent. An incompatibility occurs if only one allele could have been inherited. If none of the alleles could have been inherited, two incompatibilities are counted. The number of incompatibilities between both parents and the individual is counted by the software.

In order to take into account the error rate in the data (typing errors, mutations, null alleles…), PARENTE can accept parentage with some incompatibilities. The maximum number of incompatibilities accepted by the program is a parameter.

The program calculates the probability that the parentage link is the correct one using the allelic frequencies and the sampling rate of the populations while taking account the incompatibilities and the error rate. When there are several compatible parent pairs for an individual, the user will decide which is the best one, according to those probabilities.

*3- Calculation of the probability **P** that the potential parents are the true parents*

a) Assumptions and notations

The calculation of this probability P is computed using the following assumptions:
- Wright-Fisher type reproduction,
- Hardy Weinberg equilibrium,
- independence of loci.

The calculation also requires 2 parameters:
- the error rate , τ
- the sampling rate of the studied population, q.

We consider n alleles (A_1, A_2,…,A_n) observed at one locus. Let f_i be the frequency of allele A_i in the population. The generalisation for more than one locus can be easily derived since the loci are supposed to be independent.
For a given individual, $G=(G_1,G_2)$ is the observed genotype and $g=(g_1,g_2)$ the true genotype, with $G_1,G_2,g_1,g_2 \in \{A_i\}$. These distinctions allow us to take into account error possibilities.
In the following, upper case letters will refer to observed genotypes while lower case letters refer to true genotypes.
X, y and z are a potential mother, father and an individual respectively. I(x,y,z) corresponds to the event that x and y are the true parents of z.

b) Allele frequencies

In a first step, the program estimates the allele frequencies from the observed genotypes.

c) Possibilities of error

In the next step, the program takes into account the error possibilities (in the observed genotypes) from the following model:
-with probability $1-\tau$, there is no error and the true allele is the same as the observed allele
-with probability τ, some errors have occurred, and the true allele might be any of the alleles, so this value might be chosen according to the allele frequencies.
For example, considering we have the observation $G_1=A_1$ then:
-$Pr(g_1=A_1|G_1=A_1)=1-\tau+\tau\ f_1$
-$Pr(g_1=A_i|G_1=A_1)=\tau\ f_i$ if $i\neq1$.

Furthermore, in the case of missing data for one allele, the probabilities of the true value for this allele will be chosen according to the allele frequencies, i.e. $Pr(g_1=A_i|G_1 \text{ missing})=f_i$.

With this choice the observed allele frequencies equal the true allele frequencies and Pr(g|G)=Pr(G|g).
The program calculates Pr(g|G) according to these rules.

d) Transition from parents to offspring
Let z denote an offspring, x a candidate mother and y a candidate father. E, M & F are their respective genotypes. The probability that E will be observed in the case of reproduction between x and y is:
$Pr(E|M,F,I(x,y,z))=\Sigma_{e,m,f}Pr(e|E)Pr(m|M)Pr(f|F)T(e|m,f,I(x,y,z))$
Where T(e|m,f,I(x,y,z)) is calculated using the Mendelian segregation model.
Again, upper case letters refer to observed genotypes while lower case letters refer to true genotypes.
In the case of several loci, the probability is easily obtained by the product of all locus probabilities.

e) Calculation of P
Without any additional biological information, all the individuals have the same parentage probability, i.e. P(I(x,y,z)) does not depend on the parents x and y. This means that we use a uniform prior for I(x,y,z). We calculate the posterior probability that some candidate parents are the true parents under this assumption. At this stage our software does not take into account other priors.
n_1 and n_2 are respectively the number of potential mothers and potential father in the sample.
We define $n'_1=n_1*(1-q)/q$ and $n'_2=n_2*(1-q)/q$ as the number of potential mother and father that have not been sampled.
U is the "observed" genotype of an individual that is not sampled (all missing data). For this genotype we have: $Pr(u_1=A_i|U_1)=f_i$

The posterior probability that individuals x_0 and y_0 are the true parents of z is then :
$P=P(I(x_0,y_0,z)|E_z,M_{x0},F_{y0})=P(E_z|M_{x0},F_{y0})/K$
With $K=\Sigma_{x,y}\ P(E_z|M_x,F_y,I(x,y,z))+n'_2\ \Sigma_x\ P(E_z|M_x,U,I(x,y,z))+n'_1\ \Sigma_y\ P(E_z|U,F_y,I(x,y,z))+n'_1\ n'_2\ P(E_z|U,U,I(x,y,z))$
K is also the sum over all the possible pairs, sampled or unsampled.

4- Treatment of missing data

The input file may contain some missing data (at different levels). In the program, missing data never leads to incompatibilities.

IV. THE INPUT FILE

The data need to be stored in a text file. Each line of this file contains the information for one individual and each column corresponds to a variable (e.g. age…). A given character separates the different columns (the default character is the tab key).
Two types of data are essential :
- An identifier for each individual (without any space character) in the first column.
- The genotype of each individual in the last columns.
The following columns are optional and their order can be modified by the user. These columns are:
- The sex of the individual (**M** for male, **F** for female);

- The birth date given as a single number (e.g. year);
- The death date given with the same format as birth date;
- The identifier for the mother of the individual, if known;
- The identifier for the father of the individual, if known;

Some extra columns (e.g. the subpopulation) might follow. The data contained in these columns will not be used by the program but will be indicated in the result files.
The last columns contain the genetic data (one column for one allele; one locus is represented by 2 columns).

V. THE OUTPUT FILES

The results are stored in 4 different output files (created by the program during a run).

1- The « pair.txt » file

This file contains the data concerning the triplets found {individual; mother; father}. Each line of this file corresponds to one triplet. The first column contains the identifier of the individual, followed by its birth and death dates and the other data present in the input file. Next the file gives the name of a potential mother for this individual, followed by the dates of birth and death, additional data, the number of genetic incompatibilities between the mother and the individual and the number of common alleles between the mother and the individual. Exactly the same results are then given for the father. Finally, the last columns contain the number of incompatibilities between both parents and the individual and the probability of the candidate parents being the true parents.

2- The « mother.txt » and « father.txt » files

Both files are similar to the previous one (“pair.txt”) but the information is limited to pairs {individual; mother} and {individual; father}.

3- The « without.txt » file

This file contains the list of individuals for whom no parent has been found.

VI. DESCRIPTION OF PARAMETERS

1- Generalities about parameters

The software requires some parameters that the user may modify. These parameters are stored in a file “**param.txt**” (available with the program). In this file, each parameter is preceded by the character “#” and followed by “=” and its value. Default values are indicated in bold in brackets in the following description:

2- The input and output files

This list of parameters allows the program to define the names of the input and output files.

- **Input:** name of the input file (**input.txt**).
- **Pair:** name of the file containing the results about the triplets {individual; mother; father} found (**pair.txt**).
- **Mother:** name of the file containing the results about the pairs {individual; mother} found (**mother.txt**).
- **Father:** name of the file containing the results about the pairs {individual; father} found (**father.txt**).

- **Without:** name of the file containing the list of individuals for whom no parents were found (**without.txt**)

3- Format of the input file

This list of parameters describes the input file.

- **Missing_data:** chain of characters used for the missing data. (**?)**
  - **nb_loci :** number of loci used; if a larger number is given, the program calculates the real number of loci (**50**)
  - **pop_size** : size of the population, if a larger number is given, the program calculates the real number of individuals (**1000**)
- **header**: indicates whether the first line of the input file is a header or not. 0 indicates no header, while 1 is used if a header is present (in this case, the first line is ignored). (**1**)
- **separator**: character used to separate the different columns in the input file. Use « esp » for a space character and « **tab** » for a tab key.(**tab**)
- **sexe_col**: number of the column which contains the sex of the individual. 0 if this data is not given (**2**)
- **birth_col**: number of the column which contains the date of birth of the individual. 0 if this data is not given (**3**)
- **death_col**: number of the column which contains the date of death of the individual. 0 if this data is not given (**4)**
- **mother_col**: number of the column which contains the name of the mother of the individual (if known). 0 if this data is not given (**5)**
- **father_col**: number of the column which contains the name of the father of the individual (if known). 0 if this data is not given (**0)**
- **nb_other_col**: number of extra columns before the genetic data. (**0)**

4- Parameters for the calculation

- **age_dif_M**: minimal age difference between a male and an individual to consider the male as a potential father (age of sexual maturity). This difference of age is also taken into account as the minimum age allowing the male to be considered as a father. The unit used must be the same as the one used for the dates of birth and of death , decimal numbers are allowed(**0**)
- **age_dif_F:** minimal age difference between a female and an individual to consider the female as a potential mother (age of sexual maturity). This age difference is also taken into account as the minimum age allowing the female to be considered as a mother. The unit used must be the same as the one used for the dates of birth and of death , decimal numbers are allowed (**0**)
- **age_dif2_M**: Maximal delay, between the birth of the individual and the death of the male, to consider the male as a potential father. Negative numbers are allowed. The unit used must be the same as the one used for the dates of birth and of death, decimal numbers are allowed (**0**)
- **age_dif2_F:** Maximal delay, between the birth of the individual and the death of the female, to consider the female as a potential mother. Negative numbers are allowed. The unit used must be the same as the one used for the dates of birth and of death, decimal numbers are allowed (**0**)
- **max_incomp**: maximum number of genetic incompatibilities allowed (**1**)

- **max_missing**: maximum number of genetic data missing for the individual to be taken into account. (**10**)
- **error_rate**: estimate of the error rate in the data. Different causes of errors are possible : typing error, mutation rate, null alleles. For example, the typing error rate can be estimated by repeating the typing experiment several times, or by comparing genotypes of known pairs parent-offspring. Classical values for this error rate are comprised between 1 and 5%. The programs accepts numbers between 0 and 1. (**0.01**)
- **sampling_rate:** sampling rate of the population. 1 means that all the individuals are sampled (**1**)

5- Other parameters

- **silence :** 0 if the pairs found should be displayed at screen during a run (**1**)
- **cal_prob** : 1 if you want to display the probabilities at screen (**1**)

VII. ERROR MESSAGES

1- Warnings

The following messages are some of the possible warnings. The program keeps running but the results may be erroneous.

- **param.txt : couldn't open file**

  default value used
  The program has not been able to open the parameter file and will use all the default values. This may be due to the absence of the file "param.txt" in the same directory as the executable.

- **Detected population size=**
  The program has detected a different number of individuals that the one defined in the parameters file.
- **Detected number of columns=**
  The program has detected a different number of columns then the one defined in the parameters file. This may be due to a wrong number of loci or can be a problem with the separator characters.

- **Using *x* individuals**
  If the detected population size is lower than the population size given in parameters, the program uses the lower number.

- **Using *x* loci**
  If the number of columns detected is lower than the one calculated from the parameters, the number of loci is recalculated, but the problem may come from the labels of the other columns. In this case, the results (if produced) may be wrong.

2- Errors

The following messages are displayed at screen with some errors that terminate the program prematurely.

- **Couldn't open input file**
  The program has not been able to open the input file. If this file is not in the same directory as the executable, its name must be given with its full path.

- **Bad list of columns**
  There is some inconsistency in the list of columns e.g. a label of column is not used within this list.

VIII. EXAMPLE

A simulated example has been created with known parentage relationships. We then applied PARENTAGE to compute these relationships.

1- Example of an input file

The input file, called "example.txt" is http://www2.ujf-grenoble.fr/leca/membres/manel.html
It contains 9 individuals characterized at 4 loci.

	Sex	Year of birth	Year of death	Mother	A1	A2	B1	B2	C1	C2	D1	D2
Ind1	M	1985	2000	Ind2	102	104	136	140	211	211	155	157
Ind2	F	1977	?	?	100	102	136	136	211	215	?	159
Ind3	F	2000	?	?	100	104	136	136	211	211	157	159
Ind4	M	2000	?	?	100	102	136	140	?	215	157	157
Ind5	F	1985	1997	?	104	104	140	142	211	213	155	159
Ind6	M	1978	1996	?	104	108	?	144	211	213	155	155
Ind7	M	?	2000	?	104	104	140	144	211	211	155	159
Ind8	M	1984	2000	Ind2	102	106	136	138	209	211	155	157
Ind9	?	?	?	Ind5	104	108	142	144	211	213	155	155

2- Description of the file « param.txt »:

```
**********************************
*Output and input files *
**********************************
# input = example.txt
# pair = pair.txt
# without   = without.txt
# mother  = mother.txt
# father = father.txt

**********************************
*Size of data            *
**********************************
# missing_data = ?
# nb_loci      = 4
# pop_size     = 9
# header       = 1
# separator    = tab
# sexe_col     = 2
# birth_col    = 3
# death_col    = 4
```

```
# mother_col   = 5
# father_col   = 0
# nb_other_col = 0

************************************
*parameters used during calculation *
************************************
# age_dif_M      = 3
# age_dif_F      = 3
# age_dif2_M     = 0
# age_dif2_F     = 0
# max_incomp     = 1
# max_missing    = 5
# error_rate     = 0.01
# sampling_rate  = 0,7
# silence  = 1
# cal_prob = 1
```

3- Description of the results files

On the screen, you can read :
3 females
5 males
1 individual of unknown sex
total : 9 individuals used
18 pairs have been found
completed

a) Parents file (pair.txt) :

Ind3	2000	-1	Ind2	1977	-1	0	5	Ind1	1985	2000	0	5	0	0.4971
Ind3	2000	-1	Ind2	1977	-1	0	5	Ind7	-1	2000	1	5	1	0.0003071
Ind3	2000	-1	Ind2	1977	-1	0	5	Ind8	1984	2000	1	3	1	0.002927
Ind5	1985	1997	Ind6	1978	1996	0	4	Ind9	-1	-1	0	5	1	4.55E-05
Ind5	1985	1997	Ind7	-1	2000	0	6	Ind9	-1	-1	0	5	0	0.7301
Ind1	1985	2000	Ind2	1977	-1	0	3	Ind6	1978	1996	0	3	0	5.19E-05
Ind1	1985	2000	Ind2	1977	-1	0	3	Ind7	-1	2000	0	6	0	0.01579
Ind1	1985	2000	Ind2	1977	-1	0	3	Ind9	-1	-1	1	3	1	6.13E-05
Ind4	2000	-1	Ind2	1977	-1	0	4	Ind1	1985	2000	0	4	0	0.1007
Ind4	2000	-1	Ind2	1977	-1	0	4	Ind8	1984	2000	0	3	1	0.0006484
Ind6	1978	1996	Ind7	-1	2000	0	4	Ind9	-1	-1	0	7	0	0.475
Ind7	-1	2000	Ind5	1985	1997	0	6	Ind1	1985	2000	0	5	1	0.003789
Ind7	-1	2000	Ind5	1985	1997	0	6	Ind6	1978	1996	0	4	0	0.253
Ind7	-1	2000	Ind5	1985	1997	0	6	Ind9	-1	-1	0	4	0	0.2533
Ind7	-1	2000	Ind1	1985	2000	0	5	Ind9	-1	-1	0	4	1	0.003747
Ind7	-1	2000	Ind6	1978	1996	0	4	Ind9	-1	-1	0	4	1	3.00E-05
Ind9	-1	-1	Ind5	1985	1997	0	5	Ind6	1978	1996	0	7	0	0.7573
Ind9	-1	-1	Ind5	1985	1997	0	5	Ind7	-1	2000	0	4	1	0.005132

In this file missing data are represented by “-1”. The first line corresponds to the individual 3. He was born in 2000 and his death date is unknown. The mother of the individual 3 is the

individual 2 (no incompatibility between their genotypes and 5 common alleles). 3 fathers are possible with mother 2:
- individual 1 is a potential father (no incompatibility with the individual 3 and 5 common alleles),
- individual 7 (1 incompatibility between their 2 genotypes and 5 common alleles),
- individual 8 (1 incompatibility and 3 common alleles)

But we conclude that the first parent-pair (ind 2-7) is the right one since it has a probability of about 50% of being the right pair (taking into account the error rate, sampling rate and allelic frequencies). The probabilities of the two other couples are too weak (<1%).

All the other files could be interpreted in the same way.

b) Mothers file (mother.txt) :

Ind3	2000	-1	Ind2	1977	-1	0	5	0.8049
Ind5	1985	1997	Ind9	-1	-1	0	5	0.1061
Ind1	1985	2000	Ind2	1977	-1	0	3	0.02257
Ind1	1985	2000	Ind9	-1	-1	1	3	0.003349
Ind4	2000	-1	Ind2	1977	-1	0	4	0.1416
Ind6	1978	1996	Ind9	-1	-1	0	7	0.2944
Ind7	-1	2000	Ind5	1985	1997	0	6	0.6704
Ind7	-1	2000	Ind9	-1	-1	0	4	0.07186
Ind8	1984	2000	Ind2	1977	-1	0	3	0.005797
Ind9	-1	-1	Ind5	1985	1997	0	5	0.8258

c) Fathers file (father.txt) :

Ind3	2000	-1	unknown	Ind1	1985	2000	0	5	0.5545
Ind3	2000	-1	unknown	Ind7	-1	2000	1	5	0.002603
Ind3	2000	-1	unknown	Ind8	1984	2000	1	3	0.003715
Ind5	1985	1997	unknown	Ind6	1978	1996	0	4	0.000594
Ind5	1985	1997	unknown	Ind7	-1	2000	0	6	0.05269
Ind5	1985	1997	unknown	Ind9	-1	-1	0	5	0.8097
Ind1	1985	2000	Ind2	Ind6	1978	1996	0	3	0.0023
Ind1	1985	2000	Ind2	Ind7	-1	2000	0	6	0.6997
Ind1	1985	2000	Ind2	Ind9	-1	-1	1	3	0.002716
Ind4	2000	-1	unknown	Ind1	1985	2000	0	4	0.4993
Ind4	2000	-1	unknown	Ind8	1984	2000	0	3	0.08144
Ind6	1978	1996	unknown	Ind7	-1	2000	0	4	0.01912
Ind6	1978	1996	unknown	Ind9	-1	-1	0	7	0.6712
Ind7	-1	2000	unknown	Ind1	1985	2000	0	5	0.04797
Ind7	-1	2000	unknown	Ind6	1978	1996	0	4	0.3104
Ind7	-1	2000	unknown	Ind9	-1	-1	0	4	0.3146
Ind9	-1	-1	Ind5	Ind6	1978	1996	0	7	0.9171
Ind9	-1	-1	Ind5	Ind7	-1	2000	0	4	0.006214

This file takes into account an additional column: the mother column (indicating whether she is known or not) and the father is looking for taking into consideration this information.

d) Individuals for whom no parents have been found (without.txt) :

Ind2