MOL_COANC

Version 3.0

26th of June 2012

Jesús Fernández, David Cros and Miguel Ángel Toro

Introduction

This programme is intended to estimate the genealogical coancestry between a group of contemporaneous individuals (i.e. none of them can be ancestor of other individual in the same set) from the genotype for a number of codominant markers, following the approach presented in Fernández & Toro (2006).

The underlying idea is constructing a pedigree by creating 'virtual' ancestors of the target (genotyped) individuals which maximises the correlation between the molecular coancestry matrix calculated from markers and the genealogical coancestry obtained for that particular pedigree. As stated in the original paper, this method has several advantages over other coancestry estimators: i) as the method implies explicit reconstruction of a genealogy, it always yields congruent coancestry matrices (in opposition with pairwise methods); ii) it needs no assumptions about the allelic frequencies of the used markers and no previous knowledge of the population structure (e.g. only FS families exist); iii) it allows for complex relationships between individuals as it is able to construct 'deeper' genealogies.

Apart from the genealogical congruence, the software also provides solutions that are feasible from the molecular point of view. Proposed full-sib families are checked to make sure they comply with the Mendelian rules of allele transmission (i.e. no more than four alleles in the same family, one allele can exists in heterozygosity just with two different alleles, ...). The way of coping with genotyping errors is the possibility of relaxing the conditions of compatibility within a FS family. If there are *L* genotyped loci we can make the program to accept families with incompatibilities in *n* loci (being *n* the tolerance). This way, individuals with genotyping errors can be correctly included in its family (but wrong individuals can be included in the family too!!; be careful with the number of 'tolerant' loci).

The software allows for the possibility of including previous knowledge about the relationship (or genealogy) between some of the individuals. For example: in some situations we know the mother for the individuals (seeds grouped in the same fruit, eggs laid together, etc) so we have only to infer the fathers for each one. Program will fix any known relationship and will force solutions to fit with that partially known genealogy. The previous information can be related to genotyped individuals as well as to 'virtual' individuals if our knowledge of the population allows for (e.g. we know a group of individuals comes from a brother x sister line).

The searching engine/optimisation procedure the software uses for the determination of the pedigree with the highest correlation between coancestry matrices is a *simulated annealing* algorithm. Basic information on the characteristics of this kind of optimisation methods can be found, for example, in Kirpatrick et al. (1983) Science 220**:** 671-680.

The aim of the software is not to recover the real pedigree which generated the present population (notice, for example, that generations are discrete across the whole genealogy) but providing a pedigree that is compatible with the observed diversity and the molecular relationship between genotyped individuals. Therefore, the estimated genealogy could be used to calculate historical effective population sizes, for example.

New features

- Main novelty of this release is the possibility of dealing with hermaphroditic and monoecious species. Consequently, we can force the software to create ancestors with separate sexes (i.e. they permanently act as males or females, as in previous versions) or allow for any kind of mating design, assuming they can act as males or females each time they mate.
- In the case of non separated sexes it is also possible to allow or not for selfings, depending on the physiology of the species.
- Now the maximum number of ancestors per generation has not to be constant but can be define separately. This way the previous information on the demographic history of the population can be accounted for (e.g. number of founders, known bottlenecks or expansions) and more accurate solutions can be found. Notice that a larger number of possible ancestors will (unnecessarily) enlarge the space of feasible solutions reducing the efficiency of the optimisation algorithm.
- It is possible to establish a predefined coancestry matrix between founders. This feature permits to account for a real knowledge of the origin of the oldest known ancestors or to compare different hypothesis about the foundation of the population (i.e. from related or unrelated individuals).

Input files

Besides the genotypes of the tested individuals, algorithm needs some other parameters to work that should be entered using several files. Programme will ask you for a name (*user_defined_input* from here) for the file with the genotypes, and (optionally) for a name for a predefined coancestry matrix (*user_defined_coancestry_matrix*) and predefined known relationships (*user_defined_known_genealogy*).

- [*user_defined_input*]
Main data file. Text file with data separated by blanks (at least one), tabs or commas. Program reads data in free format, so there is no need of genotypes to be in columns (although it looks nicer). Alleles should be code for each locus separately correlatively from 1 to the number of alleles (a companion program is supplied to recode from other kind of numeric values, e.g no. repeats, fragment length, etc).

`10` ⇒ *seed for random number generator (integer)*

`200` ⇒ *no. (genotyped) individuals to estimate relationships between them (N)*

`11 5 1` ⇒ *no. of loci (L, codominant), no. loci to use for compatibility and tolerance for families' compatibility, no. of incompatibilities admitted (tolerance)*

`13 14 16 7 5 20 24 13 13 10 12` ⇒ *no. alleles per locus*

`2` ⇒ *no. previous generations to be simulated*

`0` ⇒ *comparison matrix calculated from markers (0) or taken from file (1)*

`1` ⇒ *partially known genealogy (1) or completely unknown (0)*

`0` ⇒ *given coancestry between founders (1; NB for separated sexes: male must be placed first) or unrelated founders (0)*

`TRUE FALSE 1` ⇒ *individuals both male AND female (TRUE/FALSE), selfing (TRUE/FALSE) and 1st generation with selfing allowed (a number must be placed even if selfing is not allowed (in this case type any number))*

`10 20` ⇒ *no. of virtual individuals per generation. If separate sexes have been indicated then two lines have to be included, one with the number of males per generation and other with the number of females.*

```
  5  5  4  6 14 14  4  5  1  4 18 19  1  2  7 12  3  3  4  4  4  4
```

⇒ *genotypes with one individual per row and two columns per marker (first allele first locus, second allele first locus, first allele second locus, second allele second locus, ...)*

```
  5  5  7  7 14 14  5  5  3  4  0  0  1 23  6  7  3  7  4  4  3  8
  7  8  6  8  2 14  5  5  3  4 16 19  1  9  9 12  3  5  4  4  4  8
  6  8  2  8  3  4  5  5  1  2  3 18 10 12  5 12  3  3  4  5  5  5
  3  6  7  8  4  9  5  5  1  1 18 20  1  5  8 10  7  7  4  5  3 11
  9 11  8 12  9 10  4  5  1  3  2 18  4 19  1  6  3  3  3  4  2 10
  5  5  7 12  9 14  4  5  1  3 17 19  1 10  6 12  3  3  5  9  3  8
  8 10  7 10  5 14  2  5  1  2  2 19  1 24  6  7  3 11  4  5  4  8
  5 10  4  7  5  6  5  7  1  4 12 19 10 17  6  7  3  3  7  9  3  4
  5  9  4  8  3  4  4  5  1  3  2 16 14 19  5  6  3  5  3  7  2  4
  7 12  8  8  3  4  5  7  1  1  2 15  1 16  6  7  3  7  3  7  2  3
  6  9  2  9  2  4  5  6  1  1  2  3  1  9  6  9  3  3  3  4  3 10
  5  5  8 12  3  3  4  5  1  3 17 18  4 16  7 12  3  5  4  7  3  4
  5  6  2  9  9 12  4  6  1  3  3  5  2 12  5 13  5  6  3  8  6  8
  3  8  7  8  4 14  1  5  1  2  7 20  4  9  5 12  7  7  5  9  3  5
  5  8  8 10  9 13  5  5  1  1 12 18 13 19  6 12  7 10  2  3  3  9
```

Missing values should be coded as 0 (zero).

- anneal_param.txt

Text file including the control parameters of the simulated annealing algorithm. You can not change the name of the file; therefore, if you are using different sets of parameters, remember to rename the files before using them. Meaning of each value is:

`200` ⇒ *maximum no. of allowed steps (temperatures)*
`5000` ⇒ *no. of solutions tested in each step*
`.01` ⇒ *initial temperature*
`.9` ⇒ *rate of decrease in temperature (i.e. $T_{t+1} = T_t * 0.9$)*

In order to get accuracy and speed of computation, some previous runs may be performed to find the optimal values for the particular problem dealt with.

- [*user_defined_coancestry_matrix*] (optional file)

Text file with the coancestry matrix to be used as comparison instead of the raw molecular coancestry. Full matrix must be entered (i.e. coancestry between *i* and *j* and between *j* and *i* are to be provided). Therefore, file should have as many rows as the number of genotyped individuals with the same number of values in each.

- [*user_defined_founders_coancestry_matrix*] (optional file)

Text file with the coancestry matrix between the oldest virtual ancestors (i.e. founders of the genealogy). Full matrix must be entered (i.e. coancestry between *i* and *j* and between *j* and *i* are to be provided). Therefore, file should have as many rows as the number of ancestors (whatever males, females or hermaphrodites) with the same number of values in each.

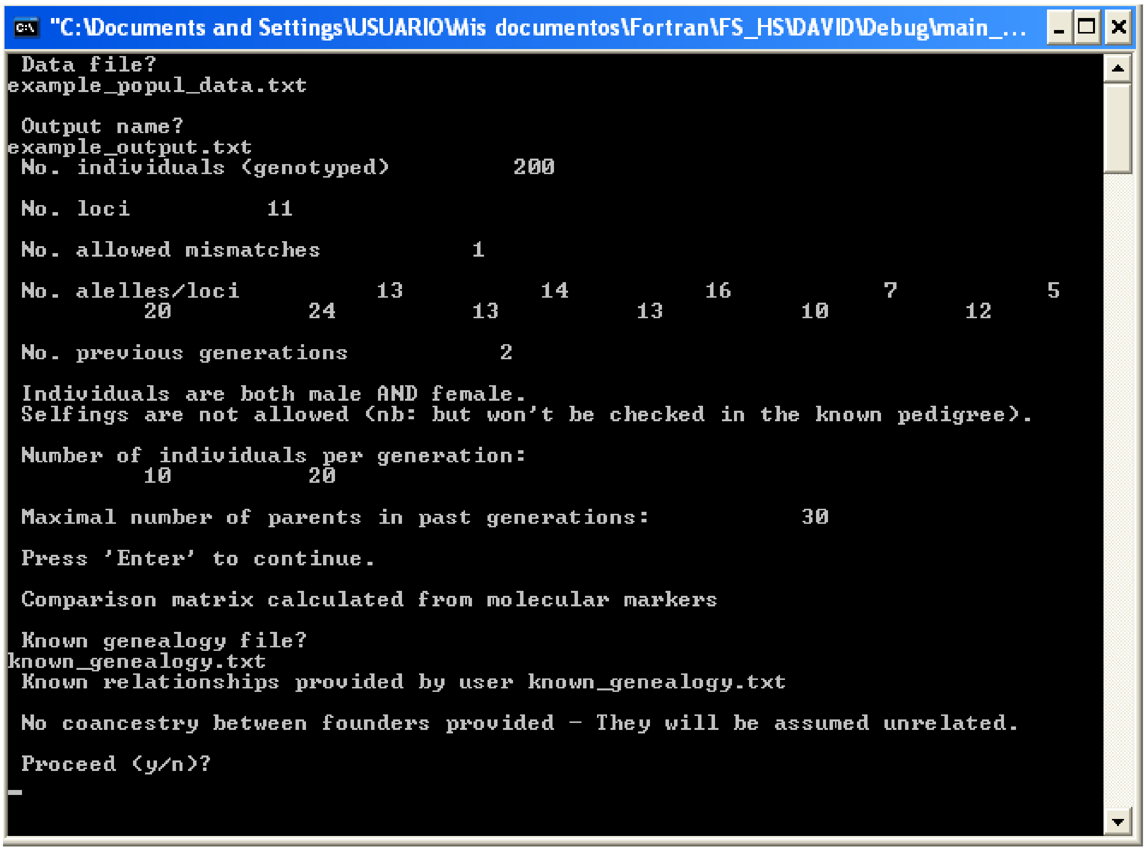
- [*user_defined_known_genealogy*] (optional file)

A full genealogy must be provided, including all virtual parents from any previous generation, placing in any generation males before females if no monoecious is assumed (be careful with congruencies so only males appear as fathers and only females as mothers). Last *N* individuals are the ones with genotype. Format is the usual for genealogies with three columns: individual, father and mother.

Individual ID	*Parent1 ID*	*Parent2 ID*
⇓	⇓	⇓
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	0	0
9	0	0
⋮	⋮	⋮
27	0	0
28	0	0
29	0	0
30	0	0
31	26	28
32	26	27
33	26	29
34	0	0
35	24	28
36	28	0
37	30	27
38	26	27
39	25	27
40	28	26
41	28	27
42	0	0
43	0	0
44	0	0
45	0	0
⋮	⋮	⋮

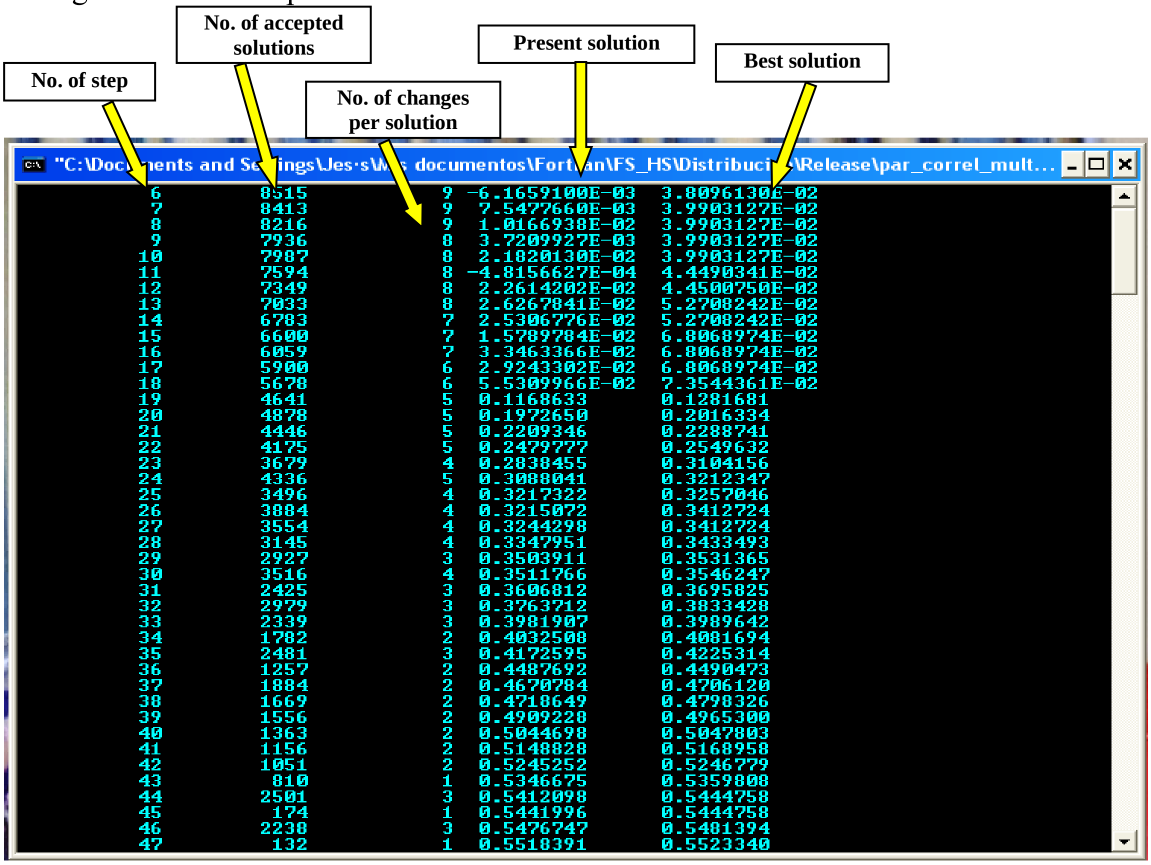
For any type of individual (virtual or genotyped) zero means that ancestor is unknown. Father, mother or both parents can be fixed. Solutions obtained from the program will contain the previously known relationships.

Before starting the calculations the programme will show the read parameters and will ask for confirmation.


"C:\Documents and Settings\USUARIO\Mis documentos\Fortran\FS_HS\DAVID\Debug\main_...

```
 Data file?
example_popul_data.txt

 Output name?
example_output.txt
 No. individuals (genotyped)          200

 No. loci            11

 No. allowed mismatches               1

 No. alelles/loci          13          14          13          16          10           7          12           5
          20          24          13          13          10          12

 No. previous generations               2

 Individuals are both male AND female.
 Selfings are not allowed (nb: but won't be checked in the known pedigree).

 Number of individuals per generation:
          10          20

 Maximal number of parents in past generations:             30

 Press 'Enter' to continue.

 Comparison matrix calculated from molecular markers

 Known genealogy file?
known_genealogy.txt
 Known relationships provided by user known_genealogy.txt

 No coancestry between founders provided - They will be assumed unrelated.

 Proceed (y/n)?
_
```

During the execution of the *simulated annealing* algorithm the programme will show some information to check if the process is running correctly and to allow for the 'fine tuning' of the control parameters. Screen will look like this


No. of accepted solutions
Present solution
Best solution
No. of step
No. of changes per solution
"C:\Doc ents and Se ings\Jes·s\M s documentos\Fort an\FS_HS\Distribuci \Release\par_correl_mult...

```
           6          8515           9 -6.1659100E-03   3.8096130E-02
           7          8413           9  7.5477660E-03   3.9903127E-02
           8          8216           9  1.0166938E-02   3.9903127E-02
           9          7936           8  3.7209927E-03   3.9903127E-02
          10          7987           8  2.1820130E-02   3.9903127E-02
          11          7594           8 -4.8156627E-04   4.4490341E-02
          12          7349           8  2.2614202E-02   4.4500750E-02
          13          7033           8  2.6267841E-02   5.2708242E-02
          14          6783           7  2.5306776E-02   5.2708242E-02
          15          6600           7  1.5789784E-02   6.8068974E-02
          16          6059           7  3.3463366E-02   6.8068974E-02
          17          5900           6  2.9243302E-02   6.8068974E-02
          18          5678           6  5.5309966E-02   7.3544361E-02
          19          4641           5  0.1168633       0.1281681
          20          4878           5  0.1972650       0.2016334
          21          4446           5  0.2209346       0.2288741
          22          4175           5  0.2479777       0.2549632
          23          3679           4  0.2838455       0.3104156
          24          4336           5  0.3088041       0.3212347
          25          3496           4  0.3217322       0.3257046
          26          3884           4  0.3215072       0.3412724
          27          3554           4  0.3244298       0.3412724
          28          3145           4  0.3347951       0.3433493
          29          2927           3  0.3503911       0.3531365
          30          3516           4  0.3511766       0.3546247
          31          2425           3  0.3606812       0.3695825
          32          2979           3  0.3763712       0.3833428
          33          2339           3  0.3981907       0.3989642
          34          1782           2  0.4032508       0.4081694
          35          2481           3  0.4172595       0.4225314
          36          1257           2  0.4487692       0.4490473
          37          1884           2  0.4670784       0.4706120
          38          1669           2  0.4718649       0.4798326
          39          1556           2  0.4909228       0.4965300
          40          1363           2  0.5044698       0.5047803
          41          1156           2  0.5148828       0.5168958
          42          1051           2  0.5245252       0.5246779
          43           810           1  0.5346675       0.5359808
          44          2501           3  0.5412098       0.5444758
          45           174           1  0.5441996       0.5444758
          46          2238           3  0.5476747       0.5481394
          47           132           1  0.5518391       0.5523340
```

For example, too rapid decrease in the number of accepted solutions (and the correspondingly number of changes per new solution) may indicate a too low initial temperature, so algorithm is stuck in the starting area. It could be also advisable to run the program with different starting points (i.e. different seeds for the random number generator).

Output files

Programme will ask you for an output name (*user_defined_output*) and several files will be created starting with such name.

- [*user_defined_output*_**mol_coanc**.txt] (optional)
If user has chosen to calculate the molecular coancestry matrix from data (option 0) this will be written in full format (*N* rows with *N* columns) to a file.

- [*user_defined_output*_**est_coanc**.txt]
This file will contain the full format coancestry matrix calculated from the estimated genealogy. As explained in the original paper, this matrix is the one with the highest correlation with the molecular coancestry matrix or the one defined as input. Only relationships between genotyped (real) individuals appears.

- [*user_defined_output*_**est_geneal**.txt]
This file will show the estimated genealogy used to calculate the coancestry matrix in the previous file.

```
200  0.5820658  ⇒ no. of (genotyped) individuals and correlation between matrices
        Individual ID    Parent1 ID       Parent2 ID
             ⇓              ⇓                 ⇓
             1              0                 0
             2              0                 0
             3              0                 0
             4              0                 0
             5              0                 0
             6              0                 0
             7              0                 0
             8              0                 0
             9              0                 0

             ⋮              ⋮                 ⋮

           397              0                 0
           398              0                 0
           399              0                 0
           400              0                 0
           401            138               219
           402            138               219
           403             54               273
           404             31               331
           405            171               214
           406            103               222
           407            161               219
           408             93               363
           409          124          219
           410          103          289

            ⋮            ⋮            ⋮
```

Note that full genealogy will be shown, including all virtual parents from any previous generation (for you to draw the genealogy or to calculate again the coancestries). Last *N* individuals are the ones with genotype. IDs of parents are not absolute codes, i.e.: if you run the same problem using different seeds for the random number generator you will surely get different parents for the same individual. The important fact is if two individuals share or not parents (irrespectively if they are called 1, 27 or 354).

Graphic output

The software has been connected to the pedigree drawing software **Pedigraph** (Garbe, J. R. and Y. Da., 2008. Pedigraph user manual Version 2.4. Department of Animal Science, University of Minnesota). Therefore, the required input files are created and the program automatically called (it should be installed in the same folder as the Mol_coanc). A .jpg file with the final pedigree will be created. Pedigraph can be downloaded from http://animalgene.umn.edu/pedigraph/.

Known limitations

The limit in the size of the problem is related with the amount of memory available in your computer. Memory management is not very efficient in this version of the programme (we will try to improve this point in future releases). Notwithstanding, we have estimated coancestries for a group of 560 individuals assuming a single generation and 1120 available virtual parents in a workstation with 2 Gb of user available RAM. A useful tip could be to reduce the number of available parents to realistic figures (is quite strange to find an example with all individuals coming from completely different parents); this will save memory and will reduce the space of solutions to explore.

If you have any particular problem when using the program please, contact with the authors in the e-mail jmj@inia.es. If you discover any bug or you would like any feature to be included in future versions, your comments will be very welcome.

How to cite this programme

Fernández, J. and Toro, M. A. 2006. A new method to estimate relatedness from molecular markers. **Molecular Ecology** 15**:** 1657–1667.