D I P L O G E N - User's Manual

Qualitative inheritance analysis of zymograms and DNA electropherograms in diploid individuals

Developed for free distribution by: Elizabeth M. Gillet
Inst. Forstgenetik, Universitaet Goettingen
Buesgenweg 2, D-37077 Goettingen; egillet@gwdg.de

September 1998

Given the banding patterns of the zymograms or DNA electropherograms of a genetically closed sample of diploid individuals, <DIPLOGEN> systematically generates all hypotheses for the mode of inheritance of these patterns that conform to certain qualitative rules for the genetic interpretation of single bands. Both dominance and codominance as modes of gene action can be detected. These rules follow from formulation of the concept of << TRANSMISSION HOMOLOGY >> within single individuals and sets of individuals (Gillet 1996).

This manual is published and <DIPLOGEN> is available on the internet under URL:
http://www.uni-forst.gwdg.de/forst/fg/index.htm

CONTENTS:

1. INHERITANCE ANALYSIS

The purpose of inheritance analysis of a genetic trait is to determine the << MODE OF INHERITANCE >> of the trait expressions. The two components of a mode of inheritance are the

(1) << MODE OF TRANSMISSION >> : number of loci, identification of the alleles at each locus;

(2) << MODE OF GENE ACTION >> : intra- and interlocus interactions between alleles (dominance, codominance, epistasis).

Where obtainable, progenies of controlled crosses or self-fertilization are used to infer mode of inheritance. In forest tree species, for example, this is often infeasible, so that other methods must be sought. <DIPLOGEN> systematically generates all possible hypotheses for mode of inheritance of isoenzyme or DNA-banding patterns that can explain the given set of banding patterns, under the assumption of their "genetic closure" (see below).

-- homomeric: the product of two genes of the same type at one locus;
-- intralocus heteromeric: consists of subunits encoded by two genes of different types at one locus (polymeric enzyme systems only);
-- interlocus heteromeric: consists of subunits encoded by the genes (of different types) at more than one locus (polymeric enzyme systems only);
-- post-translational modification (PTM):

Specification of the mode of inheritance involves identification of the origin of each band as a molecule consisting of how many subunits coded by which alleles at which loci. In some cases, even the absence of a band must be interpreted as the presence of a "null allele" (that produces a defective subunit) at some locus.

For DNA banding patterns, the number of bands per pattern can be very much larger than for isoenzymes. Other than the fact that genetic closure may require larger samples, their inheritance analysis is, however, no different than for a monomeric enzyme system allowing for "null alleles" but without PTM.

2. SAMPLING STRATEGIES

GENETIC CLOSURE

<DIPLOGEN> ideally requires as input the banding patterns of a genetically closed sample of individual banding patterns, which is explained as follows. Assuming complete genetic control of the banding patterns, each pattern is the expression of an individual's diploid genotype at the controlling loci. A sample of banding patterns is defined to be << GENETICALLY CLOSED >>, if it contains all possible patterns that can result as combinations of the genes in the sample. Thus all constructible homo- and heterozygote genotypes must be present at each locus as well as all possible interlocus combinations of these single-locus genotypes. The genetic closure of a sample can only be judged retrospectively, i.e., after inheritance analysis has been successful in identifying loci and alleles. Nevertheless, sampling strategies can be devised to increase the chances of obtaining a genetically closed sample.

SUFFICIENT SAMPLE SIZE

Given a desired probability for genetic closure of a sample, the sufficient sample size is a function of the number of genotypes and the frequency of the rarest genotype among the total collection of individuals from which a random sample is drawn. If these can be estimated, the sufficient (or minimum) sample size required to ensure with the given probability that all genotypes are detected can be calculated after Gregorius (1980). The frequency of the rarest genotype depends on the manner of association of alleles in the genotypes at each locus as well as on the association of the genotypes between loci, information which is of course not available at the outset of the study. The sufficient sample size increases for decreasing minimum genotype frequency.

SAMPLING PROGENY FROM SELF-FERTILIZATION OF A SINGLE INDIVIDUAL

Genetic closure is easiest to achieve by sampling progeny from self-fertilization of a single individual, where possible. The number of genotypes and the expected frequency of the rarest genotype depend on several unknown quantities: the number of loci at which the individual is heterozygous, the mode of gene action (codominance vs. dominance/recessiveness) between the two alleles at a heterozygous locus, the segregation proportions at each of these loci, and the recombination frequencies between these loci.

If the individual is heterozygous at m of the loci that produce the banding pattern, then the frequency of the rarest genotype is maximal, if segregation at each heterozgous locus is regular (1:1), the alleles of the different loci are randomly associated among the individual's gametes, and gametic fusion is also at random (implying random association of genotypes between loci among the progeny). In this case, the rarest genotype at each locus has the expected frequency 0.25, regardless of whether the mode of gene action is codominance or dominance. The rarest multilocus genotypes are then those that are homozygous at all loci, each of which has the expected frequency (0.25)^m. However, the minimum sample size is largest for codominance at all loci, since the number of constructible genotypes is greatest. In practice, the expected frequency must be estimated by assuming limits on segregation distortion and recombination fractions based on information gained from other systems. In general, the wider these limits are allowed to be, the larger will be the sufficient sample size.

If the above ideal conditions can be assumed for the alleles at all loci controlling the banding patterns in the parents, and the mode of gene action is codominance at all loci, then Table 1 gives sufficient sample sizes to ensure a given probability of genetic closure of a sample of the individual's progeny from self-fertilization.

TABLE 1: For sampling of progeny produced by a single individual by self-fertilization, minimum sample size is given that ensures a given probability of genetic closure of the sample under the following assumptions:
(1) the parent is heterozygous at m of the loci that control the banding pattern;
(2) segregation of the alleles at each of the m loci is regular (1:1);
(3) the genotypes at the different loci show random association among the progeny;
(4) each locus shows codominance of gene action.

Parent heterozygous at m loci	No. of genotypes in progeny	Frequency of rarest genotype in progeny =(0.25)^m	Minimum sample size such that probability of genetic closure is greater than 80%	90%	95%
1	3	0.500000	9	13	19
2	9	0.062500	57	79	104
3	27	0.015625	304	396	500
4	81	0.003906	1504	1879	2297

SAMPLING INDIVIDUALS IN POPULATIONS

Collections of individuals from large natural populations may be genetically closed. The frequency of the rarest genotype depends on the frequency distribution of multilocus genotypes among the parents of this population, the gametic phases in linkage groups in each parent, the individual gamete production (fecundity), gametic selection, and of course viability selection.

If the sampled population can be assumed to be genetically closed, and if it is possible to estimate the frequency of the rarest genotype in the population, then Table 2 gives sufficient sample sizes to ensure a given probability of genetic closure. The number of genotypes is taken to be the maximal number 1/(frequency of rarest genotype).

SEQUENTIAL SAMPLING OF INDIVIDUALS

Since sufficient sample sizes are rarely exactly calculable, a sequential sampling scheme among individuals with the potential for genetic closure may be most appropriate: sampling continues until <DIPLOGEN> succeeds in finding a hypothesis.

TABLE 2: For sampling individuals in a large population, minimum sample size to ensure a given probability of detecting all genotypes that are present at relative frequencies not less than a given minimum frequency is given. The number of genotypes actually present in the population is assumed to equal 1/(minimum genotype frequency). Word of warning: A sample of sufficient size to detect all genotypes can, however, only be genetically closed if the sampled population itself is genetically closed.

To detect all genotypes that have frequency not less than	Minimum sample size such that probability of detection of all such genotypes is greater than		
	95%	99%	99.9%
0.500	6	8	11
0.400	7	10	14
0.300	11	15	22
0.200	21	28	39
0.100	51	66	88
0.090	57	74	99
0.080	65	84	112
0.070	77	99	131
0.060	92	119	156
0.050	117	149	194
0.040	152	192	249
0.030	212	265	341
0.020	341	422	536
0.010	754	916	1146
0.009	850	1030	1285
0.008	972	1174	1462

After Gregorius (1980).

3. ELEMENTARY ZONES

Given a sample of banding patterns, the path of migration of bands is divided into << ELEMENTARY ZONES >>, abbreviated << EZONE >> in <DIPLOGEN>, such that

(1) each elementary zone contains a band of at least one banding pattern;
(2) any two bands of different patterns that appear in this elementary zone are considered to represent the "same" band (in general, identical isoenzymes or DNA fragments).

Qualitative inheritance analysis of the banding patterns consists in interpretation of the patterns of band appearance in the elementary zones.

For this purpose, elementary zones are classified into the following types:

An elementary zone is << FIXED >>, if a band appears in this zone in all of the patterns. The lack of variation in a fixed zone prohibits its interpretation.

A non-fixed elementary zone is << DEPENDENT >> on a second non-fixed elementary zone, if whenever a band appears in the one zone of any pattern, a band is also present in the second zone. A zone can be dependent on more than one zone (besides itself).

A non-fixed elementary zone i is << INDEPENDENT >>, if it is not dependent on any other elementary zone, i.e., if for each other non-fixed elementary zone j, there exists a banding pattern that exhibits a band in i but no band in j.

Two non-fixed elementary zones are << EQUIVALENT >>, if each zone is dependent on the other, i.e., if in every banding pattern bands appear either in both zones or in neither zone. The relation "equivalence", denoted "~", partitions the set of elementary zones into << EQUIVALENCE CLASSES >> of elementary zones, since it is reflexive (Z~Z), symmetric (Z~Y==>Y~Z), and transitive (Z~Y and Y~X ==> Z~X).

4. ZYMOGRAMS

Isoenzymes are defined as "electrophoretically separable variants of one enzyme ... system" (Bergmann et al. 1989). For isoenzyme banding patterns (zymograms), the development of a computer program for the formulation of hypotheses on the mode of inheritance is a complex task, due to the different ways in which isoenzymes expressed in haploid tissue correspond to genes at loci. Whereas each enzyme molecule of a monomeric enzyme system is the product of the gene at a single locus, polymeric enzymes are formed from two or more enzyme subunits, each of which is the product of the gene at a locus. Four types of enzyme molecule occur in diploid tissue.

TYPES OF ISOENZYMES

<< HOMOMERIC >> isoenzymes consist of subunits that are encoded by genes of the same type at the same locus. Monomeric isoenzymes, which consist of only a single subunit, are treated as homomerics.

<< INTRALOCUS HETEROMERIC >> isoenzymes consist of subunits encoded by two genes of different types (alleles) at the same locus.

<< INTERLOCUS HETEROMERIC >> isoenzymes consist of subunits encoded by genes at two (or more) different loci.

<< POST-TRANSLATIONAL MODIFICATION (PTM) >> is an enzyme molecule, the electrostatic charge or molecular conformation of which was modified, probably by the product of a gene considered to belong to the "genetic background" (i.e. not coding for subunits of the enzyme system being studied). PTM affects the migration velocity through the gel. If not all molecules of a particular subunit structure in an individual are modified, PTM results in the appearance of one or more additional bands in the zymogram. Two types of PTM of molecules of a given subunit structure can be distinguished within a collection of individuals in its environment: A PTM of a particular molecule will be termed << FIXED >>, if the PTM occurs in all members possessing the molecule, and otherwise << FACULTATIVE >>.

INTERPRETATION OF BANDING PATTERNS

The strategy formulated by Gillet (1996) is to identify all elementary zones that contain homomeric isoenzymes and then to partition these zones into disjoint sets such that each set represents a complete set of transmission homologous gene types, i.e., the set of all alleles of a locus. Thus, the alleles present at each locus in the sample of banding patterns are represented by a set of elementary zones, and the alleles present in any given banding pattern are revealed by the appearance of a band in either only one (individual is homozygous) or two (individual is heterozygous) or even in none (individual is homozygous for "null allele") of these zones.

>> Non-fixed elementary zones of homomerics are independent:
If the sample of banding patterns is genetically closed, then all non-fixed elementary zones representing homomerics are independent. In the other direction, all elementary zones representing homomerics are independent, with only one rare exception: The facultative PTM of a homomeric at a fixed locus will also be independent.

>> Zones of intra- and interlocus heteromerics are dependent:
The appearance of a band representing an interlocus heteromeric depends on the appearance of the two bands representing the corresponding homomerics. An exception is the case in which one of the genes is a null allele that produces the heteromeric but not the homomeric.

>> Zones of PTM are dependent:
Appearance of a band in an elementary zone representing a PTM is dependent on the appearance of the unmodified isoenzyme (as long as not all molecules are modified and the zone of the unmodified isoenzyme is not fixed). This dependence distinguishes the elementary zones of heteromerics and PTM's from those of homomerics.

5. DNA ELECTROPHEROGRAMS

TYPES OF DNA FRAGMENT

In DNA electrophoresis, each elementary zone represents a DNA fragment "encoded" by a single gene, since fragments analogous to heteromeric isoenzymes and post-translational modification are thought not to occur.

INTERPRETATION OF BANDING PATTERNS

>> Non-fixed elementary zones are independent and represent a single gene at some locus:
If the sample of banding patterns is genetically closed, then all non-fixed elementary zones are independent and represent a single gene at some locus.

Conversely, if an elementary zone in a given sample is found not to be independent, the sample cannot be genetically closed, and no hypothesis can be formulated.

If all elementary zones show independence for the given sample, the qualitative interpretation of the banding patterns is the same as for monomeric isoenzymes without post-translational modification. The "null alleles" that often occur in DNA analysis, especially in RAPD, are also analogous to the "null alleles" of isoenzyme analysis.

Thus the genetic interpretation of DNA electropherograms is conceptionally much simpler than that of isoenzyme banding patterns. The number of elementary zones can, however, be much larger (e.g. DNA fingerprints), requiring a much larger sample to ensure genetic closure.

6. GENERATING HYPOTHESES

<DIPLOGEN> systematically generates all possible modes of inheritance under the assumption that the independent elementary zones exactly correspond to the genes. First, all possible modes of transmission are generated as all possible partitions of the set of independent elementary zones, such that at least two zones are assigned to each subset. Since independent zones are non-fixed, the requirement of at least two zones per subset corresponds to the presence of at least two alleles at each non-fixed locus.

For each such partition, and thus each possible mode of transmission, all hypotheses on the mode of gene action are generated by running through all possible combinations of codominance resp. dominance in the presence of a (recessive) "null allele" at the different loci.

Each hypothesis on the mode of inheritance is subsequently tested by constructing the set of all banding patterns, considering only the independent elementary zones, that would be found in a genetically closed set of individuals and comparing them to the observed set of banding patterns, likewise considering only the independent zones. If the two sets of patterns exactly match (or, in the case of Ezone splitting, if at least 3/4 of the expected banding patterns were observed), <DIPLOGEN> prints out the mode of inheritance as a hypothesis.

SPLITTING ELEMENTARY ZONES

It frequently happens that single isoenzymes or DNA fragments that are produced by genes at different loci nonetheless migrate to the same position in the gel. Since their respective bands are usually indistinguishable (except for the rare case that differences in band intensity are interpretable), they will be assigned to the same elementary zone. Since this elementary zone has two different genetic interpretations, <DIPLOGEN> is unable to formulate a hypothesis for mode of inheritance.

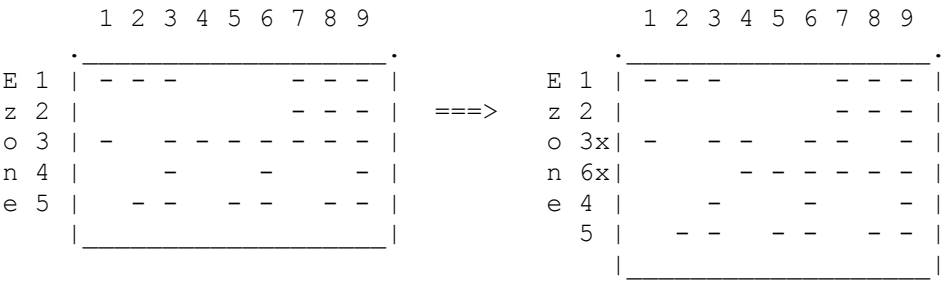
To alleviate this problem, <DIPLOGEN> provides the option of splitting one elementary zone at a time into two zones, alternately assigning the band appearing in the original zone in a banding pattern to either one or to both of the new zones. All combinations of assignment to the first new zone, the second new zone, and to both new zones are produced among all of the banding patterns that exhibit a band in the original zone.

If a banding pattern exhibits a band in the elementary zone to be split, there are three ways of assigning this band to the two new zones:

- band appears in new zone 1 but not 2 (splitting code 1)
- band appears in new zone 2 but not 1 (splitting code 2)
- band appears in both new zones 1 and 2 (splitting code 0).

If N banding patterns exhibit a band in the elementary zone to be split, there are 3^N ways to assign the band to one or both of the new zones among the N patterns. Subtracting the three trivial cases where for each banding pattern the band is always assigned to the same new zone or to both zones, there remain 3^N-3 ways to split the bands appearing in the original elementary zone over the two new zones. Considering that each of these 3^N-3 cases has a symmetric counterpart that gives the same banding pattern (i.e., when all splitting codes 1 are replaced by 2 and all 2's by 1's), a total of (3^N-3)/2 different ways to distribute the N bands over two new elementary zones exist.

EXAMPLE 1: For the sample of isoenzyme banding patterns to the left, no hypothesis can be formulated for which the sample is genetically closed. By splitting elementary zone 3 into the two zones 3x and 6x, as done in the right diagram, a hypothesis can be formulated for which the sample is genetically closed: Elementary zones 1 and 6x are alleles of locus 1 with intralocus heteromeric 2, and zones 3x and 5 are alleles of locus 2 with intralocus heteromeric 4.



7. INPUT FILE FOR <DIPLOGEN>

The input to the program is a matrix of zeros and ones that represents the different banding patterns observed in a (hopefully) genetically closed sample of diploid individuals. After the elementary zones of the patterns have been defined, each banding pattern can be described by a list of ones and zeros indicating presence or absence, respectively, of a band in the successive elementary zones.

To input m banding patterns, the user applies any text editor to prepare a data file (unformatted, ASCII characters only) consisting of m+3 lines (optionally m+2) as described in the following:

FORMAT OF INPUT FILE

Line 1: n = integer specifying number of different elementary zones

Line 2: (nI1) = usual FORTRAN format specification for reading banding patterns as a list of integers, each of width 1.
Other FORTRAN formats for reading list of integers can be specified, e.g. to include blanks of width w by wX (X-format) or define each of k integers to be of width w by kIw.

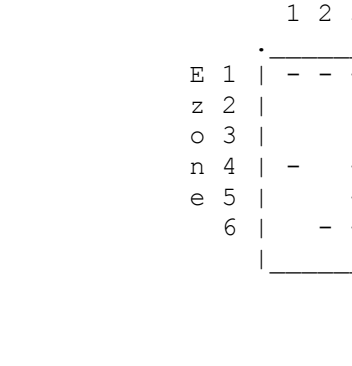
Lines 3 to m+2, one line for each banding pattern conforming to the format defined in Line 2:
A list of length n of 0's and 1's representing the banding pattern, where the entry in the j-th position of the list specifies the presence or absence of a band in the j-th elementary zone:
-> "1" signifies presence
-> "0" signifies absence

Line m+3: A "9" in the first position (according to format specification in Line 2) ends reading of the input file.
Alternatively, the file can be terminated at the end of the last line that defines a banding pattern. No carriage return may follow; otherwise, <DIPLOGEN> will read the following and any further empty lines as the banding pattern "000...0".

Banding patterns that are encountered more than once can be included in the input file as often as they appear, since <DIPLOGEN> recognizes redundant patterns and prints the number of times each pattern is encountered.

EXAMPLE 2

Schematic representation of banding patterns



Corresponding input data file, two possible variants

```
6                6
(6I1)            (3I1,1X,3I1)
100100           100 100
100001    OR     100 001
100111           100 111
001100           001 100
001001           001 001
001111           001 111
111100           111 100
111001           111 001
111111           111 111
9                9
```

8. RUNNING <DIPLOGEN>

When started, <DIPLOGEN> asks for the name of an input file that was prepared previously using a standard text editor (see Section 7 above). It then poses the following questions:

CHOOSE INPUT FILE

>> Name of input data file [default extension = .dat]: >>

Include path if data file is not located in the same directory as the program. If the file's extension is ".dat", it can be omitted and is supplied by the program. For example, the file "d:\path\infile.dat" can be given as "d:\path\infile", but "e:\zymo.tst" must be fully given.

>> ** File does not exist: XXX

If named file, here XXX, is not found, you are prompted to retry. Check path designation.

CHOOSE OUTPUT DEVICE

>> Output device? Screen only="s" or File+Screen="f"
[default="s"] : >>

- An answer of "s" causes all output to appear on the screen only - no output is saved for later reference.
- An answer of "f" causes all output to be saved in the output file and abbreviated output to simultaneously appear on the screen.

SPECIFY OUTPUT FILE

>> Name of output file [default = XXX.out] ? : >>

Press ENTER to give the output file the same path and filename (here represented by XXX) as the input file and the extension ".out". Otherwise, type complete path, filename and extension as desired.

>> ** File XXX.out already exists. Append="a", Overwrite="o"? : >>

- An answer of "a" causes new output to be appended to the end of the existing file XXX.out without changing previous contents of the file.
- An answer of "o" causes new output to be written at the beginning of XXX.out, and all previous contents of the file are lost.

SPECIFY TYPE OF BANDING PATTERNS

>> Type of pattern?:
Zymogram = "z", DNA electropherogram = "d" [default = "z"] >>

- If the answer is "d", <DIPLOGEN> treats all bands as alleles at some locus. In terms of programming technique, all bands are handled as if they were monomeric (thus homomeric) isoenzymes in a system allowing "null alleles" but without PTM.

GENERATING ADDITIONAL HYPOTHESES BY SPLITTING ONE EZONE INTO TWO OVERLAPPING EZONES

>> Do you want to search for overlapping Ezones (epistasis)? :
Yes="y", No="n" [default="n"] : >>

See Section 6 above.

>> Which Ezone should be split into two new Ezones?
Ezone N = "N", All Ezones = "0", End program = "-1"
[no default]: >>

- If the answer is a positive integer "N", then only elementary zone N is split.
- If the answer is "0", all elementary zones are split, one at a time.
- If the answer is "-1", the program is terminated.

>> If an Ezone exhibits a band in N patterns, there are (3^N-3)/2 ways to distribute the N bands over two new Ezones, such that for each of the N patterns, a band appears in at least one of the new zones.
Input maximal N not greater than nn for which Ezone splitting is to be performed [no default] : >>

- An answer of "N" causes only those Ezones to be split in which min(N,nn) or fewer banding patterns exhibit a band, where nn is originally set in <DIPLOGEN> to equal 8.

9. REFERENCES

Bergmann F, Gillet EM. 1996. Phylogenetic relationships among pine species inferred from different numbers of 6PGDH loci. Plant Systematics and Evolution 208, 25-34.

Bergmann F, Gregorius H-R, Scholz F. 1989. Isoenzymes, indicators of environmental impacts on plants or environmentally stable gene markers? In: Scholz F, Gregorius H-R, Rudin D (eds.): Genetic Effects of Air Pollutants in Forest Tree Populations. Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, pp 3-6.

Gregorius H-R. 1980. The probability of losing an allele when diploid genotypes are sampled. Biometrics 36, 632-652.

10. TECHNICAL CONSIDERATIONS

<DIPLOGEN> is written in Fortran~77. The version available on the internet under URL:

http://www.uni-forst.gwdg.de/forst/fg/index.htm

is compiled for DOS and runs with Windows, but compilation for other systems may be possible upon request.

E-Mail of author: egillet@gwdg.de