SIM – a simulation program for genetic linkage and association

SIM is a program to simulate multiple traits each with both genetic and environmental effects to be specified. The genetic components include major gene effects as well as polygenic effects and the genetic loci can be either in linkage equilibrium or in linkage disequilibrium. It uses input formats similar to LINKAGE files and outputs appropriate pedigrees from the simulation.

1 The genetic model

This program implements the genetic model

$$x = g + p + c + e$$

where x is a multivariate trait ($x_i, i = 1, ..., NTRAIT$), g represents major loci ($g_j, j = 1, ..., NMG$), p represents polygenic effects ($p_k, k = 1, ..., NPG$), c is the common environment ($c_l, l = 1, ..., NCE$) and e represents unique environments ($e_m, m = 1, ..., NUE$). The variables g, p, c, e are referred to collectively as causal components of the trait. The NMG major loci are a subset of $NLOCI$ loci for which the order and recombination fractions are $(\theta_1, \theta_2, ..., \theta_{NLOCI-1})$, and the number of alleles and allele frequencies are specified. The meaning of some model constants are summarized in table 1.

Major locus effects AA, Aa, aa are characterized by z, q, t, d, where z is the mean effect of AA, q is the allele frequency of a, t is the displacement between AA and aa and d is the dominance (see table 2).

Suppose g has mean 0 and variance 1, we have

$$z = -(tq^2 + 2pqdt)$$

and

$$t^2 = \frac{1}{(pq)^2(q + 2pd)^2 + 2pq(d - q^2 - 2pqd)^2 + q^2(1 - q^2 - 2pqd)^2}$$

so that the only free parameters are q and d.

The polygenic effect of an offspring, conditional on those of parents (p_F, p_M), is

$$p_o = \frac{p_F + p_M}{2} + \frac{u}{\sqrt{2}}$$

Table 1: Some model constants

Parameter	Meaning
$MAXLOCI$	maximum # of loci
$MAXALLELES$	maximum # of alleles at each locus
$MAXFAM$	maximum # of families
$MAXIND$	maximum # of individuals within a family
$NTRAIT$	# of traits
$NLOCI$	# of loci
NMG	# of major genes
NPG	# of polygenes
NCE	# of common environments
NUE	# of unique environments
$NDISEQ$	# of pairs in disequilibrium
β	matrix of regression coefficients for each trait

Table 2: Major locus parameters

Genotype	Frequencies	Effect(g)
AA	p^2	z
Aa	$2pq$	$z + dt$
aa	q^2	$z + t$

$u \sim N(0, 1)$ is specific for each offspring.

Suppose the degree of transmission of parental shared environment to offspring is k, then under no assortative mating, the common environment for offspring is $k(c_F + c_M) + v\sqrt{(1 - 2k^2)}$, where $v \sim N(0, 1)$ is the same within the whole sibship. The valid range for k is $(0, 1/\sqrt{2})$.

The unique environment is different among individuals and is $N(0, 1)$.

All latent variables (g, p, c, e) have mean 0 and variance 1.

The regression equation of trait x on latent variables is

$$\begin{aligned} x_i &= \beta_{ig_1} g_1 + \cdots + \beta_{ig_{NMG}} g_{NMG} \\ &+ \beta_{ip_1} p_1 + \cdots + \beta_{ip_{NPG}} p_{NPG} \\ &+ \beta_{ic_1} c_1 + \cdots + \beta_{ic_{NCE}} c_{NCE} \\ &+ \beta_{ie_1} e_1 + \cdots + \beta_{ie_{NUE}} e_{NUE} \end{aligned} \quad (1)$$

The mean and variance of x_i are then 0 and $\sum \beta_{ij}^2, j = 1, ..., NMG + NPG + NCE + NUE$. To ease the specifications, the input β's are standardized by this in order to have the unit trait variance.

1.1 Introduction of linkage disequilibrium

It is possible to include disequilibrium between multiallelic markers. Without loss of generality, consider two marker loci, with m and n alleles and allele frequencies $p_1, p_2, ..., p_m$ and $q_1, q_2, ..., q_n$, then we can arrange the haplotype frequencies into a $m \times n$ contingency table.

The $m \times n$ contingency table χ^2-squared statistic

$$\chi^2 = N(\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{h_{ij}^2}{p_i q_i} - 1)$$

where N is the total cell counts.

Under Hardy-Weinberg equilibrium, the haplotype frequencies can be obtained from the marginal allele frequencies. The above expression can be rewritten as a familar quadratic form. let $\mathbf{w}$ be the $m \times n$ diagonal matrix of direct product of the original marginals, the χ^2-squared statistic is $\mathbf{h}\mathbf{w}^{-1}\mathbf{h}' - 1$, where $\mathbf{h}$ contains the actual haplotype frequencies.

Disequilibria among different alleles are introduced by numerical optimization of the χ^2-sqaured statistic over nonnegative contraints of haplotype frequencies to get their estimates. The constraints are formed through an iterative procedure to allow for individual disequilibrium specifications.

If $m, n = 2$, this will reduce to the familiar case of 2×2 table.

2 The implementation

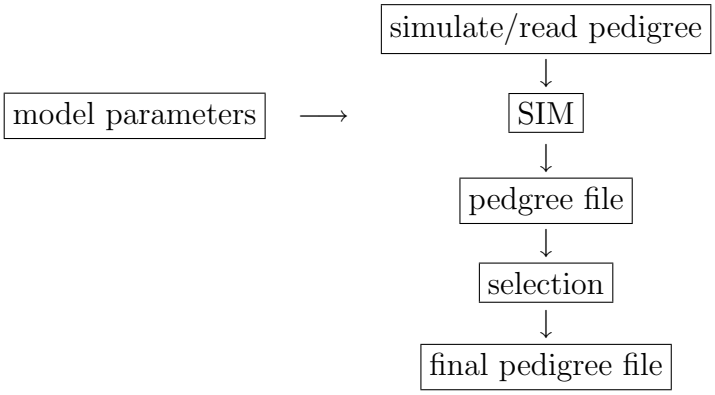
The program first gets parental haplotypes from population frequencies and then proceeds to generate the genotypes for children. Some loci are assigned as major locus with dominance parameters. The disequilibrium is simulated if the disequilibrium matrix of haplotype frequencies is specified.

The usual way of obtaining allele in a specific locus is as described in Weir (1990). Consider a 4-allele locus with allele frequencies p_1, p_2, p_3, p_4, when drawing a (pseudo)random number from [0,1] we could divide the unit interval into four segements, with boundaries $0, p_1, p_1+p_2, p_1+p_2+p_3, 1$, these segements therefore have lengths p_1, p_2, p_3, p_4. We assign allele number i if

the random number falls within the ith interval. Similar logic is applicable to simulation of recombination events. We can achieve this by using other algorithms for generating random variable from multinomial distribution.

The program has a setup that would be suitable for taking information from pedigree files similar to LINKAGE and SIMULATE. Figure 1 depicts the steps.

Figure 1: Flowchart of SIM



Currently the step SIM processes one family at a time, so that certain selection procedure can be adopted easily. A separate SAS program was written to generate the disequilibrium coefficients from linearly constrained optimization. To account for possible extensions involving many other statistical distributions, the freeware RANLIB was used. Finally, to keep the program in its pure C form, the Numerical Recipes utitlity nrutil.h/nrutil.c has been used to create dynamic arrays and matrices, although this could be easily achieved with the **new/delete** construct in C++.

A compiling control file Makefile has been created and tested under Sun Solaris and DEC Alpha with GNU C++ compiler, so that a simple command **make** would generate the executable. It has also been tested under PC with Symantec C++7.2. Prolix output could be obtained by specifying #undef DEBUG statement in include/sim.h.

The format of the command would be as follows:

SIM locusfile pedigreefile controlfile outputfile

A simple example is provided here. The locus file, pedigree file, control file, and output file are loc.tst, ped.tst, problem.dat and sim.out, respectively.

loc.tst

```
3 0 0 0 1 0 2 2 2 0 <<NLOCI,riskloci,SEXLINK,program,NTRAIT,NMG,NPG,NCE,NUE,NDISEQ
0.0 0.0 0.0        << MUT LOCUS, MUT RATE, HAPLOTYPE FREQUENCIES (IF 1)
1 2 3              << locus order
1 2                << affection, No. of alleles
 0.010000 0.990000 << gene frequencies
 3                 << No. of liability classes
0.44 0.44 0.0166
0.44 0.44 0.0166
0.44 0.44 0.0166
3 2                << allele numbers, No. of alleles
 0.67 0.33
3 3                << allele numbers, No. of alleles
 0.32 0.32 0.36
0 0                << SEX DIFFERENCE, INTERFERENCE (IF 1 OR 2)
0.5 0.12           << recombination fraction r[NLOCI-1]
1 0.10000 0.45000  << REC VARIED, INCREMENT, FINISHING VALUE
0.2 0.4  0 0   0 0 << simulated model beta[NTRAIT][NMG+NPG+NCE+NUE]
0.2 0.2            << path coefficient kce[NTRAIT][NCE]
```

Indeed it is very similar to **LINKAGE** parameter file, except extra parameters are specified in the first line and after the usual **LINKAGE** parameter file finishes if there are disequilibrium pairs. However the dominance parameter does need to be specified, with an extra line after the locus type and allele frequency line.

```
4 2 << locus type, number of alleles
0.01 0.99 << allele frequencies
0.5 << d, the dominance
```

The regression coefficients, path coefficients of common environment, and haplotype frequencies of disequilibrium pairs are specified as separate blocks after a typical end of LINKAGE parameter file, i.e., after "variation of recombination". For each trait, the regression coefficients are arranged into one line, so are the path coefficients of common environment transmission. For a pair of loci in disequilibrium the haplotype frequencies, the ordinal number of the first locus is also necessary, so the whole information is specified as follows, Note that the loci are numbered from 0.

```
first locus number
the disequibrium matrix
```

If there are more than one pairs then this is repeated for each pair. It might be redundant to input the allele frequencies for loci in the pair but it is just a matter of convenience.

ped.tst

```
1  1  0  0 2  2 1   1 1 1
1  2  0  0 1  1 1   1 1 1
1  3  2  1 2  2 3   1 1 1
1  4  2  1 1  2 1   1 1 1
1  5  2  1 1  2 1   1 1 1
1  6  2  1 1  2 1   1 1 1
1  7  0  0 1  0 1   0 0 0
1  8  0  0 2  0 1   0 0 0
1  9  7  8 2  1 1   1 1 1
1 10  2  1 2  2 1   1 1 1
1 11  7  8 1  1 1   1 1 1
1 12  6  9 1  1 1   1 1 1
1 13  6  9 1  1 1   1 1 1
1 14  6  9 1  2 1   1 1 1
1 15 11 10 1  1 1   1 1 1
1 16 11 10 1  2 1   1 1 1
... ...
6  1  0  0 1  1 1   0 0 0
6  2  0  0 2  1 1   0 0 0
6  3  0  0 1  1 1   0 0 0
6  4  1  2 2  2 1   1 1 1
6  5  1  2 2  2 1   1 1 1
6  6  1  2 1  1 1   1 1 1
6  7  0  0 2  1 1   0 0 0
6  8  1  2 1  1 1   1 1 1
6  9  1  2 1  1 1   1 1 1
6 10  0  0 2  1 1   1 1 1
6 11  3  4 1  1 1   1 1 1
6 12  3  4 1  2 1   1 1 1
6 13  3  4 1  2 1   1 1 1
```

```
6 14   3   4 1   2 1    1 1 1
6 15   3   4 2   1 1    1 1 1
6 16   3   4 1   2 1    1 1 1
6 17   3   4 2   1 1    1 1 1
6 18   6   7 1   2 3    1 1 1
6 19   6   7 2   1 1    0 0 0
6 20   6   7 2   2 2    1 1 1
6 21   9 10 1   2 1    1 1 1
6 22   9 10 1   1 1    0 0 0
6 23   9 10 1   1 1    1 1 1
6 24   9 10 1   1 1    1 1 1
6 25   9 10 2   1 1    1 1 1
6 26   9 10 2   1 1    1 1 1
```

We see that each line contains pedigree id, individual id, parent ids, sex and marker availability codes taking value of 1 if the genotype is available or 0 otherwise. The pedigree file contains the pedigrees in pre-MAKEPED (or post-MAKEPED format, then #USEMAKEPED in the header file/sim.h has to be activated. This is only useful when there are loops in the family and sex-linked disorders) and with indicators specifying if an individual has genotype at that locus or not. Note it has the same requirement as in SIMULATE, i.e., the order has to be such that parents' ids precede their offsprings.

problem.dat

```
10          << number of replicates
1232 2122   << random number seeds
```

This file contains the number of replicates and two random number seeds. Owing to RANLIB, the eligible ranges for random number seeds are huge, in [1, 2,147,483,562] and [1, 2,147,483,398], repectively.

Two demo files named sim_loc.tst and sim_ped.tst show all the features, they can be used by the command:

SIM sim_loc.tst sim_ped.tst

A driver program **diseq** is provided to simulate nuclear families assuming linkage disequilibrium. The sibship size distribution is based on the popular negative binomial distribution (Cavalli-sforza and Bodmer 1971) or truncated Poisson distribution.

Remark A version of CHRSIM, SIMULATE, together with the LINKAGE utility routine makeped.c (available from Rockefeller) have been included as separate programs. The algorithms in CHRSIM has been extended to be able to include random markers based on recent estimates of chromosome lengths and Genethon map (Dib et al 1996), which would make it more flexible and suitable for genome scanning research.

3 Testing of the program

The first test could involve test of HWE. We can also consider further genetic analysis. One possibility is to use POINTER to recover the model for nuclear families, where the maximum-likelihood estimation was performed by iterating upon the following: V, variance of x; u, mean of x; d, degree of dominance at major locus; q, gene frequency at major locus; t, displacement at major locus; H, polygenic heritability (C/V); and B, relative variance due to common environment (S/V). If only affection status can be specified, it is defined on x, hence $W = 0$; mean and variance of x are arbitrary and can be taken as $V = 1$ and $u = 0$. Affection occurs whenever liability x is greater than some threshold. For such a specification to be independent of any particular model of inheritance in segregation analysis, one must provide an estimate of the prior probability of affection in the reference population.

Other programs such as PAP and MORGAN may also be used.

Another example is the "disequilibrium and causation differentiation". This is pending to be further explored.

4 Some notes

The first is about sibship distribution. As indicated by Morton et al (1983), phenotype specification may concern a quantitative measurement, or affection. Ideally some demographic model can be incorporated, as in POPGEN, so age of onset can be considered.

The distribution of family sizes in the population was described in Cavalli-Sforza and Bodmer (1971), where they found negative binomial distribution gives better fit to the data they used. While Ewens (1982) and Morton (1982) showed that family-size distribution should not affect the result of segregation analysis, the distribution of family sizes can affect power calculations because

a data set containing 100 two-child families contains less information than, say 100 five-child families.

The program could be used for provding data in multiple traits analysis.

The program generates genetic data purely from gene dropping through families and does not consider conditioning of relatives phenotype as in SLINK. It is thus fast but restricted in terms of analysis. The program is very flexible in terms of disequilibrium specifications.

Another program, developed by Kaplan et al for linkage disequilibrium analysis in founder population, beared the same name.

Would it be worthwhile to convert the SAS optimizer to CFSQP ?

Would it be optimal if size controlled by command-line parameters ? (as in SPLINK/TRANSMIT, 21-4-1999)

5 Program history

- 6/3/1997 first draft
- 29/5/1997 changed to get families first, setup for read-ins
- 6/97 add ad-hoc program used for disequilibrium problem
- 26/8/1998 improved code and documentation (where is it now ?)
- 9/3/1999 submitted as attachment with minor change to GAW12
- 22/4/1999 expanded documentation

6 References

1. Boyle, CR and Elston, RC (1979) Multifactorial genetic models for quantitative traits in humans. Biometrics 35: 55-68.
2. Cavalli-Sforza, LL and Bodmer, WF(1971). "The Genetics of Human Populations. San Francisco, W.H.Freeman. pp310-313.
3. Curtis, D. and Sham, PC (1995) Model-free linkage analysis using likelihoods. Am. J. Hum. Genet. 57:703-716.

4. Elston, RC (1980) Segregation analysis. in Current Developments in Anthropological Genetics, vol I, edited by Mielke, JH, Crawford, MH, New York, Plenum Press, pp327-354.

5. Ewens, WJ (1982) Aspects of parameter estimation in ascertainment sampling schemes Am. J. Hum. Genet. 34: 853-865.

6. Greenberg, DA (1984) Simulation studies of segregation analysis: application to two-locus models. Am. J. Hum. Genet. 36:167-176.

7. Hasstedt, SJ, Meyers, DA, Marsh, DG (1983) Inheritance of immunoglobin E: genetic model fitting. Am. J. Med. Genet. 14: 61-66.

8. Khoury, MJ, Beaty,TH and Cohen, BH (1993) Fundamentals of Genetic Epidemiology. Oxford University Press, Inc.

9. Lalouel, JM and Morton, NE (1981) Complex segregation analysis with Pointers. Hum. Hered. 31: 312-321.

10. Morton, NE and MacLean, CJ (1974) Analysis of family resemblance. III. Complex segregation analysis of quantitative traits. Am. J. Hum. Genet. 27:365-84.

11. Morton, NE(1982) Trials of segregation analysis by deterministic and macro simulation. Am. J. Hum. Genet. 34:187A.

12. Morton, NE, Rao, DC and Lalouel, JM (1983) Methods in Genetic Epidemiology. Karger.

13. Press, WH, Teukolsky, SA, Vetterling, WT, and Flannery, BP (1992) Numerical Recipes. The Art of Scientific Computing. Second Edition. Cambridge University Press. http://cfata2.harvard.edu/nr/.

14. Risch, N(1984) Segregation analysis incorporating linkage markers. I. single-locus models with an application to Type I diabetes. Am. J. Hum. Genet. 36:363-386.

15. Schork, NJ (1992) Detection of genetic heterogeneity for complex quantitative phenotypes. Genet. Epidemiol. 9: 207-223.

16. Sham, PC(1998) Statistics in Human Genetics. Edward Arnold.

17. Terwilliger, JD and J Ott (1994) Handbook of Human Genetic Linkage. The Johns Hopkins University Press.

18. Thompson, EA and Cannings, C(1979) Sampling schemes and ascertainment. In The Genetic Analysis of Common Diseases: Applications to Predictive Factors in Coronary Heart Disease. C.F. Sing and M Skolnick, (eds) Alan Liss, New York.

19. Thompson, R(1977a) The estimation of heritability with unbalanced data. I. Observations on parents and offspring. Biometrics 33: 485-495.

20. Thompson, R(1977b) The estimation of heritability with unbalanced data. II. Data available on more than two generations. Biometrics 33: 497-504.

21. Weir, BS(1990) Genetic Data Analysis - Methods for Discrete Population Genetic Data. Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts.

7 Contact information

I would be very pleased to get your views about the program. If you find any problem please feel free to contact me at the following address.

Jing Hua Zhao

Department of Psychological Medicine, Institute of Psychiatry, De Crespigny Park, Denmark Hill, London SE5 8AF, UK. Tel +44 (171) 9193534, Fax +44 (171) 7019044

j.zhao@iop.kcl.ac.uk