

Modeling gene-gene-interactions using graphical chain models

Short title: Modeling gene-gene-interactions

Ronja Foraita, Karin Bammann, Iris Pigeot

Bremen Institute for Prevention Research and Social Medicine (BIPS)
University of Bremen

May 15, 2007

Objective: To investigate whether graphical chain models are suitable to detect gene-gene-interaction under different biological models.

Methods: We conducted a simulation study comparing graphical chain models with logistic regression models regarding their ability to detect underlying biological interaction models. For both methods, we attempted to capture simulation data following 12 different biological models. We used 10 statistical models for both methods. Of the 12 different biological models, four contained no interaction effects, two were multiplicative, and six were epistasis models. For each situation, the choice for a statistical model was based on global model fit as judged by two different information criteria, the BIC and the AIC.

Results: Both methods failed in most of the scenarios to capture the gene-gene-interaction present in the simulation data. Only in very specific cases, when disease risk was high and both genes had a dominant effect, present gene-gene-interaction was detected.

Conclusions: Graphical chain models are, similar to logistic regression models, not able to capture gene-gene-interactions for arbitrary biological models underlying the data.

Key words: Case-control study; Epistasis; Gene-gene interaction; Graphical chain models

1 Introduction

The meaning of gene-gene interaction has led to much confusion caused by different definitions of interaction or epistasis as it is usually referred to in genetics, statistics or epidemiology. Thus, different models are used in the different disciplines to capture gene-gene interactions. The problem is that the degree to which statistical analysis can

elucidate the underlying biological mechanisms may be limited and may require prior knowledge of the underlying etiology. In addition, it may also depend on the study design and the applied statistical methods. For instance, Vieland and Huang [1] have shown that data from affected sib-pairs cannot be used to distinguish heterogeneity from epistasis.

In a recent paper, North et al. [2] presented a simulation study that investigated the capability of logistic regression models to detect gene-gene interaction in a case-control study under different biological models. They assumed two independent susceptibility loci, i. e. the linkage disequilibrium was assumed to be zero ($LD = 0$), with joint influence on the disease risk. Their results showed that logistic regression models cannot convincingly reflect the biological mechanisms of interaction. This leaves the question open whether other statistical models may be more appropriate to reflect the biological meaning of gene-gene interactions as already suggested by North et al. [2]. Graphical chain models may offer a reasonable alternative. Although they are also based on regression models they allow in contrast to simple regressions to reflect complex multivariate association structures including intermediate variables and interactions in a graph. This would allow to capture biological pathways and to simultaneously account for environmental factors. The advantages of graphical models compared to simple logistic regression models became for instance apparent in Didelez et al. [3]. Thus, we study the properties of graphical chain models in small settings with the aim to generalize this approach to complex disease models.

The paper is organized as follows. In Section 2, graphical chain models, biological models of two-loci-interaction and the applied statistical models are introduced. Here, we focus on the different meanings of gene-gene interaction. The statistical models will then be investigated whether they are able to capture the biological interaction by means of simulation studies. Section 3 describes the design of the simulation study. The results are presented in Section 4 where it is especially investigated whether graphical chain models perform better in this context than the logistical regression models used by North et al. [2]. The above findings are critically reflected in the discussion.

2 Methods

Let us first define two-loci interaction in a biological sense (cf. [4]). *Locus heterogeneity* means that two genes have an influence on the disease but act separately. *Epistasis* means that the two genes interfere with each other in their influence on the disease. Bateson [5] originally introduced the term *epistatic* to describe a masking effect in which a factor at one Mendelian locus prevents another from manifesting its effect. In quantitative genetics, the term epistatic has classically been used as introduced by Fisher [6]. It refers to a deviation from additivity in the effects of the alleles of different loci with respect to prediction of a quantitative phenotype. This is similar to the usual concept of statistical interaction. Here, however, the choice of the scale becomes important, since two variables that interact, for example, on a multiplicative scale may be purely additive on a logarithmic scale.

2.1 Biological interaction models

Consider a disease phenotype Y and a sample of cases ($Y=1$) and controls ($Y=0$) selected from an arbitrary population. We assume that genotypes are obtained for each individual for a set of biallelic, autosomal candidate loci. For each candidate locus, say A and B , G_A and G_B give the number of disease alleles, i. e. G_A, G_B take possible values 0,1, or 2. In the following, let allele a denote the wild-type allele, and allele A the disease allele at locus A , b and B are defined analogously. For binary traits, we are interested in the probability of being affected given a certain joint genotype: $P(Y=1|G_A=i, G_B=j)$ is usually referred to as joint penetrance f_{ij} , $i, j = 0,1,2$. That is, f_{ij} is the penetrance for a genotype with i copies of allele A at locus A and j copies of allele B at locus B , respectively.

Based on these penetrances, Risch [7] transferred the three main biological models of gene-gene interaction to mathematical models applying penetrance factors respectively summands depending on the underlying model. The additive model assumes that no interaction is present and thus describes the joint penetrance as sum of the two penetrance summands, i.e. $f_{ij} = a_i + b_j$.

The heterogeneity model $f_{ij} = a_i + b_j - a_i \cdot b_j$ corresponds to the standard probability equation for calculating the probability of the union of two overlapping events where stochastic independence of these two events is assumed. Risch [7] showed that the heterogeneity model can often be well approximated by the additive model when used to model relative risks of disease in families. The additive and the heterogeneity models are assumed to represent biological models without interaction. The multiplicative model $f_{ij} = a_i \cdot b_j$, in contrast, represents biological interaction, where the penetrance can be written as product of the two penetrance factors [7,8]. Another approach to capture genetic interactions is based on so-called pure epistasis models that mainly model potential masking effects of one or both genes. Thus, various types can be distinguished depending on the inheritance model (cf. [9]). For two dominant genes the biological mechanism can be described by an epistasis model if only the occurrence of at least one disease allele at both loci leads to an increased penetrance (for an example see the relative penetrance matrix for Epi^{dd} in Table 5). In case of one dominant and one recessive gene the biological model reflects epistasis if the recessive gene masks the dominant one. Thus, the dominant gene only leads to an increased penetrance if two disease alleles occur at the recessive gene locus (for an example see the relative penetrance matrix for Epi^{rd} in Table 5). For two recessive genes, an epistasis model reflects that both genes mask each other. In this case, the penetrance can only be increased if two disease alleles occur at each locus (for an example see the relative penetrance matrix for Epi^{rr} in Table 5). These three epistasis models are based on the definition of Bateson.

2.2 Statistical models

The models proposed by Risch can be written as linear regression models using the parametrization as introduced by Cordell [10] and later used by North et al. [2]. At each candidate locus, G_A, G_B are subdivided into two design variables. The additive effects $x_1, x_2 \in \{-1, 0, 1\}$ are linear transformations of the genotype and reflect the number of disease alleles. The dominance effects $z_1, z_2 \in \{-0.5, 0.5\}$ distinguish homozygous from heterozygous genotypes. For instance, x_i is set to -1 and z_i to -0.5 for a homozygote aa , $x_i = 0$ and $z_i = 0.5$ for a heterozygote aA and $x_i = 1, z_i = 0.5$ for the genotype AA . Using this parameterization, an additive biological penetrance model corresponds to the following statistical model:

$$y = \alpha + \beta_1 x_1 + \gamma_1 z_1 + \beta_2 x_2 + \gamma_2 z_2.$$

This relationship can be exploited such that the penetrance summands can be calculated from the regression parameters of the linear model as $a_0 = -\beta_1 - 0.5\gamma_1$, $a_1 = 0.5\gamma_1$, $a_2 = \beta_1 - 0.5\gamma_1$ (b_j can be derived analogously).

2.2.1 Graphical chain models

Graphical chain models are an adequate tool to display a multivariate association structure in a graph. Variables are represented as vertices and associations between each pair of these variables as edges. Undirected edges represent symmetric associations, whereas directed edges pointing from an influential variable to a potential response represent asymmetric associations. In case that the so-called Markov properties hold, missing edges can be interpreted as conditional independences, as illustrated in Figure 1. A graphical chain model consists of several boxes usually ordered from right to left where the pure explanatories are summarized in the box on the right and the pure responses in the box to the left. In between are boxes with intermediates that are responses to variables on the right and explanatories to variables on the left. Edges within boxes are always undirected and edges between boxes are always directed pointing from variables in boxes on the right to variables in boxes on the left. For more details we refer to [11,12].

here Figure 1

Typically, a first rough structure is obtained from subject-matter knowledge where the variables are categorized as pure explanatories, pure responses and different levels of intermediates. The specific associations between the variables have then to be derived from the data using an appropriate selection strategy [12].

Using graphical chain models, it has first to be decided on which information should be displayed in the graph. To ensure comparability with logistic regression models we follow the idea to divide each genotype into its dominance and additive effect. In addition, it has to be accounted for the fact that we are interested in case-control studies which implies that the disease status of an individual is given and the relation to

the exposure has to be clarified during the analysis. Thus, although perhaps not common, the disease status Y is depicted as explanatory variable in the right box whereas the genetic exposures X as additive effects and Z as dominance effects are represented as response variables in the left box.

2.2.2 Model equations

Let us now briefly summarize the statistical models to be further investigated where for the sake of simplicity we restrict ourselves to representing only the additive effects of two marker genotypes. We consider logistic regressions and graphical chain models to model the statistical association structure. For illustrative purposes, the latter are additionally displayed as graphs. Since we assume that all variables are discrete, we consider graphical loglinear models. That is, we start from an $I \times J \times K$ -table where the cell probabilities are denoted as $p_{ijk}, i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K$. The simplest case is that of a null model, which is given as

$$\text{logit}(y) = \alpha$$

as logistic regression and

$$\log(p_{ijk}) = \lambda + \lambda_i^{X_1} + \lambda_j^{X_2} + \lambda_k^Y$$

as graphical model. Since no associations are present in the mean model the corresponding independence graph does not contain any edges (see Figure 2a). Main effects can be included as follows in the above models:

$$\begin{aligned} \text{logit}(y) &= \alpha + \beta_1 x_1 + \beta_2 x_2 \\ \log(p_{ijk}) &= \lambda + \lambda_i^{X_1} + \lambda_j^{X_2} + \lambda_k^Y + \lambda_{ik}^{X_1 Y} + \lambda_{jk}^{X_2 Y}. \end{aligned}$$

The corresponding graph now contains edges pointing from Y to X_1 and X_2 , but still X_1 and X_2 are disconnected (see Figure 2b). If we now include an additional interaction effect, X_1 and X_2 are no longer disconnected but connected by an undirected edge (see Figure 2c). The corresponding models read as follows:

$$\begin{aligned} \text{logit}(y) &= \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \\ \log(p_{ijk}) &= \lambda + \lambda_i^{X_1} + \lambda_j^{X_2} + \lambda_k^Y + \lambda_{ik}^{X_1 Y} + \lambda_{jk}^{X_2 Y} + \lambda_{ij}^{X_1 X_2} + \lambda_{ijk}^{X_1 X_2 Y}. \end{aligned}$$

here Figure 2

In the following, we perform a simulation study based on the design of North et al. [2] to investigate whether statistical models of the type introduced above are able to reflect the underlying biological models.

3 Simulation Study

Different designs are considered for the simulation study where in general we assume that loci A and B are in linkage equilibrium. For the sake of comparison, we set the disease allele frequencies at both loci to 0.1 ($p_A = p_B = 0.1$) [9] and the disease prevalence to $prev = 0.01$ [2]. Let us denote the basic risk for developing the disease of interest with c regardless of the joint genotype. We then introduce a factor $\rho_t, t \in \{1, \dots, T\}$, to reflect a potential risk increase. For the Risch models, this factor depends on the number of disease alleles such that e. g. the penetrance vectors a_1 and a_2 can be derived from a_0 by multiplying a_0 with ρ_1 and ρ_2 , respectively.

For each biological interaction scenario, we simulate a model of low and of high risk increase. Table 5 in the appendix shows the marginal genotype relative risk for each locus, which is defined e. g. for locus A as $GRR_i = \frac{P(Y=1|G_A=i)}{P(Y=1|G_A=0)}$, and the relative

penetrance, calculated as $\frac{f_{ij}}{\min(f_{ij})}$. The specific simulation steps are as follows:

1. *First step.* Depending on the above assumptions and on the specific biological model, either the penetrance vectors (a_0, a_1, a_2) and (b_0, b_1, b_2) or the constant c are calculated. The results are used to build the penetrances f_{ij} and $1 - f_{ij}$.
2. *Second step.* Given the prevalence $prev$, the penetrances f_{ij} and $1 - f_{ij}$ and the joint genotype $G_l, l \in \{0, \dots, 8\}, l = 3i + j, i, j \in \{0, 1, 2\}$, the conditional probabilities of a certain joint genotype given the disease status, i. e. $P(G_l | Y = k), k \in \{0, 1\}$, are calculated. These probabilities are each multiplied with 10000 to obtain a population with 10000 affected (cases) and 10000 unaffected subjects (controls).
3. *Third step.* A random sample of size $N = 1000$ of the subpopulation with joint genotype G_l is drawn with replacement. This sampling step is replicated independently $n = 100$ times.

Note, that the simulation design of North et al. [2] did not include a random selection of the population to be investigated (Step 3). The generated data samples are used to examine whether the statistical models introduced above correspond to biological interaction models. For this purpose, ten statistical candidate models with different parametrization are formulated (see Table 1). In the following, statistical models are denoted by S with the special model type as index. The first seven models consist only of main effects representing biological models without gene-gene interactions. The last three models include interaction terms and shall therefore capture gene-gene interactions.

All statistical models are translated into graphical chain models and logistic regression models (see Table 1).

here Table 1

Each statistical scenario is fitted using a logistic regression model and the complementary graphical chain model where the statistical software packages R, MIM [12] and the interface mimR [13] are applied.

The global model fit is assessed by the Bayesian Information Criterion (BIC) and by Akaike's Information Criterion (AIC). The results obtained from the latter are described only briefly. Since on the one hand different model parameterizations do not allow to compare the absolute BIC resp. AIC-values and on the other hand only the relative values are relevant to select the best model out of a given set of candidate models, we consider the differences between the value of the actual model and the minimum of all BIC resp. AIC-values in a given set, which means e. g. for the BIC

$$\Delta_m^{BIC} = BIC_m - \min(BIC_1, \dots, BIC_{10}), m = 1, \dots, 10.$$

According to Burnham and Anderson [14] a selected model with $\Delta_m^{BIC} \leq 2$ or $\Delta_m^{AIC} \leq 2$, respectively, should be considered as particularly useful. Such models are said to have a substantial level of empirical support.

4 Results

The results of our simulation study are presented in Table 2. Statistical models that are assumed to reflect best the biological mechanisms, briefly denoted as "true" models are shaded gray. As a general result, the logistic regression models typically leads to a higher variety in selected statistical models as compared to graphical chain models. In addition, more models being consistent with the assumed biological mechanisms are selected based on logistic regression models compared to graphical chain models.

For the null model the mean log(OR) and the bias in log(GRR)s are close to 0 (mean log(OR) = -0.0028 using allele *b* as reference; bias log(GRR): -0.0027, -0.059 using genotype *bb* as reference). The bias in log(GRR) is the difference between the log of simulated GRR and the log of true GRR. The type I error rate (see Table 4) shows that LRM is too conservative, GCM comes up with good empirical error rates for the 0.01-level, but shows a slightly anti-conservative behavior for the interaction models.

Let us now go into more details, where we distinguish Risch models and epistasis models.

For the low risk additive model Add_L we presume S_{A2} and S_{D2} as reasonable "true" models, since an risk increase is observed for the marginal GRR at locus *B* in the relative penetrance matrix (see Table 5). Here, the logistic regression model (LRM) mostly selects the additive effect model S_{A2} (92%) whereas the graphical chain model (GCM) prefers the dominance effect model S_{D2} (86%). The "true" models for the high risk additive model Add_H are S_{A1} and S_{D1} . The tenfold risk increase provokes the GCM to chose S_{D1} in 98%.

Like the additive model, the heterogeneity model Het_L is regarded as biological model without gene-gene interaction. The “true” models for the low risk scenario are S_{A1} , S_{D1} and for the high risk model S_{D2} and S_{D12} , since Het_H includes an explicit dominance effect. Looking at Het_L , LRM finds again in contrast to GCM models without dominance effects. This changes for an increased risk and two recessive loci. LRM captures the “true” models S_{D2} and S_{D12} in 100%. GCM fails and selects S_{D2} in only 11%, but selects additive effect models in 89%.

The results for the multiplicative model show that LRM selects the additive effect model S_{A12} for the low risk model and S_{D12} for the high risk model. LRM finds as well some interaction models. GCM decides for S_{D12} regardless of whether a low or high risk model applies.

For the dominant low risk epistasis model Epi_L^{dd} , most frequently the mean model is selected. For the high risk model Epi_H^{dd} gene-gene interaction can be found. GCM as well as LRM fully reflect the underlying biological model. Substituting one dominant genotype for a recessive genotype, LRM and GCM have both difficulties to ascertain interaction effects for the Epi_L^{rd} . Almost all model replications favor the mean model. The high risk model performs better. Although the epistasis effect cannot be revealed, both approaches select main effect models S_{D1} resp. S_{A1} and hence find evidence that only locus A is involved in the pathway of disease. Both epistasis models with two recessive loci, Epi_L^{rr} and Epi_H^{rr} , show only minor variation in the marginal GRR's. Since only the joint genotype $AABB$ leads to an increase of risk, the choice of allele frequencies and the number of observations lead to a lack of statistical power for detecting any interaction effect. Both, LRM and GCM are not able to identify this relationship, neither for a 10- nor 100-fold risk increase.

In addition to the BIC, the AIC was computed for all analyses (see Table 3). The results come up with some remarkable differences. In general, the AIC selects much more models that are consistent with the data. This leads particularly for the LRM to a broader variety of selected statistical models with less focus on one best, whereas the GCM selects a few statistical models with stronger focus on one best model. Furthermore, the AIC often prefers statistical models including interaction terms, even for additive or heterogeneity models. Using the AIC, GCM reflect the underlying biological model much better than LRM.

here Tables 2, 3

5 Discussion

In our simulation study, we investigated whether graphical chain models are able to detect underlying biological models of gene-gene interaction in association studies. It should be noted that the simulated biological models are unrealistically simplistic and thus of limited value with respect to the understanding of the underlying complex

biological pathway, as also pointed out by Cordell [4]. But, despite their simplicity, accounting for interaction may increase the power to detect genetic effects (see [4] for practical examples). For this reason, the statistical model used to capture the simplified biological model should be able to reflect the biological interaction by its statistical interaction terms. Unfortunately, like in the study of North et al. [2], our results showed that there is a severe gap between a fitted graphical chain model and the simulated biological model. The simulation results should be interpreted with caution regarding the true biology since the whole simulation study is based on two modeling steps: first, the complex biological pathway is modeled by a simplified biological model and second, a statistical model is used to capture this biological model. Based on the simulation results, it is, therefore, not possible to assess whether a statistical model is able to capture the true biology, but only whether it is able to capture the simplified biological model.

For the Risch models of biological interaction, it can be seen that the additive and heterogeneity models can be reflected with both methods, logistic regression models and graphical chain models. The multiplicative model, however, cannot be identified by either of these. It is well known that a multiplicative model is equivalent to an additive model on a logarithmic scale, that is, in log-linear models loci A and B are modeled as statistically independent additive effects on the logarithmic scale. Since interaction effects are present on the penetrance scale, logistic regression models and graphical chain models that are based on log-linear models will have difficulties to find these interaction effects when analyzing multiplicative penetrance models. In addition, if all penetrances are small, the approximation $a_i + b_j = \log(f_{ij}/(1 - f_{ij})) \approx \log(f_{ij})$ and so $f_{ij} \approx \exp(a_i) \cdot \exp(b_j) = \tilde{a}_i \tilde{b}_j$ indicates that an additive model for the log odds should be equivalent to a multiplicative model for the penetrances (see [15]). In general, the LRM frequently decides for the more parsimonious additive effect models, whereas the GCM prefers dominance effect models to additive effect models. Surprisingly, if a dominance effect is explicitly included in the data, as is the case in the biological Het_H model, the LRM detects this effect while the GCM does not. The reason for this behavior is not well understood.

For the epistasis models with their underlying biological interaction, statistical interaction is only found for one scenario. Especially, if a recessive locus is involved, both approaches fail to detect interaction effects. In this case, the epistasis models are often falsely classified as mean models, which leads to the misinterpretation that none of the analyzed loci has an effect on the disease. The results from the epistasis models are therefore not convincing for both approaches.

The question is, whether this finding was due to a lack of statistical power. Especially, when using the conservative BIC, the number of observations ($N = 1000$) may be too small to identify epistasis models. The AIC might be more appropriate in this case. Since the results of BIC and AIC are not concordant, the choice of the appropriate information criterion is crucial. Several authors have compared both criteria [14,16,17,18]. The BIC is based on the assumption that a “true” model exists, although it does not have to be in the set of candidate models. Moreover, the BIC is a dimension-

consistent criterion which means that it searches for a true model that is independent from the sample size. In contrast, the AIC does not intend to select any “true” model, but selects the best model within a given set of candidate models. As non dimension-consistent information criterion this includes that the selected best model varies with the sample size. According to Hurvich and Tsai [16], if the truth is infinite-dimensional, the AIC selects the best finite-dimensional approximating model in large samples. We decided for the BIC since we have simulated data from underlying biological models with given information about presence or absence of gene-gene interaction.

It has been shown that analyzing case-control studies using graphical chain models or logistic regression models fails in detecting interaction on a penetrance scale if the model assessment is based on AIC or BIC. Hence, other approaches may be more appropriate as for instance nonparametric methods like the multifactor dimensionality reduction [19,20]. Another alternative would be not to model interaction but to measure it directly. For this purpose, we will develop an interaction measure on the penetrance scale where directed acyclic graphs have demonstrated to be a good starting point.

Acknowledgment

We gratefully acknowledge the financial support of this research by the grant PI 345/2-1 from the German Research Foundation (DFG). We also thank two anonymous referees for their helpful comments.

References

- [1] Vieland VJ, Huang J: Two-locus heterogeneity cannot be distinguished from two-locus epistasis on the basis of affected-sib-pair data. *Am J Hum Genet* 2003;73:223-232.
- [2] North BV, Curtis D, Sham PC: Application of logistic regression to case-control association studies involving two causative loci. *Hum Hered* 2005;56:79-87.
- [3] Didelez V, Pigeot I, Dean K, Wister A: A comparative analysis of graphical interaction and logistic regression modelling: self-care and coping with a chronic illness in later life. *Biometrical J* 2002;44:410-432.
- [4] Cordell HJ: Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* 2002;11:2463-2468.
- [5] Bateson W: *Mendel's Principles of Heredity*. Cambridge, Cambridge University Press, 1909.
- [6] Fisher RA: The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edin* 1918;52:399-433.
- [7] Risch N: Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet* 1990;46:222-228.
- [8] Hodge S: Some epistatic two-locus models of disease. I. Relative risks and identity by descent distributions in affected sib pairs. *Am J Hum Genet* 1981;33:382-395.
- [9] Howson JM, Barratt BJ, Todd JA, Cordell HJ: Comparison of population and family-based methods for genetic association analysis in the presence of interacting loci. *Genet Epidemiol* 2005;29:51-67.
- [10] Cordell HJ, Todd JA, Bennett ST, Kawagushi Y, Farrall M: Two-locus maximum lod score analysis of a multifactorial trait: joint consideration of IDDM2 and IDDM4 with IDDM1 in type 1 diabetes. *Am J Hum Genet* 1995;57:920-934.
- [11] Lauritzen SL: *Graphical Models*. Oxford, Clarendon Press, 1996.
- [12] Edwards D: *Introduction to Graphical Modelling*, ed 2. New York, Springer, 2000.
- [13] Højsgaard S: The mimR package for graphical modelling in R. *J Stat Softw* 2004;11:21-72.
- [14] Burnham KP, Anderson DR: *Model Selection and Multimodel Inference*, ed 2. New York, Springer, 2002.
- [15] Cordell HJ, Todd JA, Hill NJ, Lord CJ, Lyons PA, Peterson LB, Wicker LS, Clayton

DG: Statistical modeling of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type 1 diabetes. *Genetics* 2001;158:357-367.

[16] Hurvich CM, Tsai C-L: Autoregressive model selection in small samples using a bias-corrected version of AIC; in Bozdogan H (ed): *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*. Dordrecht, Kluwer Academic, 1994, vol 1, pp 137-157.

[17] Kuha J: AIC and BIC. Comparisons of assumptions and performance. *Sociol Method Res* 2004a;33:188-229.

[18] Kuha J: AIC and BIC. Comparisons of assumptions and performance (Publisher's errata). *Sociol Method Res* 2004b;33:417-418.

[19] Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001;69:138-147.

[20] Hahn LW, Ritchie MD, Moore JH: Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interaction. *Bioinformatics* 2003;19:376-382.

Appendix

here Table 5

Table 1

Stat. model	Covariates used in logistic regression model								Graphical chain model	Interaction effects
S_M									here Figure <GM_Mean.eps>	no gene gene-interaction
S_{A1}	x_1								here Figure <GM_Add1.eps>	
S_{A2}		x_2							here Figure <GM_Add2.eps>	
S_{A12}	x_1	x_2							here Figure <GM_Add.eps>	
S_{D1}	x_1		z_1						here Figure <GM_Dom1.eps>	
S_{D2}		x_2	z_2						here Figure <GM_Dom2.eps>	
S_{D12}	x_1	x_2	z_1	z_2					here Figure <GM_Dom.eps>	gene-gene interaction
S_{I1}	x_1	x_2	z_1	z_2	x_1x_2				here Figure <GM_Int1.eps>	
S_{I2}	x_1	x_2	z_1	z_2	x_1x_2	x_1z_2	x_2z_1		here Figure <GM_Int2.eps>	
S_{I3}	x_1	x_2	z_1	z_2	x_1x_2	x_1z_2	x_2z_1	z_1z_2	here Figure <GM_Int3.eps>	

Table 2

			Percentage of selected statistical models with $\Delta_m^{BIC} \leq 2$ (%)									
				additive effect models			dominance effect models			interaction effect models		
Biological models	Method	Total	S_M	S_{A1}	S_{A2}	S_{A12}	S_{D1}	S_{D2}	S_{D12}	S_{I1}	S_{I2}	S_{I3}
Add_L	LRM	107	4	0	92	1	0	4	0	0	0	0
	GCM	107	7	0	6	0	0	86	1	0	0	0
Add_H	LRM	135	0	25	0	19	50	0	6	1	0	0
	GCM	101	0	0	0	0	98	0	2	0	0	0
Het_L	LRM	123	0	75	0	16	9	0	0	0	0	0
	GCM	101	0	0	0	0	97	0	3	0	0	0
Het_H	LRM	110	0	0	0	0	0	89	11	0	0	0
	GCM	114	0	0	82	7	0	11	0	0	0	0
$Mult_L$	LRM	110	0	0	5	90	0	0	5	1	0	0
	GCM	105	0	0	0	0	0	8	92	0	0	0
$Mult_H$	LRM	115	0	0	0	21	0	0	75	4	0	0
	GCM	100	0	0	0	0	0	0	100	0	0	0
Epi_L^{dd}	LRM	146	59	16	17	3	2	1	0	1	0	0
	GCM	114	83	5	4	0	4	4	0	0	0	0
Epi_H^{dd}	LRM	101	0	0	0	0	0	0	0	99	1	0
	GCM	100	0	0	0	0	0	0	0	0	0	100
Epi_L^{rd}	LRM	111	87	6	1	1	5	0	0	0	0	0
	GCM	105	93	7	0	0	0	0	0	0	0	0
Epi_H^{rd}	LRM	100	0	0	0	0	100	0	0	0	0	0
	GCM	118	0	75	0	1	24	0	0	0	0	0
Epi_L^{rr}	LRM	106	94	3	3	0	0	0	0	0	0	0
	GCM	100	100	0	0	0	0	0	0	0	0	0
Epi_H^{rr}	LRM	109	87	6	1	0	2	5	0	0	0	0
	GCM	105	91	2	5	0	0	2	0	0	0	0

Table 3

			Percentage of selected statistical models with $\Delta_m^{AIC} \leq 2$ (%)									
				additive effect models			dominance effect models			interaction effect models		
	Method	Total	S_M	S_{A1}	S_{A2}	S_{A12}	S_{D1}	S_{D2}	S_{D12}	S_{I1}	S_{I2}	S_{I3}
Add_L	LRM	295	0	0	28	28	0	28	5	4	3	3
	GCM	122	0	0	2	0	0	73	18	0	1	7
Add_H	LRM	256	0	3	0	3	14	0	19	25	13	11
	GCM	142	0	0	0	0	37	0	23	0	12	27
Het_L	LRM	285	0	19	0	26	21	0	9	12	5	4
	GCM	126	0	0	0	0	53	0	36	0	4	7
Het_H	LRM	216	0	0	0	0	0	18	26	21	5	3
	GCM	202	0	0	23	29	0	26	17	1	1	3
$Mult_L$	LRM	204	0	0	0	19	0	0	21	19	5	4
	GCM	111	0	0	0	0	0	0	85	0	4	12
$Mult_H$	LRM	213	0	0	0	1	0	0	31	33	7	2
	GCM	105	0	0	0	0	0	0	90	0	4	7
Epi_L^{dd}	LRM	278	7	11	11	15	8	7	0	23	7	4
	GCM	155	7	5	5	0	6	8	5	3	18	45
Epi_H^{dd}	LRM	163	0	0	0	0	0	0	0	29	16	10
	GCM	100	0	0	0	0	0	0	0	0	0	100
Epi_L^{rd}	LRM	348	23	22	17	10	26	6	5	4	3	2
	GCM	246	27	28	5	5	21	4	2	1	2	3
Epi_H^{rd}	LRM	166	0	0	0	0	31	0	10	8	6	2
	GCM	182	0	18	0	9	24	0	9	12	15	13
Epi_L^{rr}	LRM	355	27	27	25	10	11	10	3	2	3	3
	GCM	193	41	13	12	3	12	10	4	3	1	0
Epi_H^{rr}	LRM	330	18	16	14	5	15	13	3	7	12	9
	GCM	131	6	3	2	0	2	4	1	65	8	10

Table 4

	Type I error			
Stat. model	0.05		0.01	
	LRM	GCM	LRM	GCM
S_{A2}	43	53	8	10
S_{A12}	58	53	5	10
S_{D2}	19	53	0	10
S_{D12}	31	53	1	10
S_{Int1}	33	33	4	1
S_{Int2}	19	67	4	10
S_{Int3}	9	66	3	10

Table 5

Biological model	Relative penetrance matrix			Locus	Marginal GRR		
					0	1	2
Add_L	1	2	3.9	A	1	1	1.2
	1.1	2.1	4	B	1	2	3.9
	1.2	2.2	4.1				
Add_H	1	1.8	2.7	A	1	17	33.8
	19.7	20.5	21.4	B	1	1.2	1.4
	39.3	40.2	41				
Het_L	1	1.3	1.8	A	1	3.7	7.1
	3.9	4.1	4.7	B	1	1.2	1.5
	7.5	7.7	8.3				
Het_H	1	1	16	A	1	1	2.5
	1	1	16	B	1	1	16
	3	3	18				
$Mult_L$	1	5	10	A	1	2	4
	2	10	20	B	1	5	10
	4	20	40				
$Mult_H$	1	50	100	A	1	20	40
	20	1000	2000	B	1	50	100
	40	2000	4000				
Epi_L^{dd}	1	1	1	A	1	1.2	1.8
	1	2	4	B	1	1.2	1.6
	1	5	10				
Epi_H^{dd}	1	1	1	A	1	5	11
	1	20	40	B	1	5	9
	1	50	100				
Epi_L^{rd}	1	1	1	A	1	1	2.4
	1	1	1	B	1	1	1.1
	2	4	10				
Epi_H^{rd}	1	1	1	A	1	1	24.4
	1	1	1	B	1	1.2	1.7
	20	40	80				
Epi_L^{rr}	1	1	1	A	1	1	1.1
	1	1	1	B	1	1	1.1
	1	1	10				
Epi_H^{rr}	1	1	1	A	1	1	2
	1	1	1	B	1	1	2
	1	1	100				

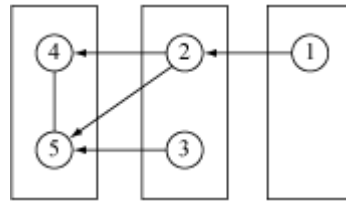


Figure 1

File name: <CG.eps>

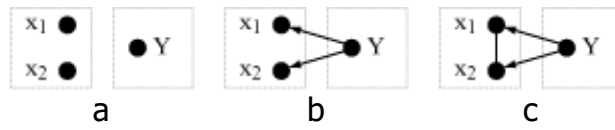


Figure 2

File names:

<GM-LRG1.eps>

<GM-LRG2.eps>

<GM-LRG3.eps>

Unnumbered figures used in Table 1:

<GM_Mean.eps>	
<GM_Add1.eps>	
<GM_Add2.eps>	
<GM_Add.eps>	
<GM_Dom1.eps>	
<GM_Dom2.eps>	
<GM_Dom.eps>	
<GM_Int1.eps>	
<GM_Int2.eps>	
<GM_Int3.eps>	

Table 1	Set of statistical candidate models investigated in the simulation study.
Table 2	Resulting BIC-values of the simulation study. Each biological model (see Table 5) has been simulated under low and high risk 100 times. LRM = logistic regression model, GCM = graphical chain model, Total = number of models satisfying $\Delta_m^{BIC} \leq 2$. Since more than one model may fulfill this criterion in each replication the total number of selected models may exceed the number 100 of model replications. Bold numbers indicate the maximum. Gray shadowed cells mark "true" models, that are statistical models assumed to reflect best the biological mechanism.
Table 3	Resulting AIC-values of the simulation study. See Table 2 for further explanation.
Table 4	Type I error for the null model. The table shows the number of falsely rejected null hypothesis of 1000 replications for the relevant statistical models.
Table 5	Simulated biological models with constant prevalence ($prev=0.01$), disease allele frequencies $p_A = p_B = 0.01$ and given risk increase ρ . The models Add, Het, Mult and $Epi^{dd}, Epi^{rd}, Epi^{rr}$ are described in Section 2.1. The additional index letters L and H symbolizes low and high risk increase.

Figure 1	Graphical chain model with following conditional independences $X_1 \perp\!\!\!\perp X_3 \{X_2\}$, $X_1 \perp\!\!\!\perp X_4 \{X_2, X_3, X_5\}$, $X_1 \perp\!\!\!\perp X_5 \{X_2, X_3, X_4\}$, $X_2 \perp\!\!\!\perp X_3 \{X_1\}$, $X_3 \perp\!\!\!\perp X_4 \{X_1, X_2, X_5\}$ of a five-dimensional random vector X .
Subfigure 2(a)	Mean model
Subfigure 2(b)	Main effect model
Subfigure 2(c)	Interaction effect model
Figure 2	The independence graph for three different parameterizations.