Di Serio, Vicard:

Graphical chain models for the analysis of complex genetic diseases: an application to hypertension

Sonderforschungsbereich 386, Paper 288 (2002)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Graphical chain models for the analysis of complex genetic diseases: an application to hypertension

C. Di Serio[1], P. Vicard[2]

[1]Università "Vita-Salute" San Raffaele, Milan, Italy
[2]Università Roma Tre, Rome, Italy

Abstract

A crucial task in modern genetic medicine is the understanding of complex genetic diseases. The main complicating features are that a combination of genetic and environmental risk factors is involved, and the phenotype of interest may be complex. Traditional statistical techniques based on lod-scores fail when the disease is no longer monogenic and the underlying disease transmission model is not defined. Different kinds of association tests have been proved to be an appropriate and powerful statistical tool to detect a "candidate gene" for a complex disorder. However, statistical techniques able to investigate direct and indirect influences among phenotypes, genotypes and environmental risk factors, are required to analyse the association structure of complex diseases. In this paper we propose graphical models as a natural tool to analyse the multifactorial structure of complex genetic diseases. An application of this model to primary hypertension data set is illustrated.

Keywords: complex disorders, conditional independence, genotype, graphical chain models, phenotype.

1. Statistical problems and strategies in approaching multifactorial genetics disease

Modelling genetic diseases is in general a hard task. It is difficult to make inference on the expressions of a quantitative observed trait describing the disease (phenotype) starting from the genetic information (genotypes). This is particularly true when dealing with multifactorial genetic diseases. Indeed, genetic diseases can be roughly divided into two main categories: *Mendelian* and *multifactorial* genetic disorders. Mendelian disorders (e.g. Down syndrome, Huntington disease) are rare and mainly monogenic, meaning that the disease is due to a single gene mutation. The phenotype for these diseases can be clearly identified and the distinction between affected and unaffected population is clear-cut. Mutations are rare and recent so that a causal gene-disease transmission mechanism can be identified. This means that it is not appropriate to talk about *genetic predisposition* of individuals: either they carry the deleterious mutation and will become ill, or they do not. Linkage analysis based on lod-score tests is the main statistical tool to detect genetic risk factors in these disorders. For further details we refer to Ott (1991).
In this paper we focus on modelling *multifactorial* genetic diseases. Examples of complex diseases are hypertension, multiple sclerosis, schizophrenia, diabetes and other common disorders. They are far more common than Mendelian disorders, and there is no defined pattern of segregation in families. In other words no clear cut-off between affected and unaffected is identified in the population. This leads to a misleading definition of the phenotype which is a primary difficulty for making inference in these settings. These diseases are called multifactorial because their *complex* structure involves a combination of genetic and environmental risk factors, where the *genetic predisposition* is due to the lack of a clear gene-disease transmission mechanism. They are not transmitted but only *promoted* by a collection of factors some of which are hereditary.
Specifically in the paper we address the following issues:

i) defining the complex phenotype;
ii) identifying genetic risk factors;

iii) modelling their role;
iv) including interactions between candidate genes (epistatic interactions) as well as between genes and environmental factors.

Linkage analysis (Clarget-Darpoux, 1998) may help to indicate a *promoter* region where a "candidate gene" (functionally related to the disorder) lies. Association studies are commonly used (Risch, 1990, Bickeboeller and Clarget-Darpoux, 1995) to investigate a possible association link between the candidate gene and the multifactorial disease. However, a (2×2) contingency table approach is not appropriate to identify direct and indirect influences of both etiological predisposing factors as well as the association structure among environmental and genetic determinants.

Here we propose the use of graphical chain models to analyse multifactorial genetic diseases. Graphical chain models are probability models for multivariate random observations whose independence structure is encoded in a graph able to represent both associative and causal relations. These models are already widely used in social sciences as efficient tools to analyse observational studies (Pigeot et al. 2000). In the paper we will show their suitability in evaluating the complex association structure among prognostic factors in a complex diseases framework.

We start our analysis by illustrating how graphical chain models methodology is consistent with the genetic representation. The search for the pathogenetic disease mechanism (Figure 1) starts from the top of complexity, at whole organism level, down to the DNA level. The disease phenotype is separated into several intermediate phenotypes at different biological organisation levels.

The genetic mechanism in Figure 1 indicates that:

- environmental factors and polygenes describe a large number of potential background influences.
- The final phenotypes can be identified as the mixed multivariate final responses.
- Intermediate phenotypes act as variables mediating between the final phenotypes and the potential background influences.

This hierarchy leads to identify a recursive structure in the genotype-phenotype chain which can be expressed in statistical terms. A first crude approximation to the biological chain in Figure1 is shown in Figure 2.

(Figure 1 here)
(Figure 2 here)

The paper is structured as it follows. In Section 2 an introduction to graphs and graphical chain models is given. In Section 3 background on the relevant variables of the hypertension study is introduced. A preliminary analysis is also provided. The graphical chain model for hypertension is derived and shown in Section 4. Results and conclusions are presented in Sections 5 and 6, respectively.

2. Graphical chain models

The framework introduced above illustrates a complex situation where the power of the graphical representation can be exploited. In multifactorial genetic diseases phenotypic, genetic, environmental and demographic variables are all involved.

Graphical chain models (see Cox and Wermuth, 1996) are able to describe all these variables and their dependence structure with a single graph.

A *graph* is a pair *G=(V, E)*, where *V* is the set of vertices, $V = \{1, \ldots, q\}$, and *E* is the set of edges, *i.e.* a subset of $V \times V$ of ordered pairs of distinct vertices, $E \subseteq V \times V$. Consider two vertices *v* and *w* connected by an edge. If both (*v*, *w*) and (*w*, *v*) are in *E*, the edge is undirected (and it is represented with a line); if (*v*, *w*) is in *E* but (*w*, *v*) $\notin$ *E*, the edge is directed (and it is represented with an arrow pointing to *w* from *v*).
A *chain graph* contains both directed and undirected edges. Moreover the vertex set *V* of a chain graph can be partitioned into ordered blocks (or chain components) $B_1, \ldots, B_k$ such that all edges between nodes in the same block are undirected, and all edges between nodes in different blocks, say B_i and B_j with $i<j$, are arrows pointing from a node in B_i to a node in B_j. Notice that all arrows between any two components must have the same direction, e.g. see Figure 3.

(Figure 3 here)

In the graphical model setting, the nodes represent random variables. In a chain graph the variables are arranged in ordered blocks. Variables in the same block are considered on an equal footing and their association structure is taken to be symmetric. When two variables belong to different blocks, the variable in the lower-numbered block is considered logically antecedent (or "causal", roughly speaking) to the variable in the higher-numbered block. For example in Figure 3, $X_\delta \in B_2$ is logically antecedent to $X_\beta \in B_3$. Since chain graphs contain both directed and undirected edges they represent at the same time and by means of one single picture the association structure and the "causal" relations. Notice that in this paper the adjective "causal" is used mainly to qualify an explanatory-response variable relation.
When both discrete and continuous variables are analysed – as in this paper (see Figures 1 and 2) – the vertices of the graphs are partitioned in two groups denoted with Δ and Γ so that $V=\Delta \cup \Gamma$ with $\Delta \cap \Gamma = \varnothing$. These graphs are named *marked* graphs. The set of discrete vertices is denoted with Δ and the set of continuous nodes is denoted with Γ. Discrete nodes are represented by dots and continuous nodes by circles. Figure 3 is an example of a marked graph.
A crucial point in the use of graphical models is the possibility to describe and to read independencies (marginal and/or conditional) from the graph. The relevant information is contained in the graph without boxes, which specifies the part of the statistical model providing the set of independencies. In particular, the absence of an edge between two nodes, *v* and *w*, in the same block B_j, or an arrow missing from a node *v* in B_i to a node *w* in B_j, $i<j$, implies that the associated variables are independent conditionally on all the variables in the blocks $B_1, \ldots, B_j$ except *v* and *w*. Formally we write $v \perp w \mid B_1 \cup \ldots \cup B_j \setminus \{v, w\}$ (following the notation of Dawid, 1979). For example in Figure 3, the arrow missing from $X_\delta \in B_2$ to $X_\alpha \in B_3$ means that X_δ and X_α are independent given all the other variables in B_3, B_2 and B_1, $X_\alpha \perp X_\delta \mid B_1 \cup B_2 \cup B_3 \setminus \{X_\alpha, X_\delta\}$. This property is known as the *pairwise Markov property for chain graphs*. Other Markov properties for chain graphs have been defined (Frydenberg, 1990) which are equivalent when the joint distribution is positive. For a detailed and rigorous account on this, see Lauritzen (1996).
According to a proposed block structure, the joint density function can be recursively factorised. So, given the ordered blocks $B_1, \ldots, B_k$, $f(x) = f(x_{B_1}, x_{B_2}, \ldots, x_{B_k})$ factorises as

$$f(x) = \prod_{t=1}^{k} f\left(x_{B_t} \middle| x_{B_1}, x_{B_2}, \ldots, x_{B_{t-1}}\right) \qquad (1)$$

where $f\left(x_{B_t} \middle| x_{B_1}, x_{B_2}, \ldots, x_{B_{t-1}}\right)$ is the conditional distribution of the *t*-th block given the first (*t*-1) blocks. This distribution can be further simplified exploiting the conditional independence properties embodied in the graph.
A *chain graph model* is a family of multivariate distributions satisfying any of the chain graph Markov properties embodied in the graph. Regarding the distributional assumptions, multivariate response models for mixed variables are needed. The distribution should also be positive to guarantee equivalence of the Markov properties. The *conditional Gaussian* (CG) distribution (see Wermuth and Lauritzen, 1990 and Lauritzen, 1996) satisfies the above requirements. It can be used except when there are discrete responses and continuous covariates. In this case we can use the *CG-regression* model, which describes the dependence of a CG distribution over response variables on explanatory variables. CG regression is the basic distributional element that can be used in the construction of chain graph models. While the Markov properties over chain graphs give a key to read independencies from the given chain graph, associating a CG distribution to the graph provides a connection between zero-valued parameters and absence of certain edges in the graph.

2.1 Graphical chain models and the genotype-phenotype chain

Chain graphs provide a suitable representation of the genotype-phenotype chain (Figure 1). The subject-matter knowledge helps us to determine a block ordering (dependence chain). Starting from the biological chain, we can identify three main blocks: B_1 for the genotypes and socio-demographic and environmental variables (background variables); B_2 for the intermediate phenotypes and other environmental variables (intermediate variables) and B_3 for the phenotypes (response variables). In this sense we see that chain graphs constitute a *natural* tool to represent the genotype-phenotype chain.
To investigate and explore the dependence and independence structure, a joint distribution function is associated to the graph and factorised recursively according to the block structure. For instance, by the pairwise Markov property, we can read from Figure 4 that Gen1 is not informative for the final phenotype *FP2*, once the intermediate phenotype (*IP*), environmental factors (*Env*) and Gen2 are known. We have $FP2 \perp Gen1 | (Env, Gen2, IP, FP1)$.

(Figure 4 about here)

According to the block ordering in Figure 4, the joint probability function can be factorised as follows

$$f(FP1, FP2, IP, Gen1, Gen2, Env) =$$
$$= f(FP1, FP2 \mid IP, Gen1, Gen2, Env) \cdot f(IP \mid Gen1, Gen2, Env) \cdot f(Gen1, Gen2, Env).$$

By the recursive factorisation property a structurally complex problem is split into computationally simpler subproblems, which can be analysed separately using appropriate well-established methodologies. The resulting dependence model is equivalent to the one obtained from the joint distribution. This procedure leads to a gain in efficiency. Thus graphical chain models are also a *useful* tool for the analysis of complex disease association structure.
To provide a statistical representation of the association structure, a dependence chain can be postulated starting from a biological chain. Then an *inferential engine* can be related to this representation to explain and obtain information about the data generating

process. Using the Cox-Wermuth strategy (Cox and Wermuth, 1996) we: 1) discover statistically significant “causal” relations, interactions and associations, *i.e.* we identify the arrows and edges to be included in the graph; 2) propose a complex quantitative trait predictor from discrete and continuous measurements of intermediate phenotypes, environmental factors and socio-demographic variables; 3) represent a number of indirect paths to the response of primary interest.
A causal interpretation of graphical chain models applied to the complex genetic diseases although natural, nevertheless is not automatic and can be misleading. In chain graphs, directed edges are interpreted as causal associations whereas undirected edges represent non-causal association. An ambiguous interpretation of the graph however may arise due to the different nature of non-causal associations. There are situations where an undirected edge is needed for an association between two variables thought to be causal but in which the causality direction is unknown. For a detailed discussion about causality in chain graphs we refer to Lauritzen and Richardson (2002). Furthermore, an explanatory-response variable causal relation may be appropriate only if intervention is allowed and the variables can be manipulated (as for instance in clinical trials). In this paper we fit the graphical chain model to an observational cross-sectional study. According to Pearl (2000) and Edwards (2000), the effects of interventions are difficult to account for in observational studies due to potentially unobserved confounders. These considerations lead to interpret every causal conclusion of the analysis as explorative results for further analyses.
The Cox-Wermuth strategy explores conditional relations by means of a series of univariate regressions instead of multivariate CG-regressions, reducing the complexity of the analysis. A drawback is that it does not necessarily fit a CG distribution so that one has to be careful in interpreting missing edges as conditional independence statements. Thus, the conclusions in terms of conditional independencies suggested by the Cox-Wermuth strategy used in the following analyses, have to be read in exploratory terms only.
The TM algorithm (Edwards and Lauritzen, 2001) has been proposed to fit CG-regression models. It is, however, computationally intensive for multivariate mixed response models.

3. Essential hypertension as a complex disease

Essential hypertension and blood pressure regulation are complex and multifactorial (for clinical details see Cusi and Bianchi, 1998). More than 40 years of epidemiological studies have identified different environmental factors associated with the development of essential hypertension (age, diet, exercise and stress). While much is understood about environmental factors, the genetic factors are still largely unknown. Genetic analysis of essential hypertension, as that of many other complex diseases suffers from three main complications:

a) each gene may have only a small quantitative effect on the disorder.

b) It is likely that essential hypertension is genetically heterogeneous. Different forms of essential hypertension share only the same final phenotype, but have different pathogenetic mechanism.

c) Epistasis, or gene interactions, is very likely to be present in essential hypertension and is an aspect of what is called context dependency.

In this analysis we deal with cross-sectional observations of 285 patients, equally distributed over sex, randomly selected from a group of 44 to 64-years old Vobarno population (n=8000). Phenotypic data, information on environmental factors, and DNA

were obtained in this sample; clinical and biomedical investigations were performed on the relationships among phenotypes and certain candidate genes (Castellano et al. 1995). Within these patients some are treated (indicated by variable "*Therapy*" equal to 1) for being likely to develop hypertension and other related pathologies (ischemic heart disease, hyperlipidemy). Those who undergo a treatment are classified as "at risk" for hypertension. The other patients (with "*Therapy*" = 0) either are not thought to be at risk and do not need to be treated or are not compatible with the treatment. The variable *Therapy* can then be considered as an indicator for the hypertensive risk set. Thus, it can be used as outcome to investigate the effect of some biological indicators on the probability of becoming at risk for hypertension.
Family clinical history is accounted for by means of the following indicator variables representing presence or absence in the family of: stroke (*FamSTR*); hypertension (*FamHYP*); mellitus diabetes (*FamDIA*); dyslipidemia (*FamDYS*). Even though these variables can provide some genetic information, they may also be associated with environmental factors.
The construction of the graphical chain model is done following the Cox and Wermuth variable selection strategy (1996) whose first step consists in postulating a dependence chain.
Before postulating this chain, the relevant variables are selected from the complete set of observed variables. Preliminary multivariate analyses are performed in order to determine phenotypes and genotypes of interest. The resulting variables will represent the vertices of the chain graph, as described in Section 2. Boxes including those variables, which can be taken as being on an equal footing, define the chain components.

3.1 Preliminary screening

Thirteen metric variables have been analysed to screen the most important quantitative traits. A hierarchical cluster analysis was performed on them to reduce heterogeneity in the data. The best results[1] based on six traits partitioned the whole data set into two clusters. The selected phenotypes are: *BMI*, systolic blood pressure (*SBP*), plasma glucose (glycaemia level as indicator, denoted with *Glyc*), triglycerides (*Trig*), cholesterol (*Chol*), uric acid (*Uric*).
A biological explanation consistent with the cluster analysis results can be suggested: a high *BMI* may be associated with an insulin resistance phenomenon that produces an increase in cholesterol, plasma glucose, uric acid and triglycerides. As a consequence, an increased activity in the sympathetic nervous system is related to higher blood pressure level. This phenomenon is known as **insulin resistance mechanism**. Its importance will be shown later.
The excluded variables are quantifications of heart frequency, pulsations, ventricular mass and alternative measurements of body mass. They have been classified as redundant since they provide similar information as the included traits. Therefore they do not play the role of confounders for further analyses. Genetic literature (Johnson *et al*., 2003) supports the results of this preliminary cluster analysis since the selected phenotypes are proved to be those more related to a genetic basis.
The univariate statistics for the clustered traits are shown in Table 1.

[1] the silhouette index (Kaufman and Rousseeuw, 1990) has been used as a goodness of clustering indicator.

Table 1 Univariate statistics for the clustered traits

Descriptive statistics

Variable	Minimum	Maximum	Mean	Std. Deviation
BMI	18.75	40.18	26.0275	3.2823
SBP	95.00	166.00	123.9846	11.8724
Glyc	56.00	172.00	93.4089	14.6304
Chol	121.00	364.00	227.1227	42.8741
Trig	60.00	600.00	117.7584	74.2543
Uric	2.60	10.30	5.4184	1.2982

The data set includes information on eleven different genotypes considered as possible *candidate genes*. In *candidate gene* studies a starting point is to detect whether one or more genes can be considered as "natural" candidates for a complex disease, *i.e.* functionally related to the disease. A factor analysis is performed to identify eventual groups of polyphormisms which are likely to act together in regulating certain functions. This analysis is also useful to evaluate the explicative contribution of the analysed genes in terms of explained variation (in population studies the contribution ranges from 30% to 65% of the total variation, Cavalli Sforza and Bodmer, 1973).
The results of the factor analysis, shown in Table 2, will be useful to interpret the final chain. Three main factors are identified explaining 30% of the total genetic variance. They can be interpreted within a correspondent genetic "partition". Factor 1 and Factor 2 will be called "adrenergic receptor factors" (ARF) given that the factor components are mainly six single nucleotide polymorphisms (*Snps1*, *Snps2*, *Snps3*, *Snps4*, *Snps5*, *Snps6*) of the adrenergic receptors genes. Specifically, *Snps3* and *Snps2* play a relevant role with respect to ARF1; their loadings on ARF1 are 0.987 and –0.568, respectively. Since these Snps are two polymorphisms on the so called β-adrenergic receptor which alone does not commonly show a main effect on hypertension, the results of the factor analysis would suggest a synergistic effect of two functional variants of the gene acting in opposite directions. In simple words, *Snps3* could act as an increasing mutation risk for final or intermediate hypertension phenotype and *Snps2* as detrimental. This is consistent with the literature (Bengtsson et al., 2001) since the two mutations are related to different pathogenic mechanisms. In addition the important role played by *Snps6* on Factor 2 (loading on ARF2 = 0.992), a polymorphisms on α-adrenoreceptor, highlights the well-known association of *Snps6* with hypertension through dyslipidemia and cholesterol. The third factor (Raas factor) is mainly related to single nucleotide polymorphisms of the renin angiotensin aldosterone system genes (Raas). This could be of particular interest because Raas genes regulate the sympathetic nervous system.

Table 2 Results of the factor Analysis

	ARF1	ARF2	RAAS
Adducine			0.139
Snps1	-0.185		
Snps2	-0.568		
Snps3	0.987	-0.136	
Snps4	-0.110		
Snps5		0.518	
Snps6	0.127	0.992	
Raas1	-0.107		0.991
Raas2			
Raas3			
Raas4			0.424

The genotype distributions within the sample are given in Table 5 (Appendix 1). The three levels indicate the more frequent (wild type) homozygous (level 0), the heterozygous (level 1) and the less frequent (mutated) homozygous (level 3) respectively. For details on the genotype definitions see Castellano, Di Serio *et al*. (2002).

4. A graphical chain model for essential hypertension

We now provide a graphical model representation of the genotype-phenotype chain dependence structure starting from the hypertension data set illustrated above. The analysis is articulated in the following four main steps of the Cox-Wermuth strategy:

- postulation of a dependence chain
- screening for interaction and non-linearities
- system of univariate regressions
- variable selection strategy for one univariate regression.

The statistical analysis is performed with the software GraphFitI[2] (Blauth *et al.*, 2000) designed to fit a graphical model to a multivariate data set. The software implements the Cox-Wermuth strategy and it performs a preliminary screening of the variables and different types of regression analyses depending on the measurement scale of the data.

4.1 Postulated dependence chain

Postulating the first rough dependence chain is a crucial step for graphical chain model derivation. The first chain is mainly based on both prior knowledge of the biological mechanisms and some preliminary statistical analyses. The researcher's initial idea and subject-matter knowledge play a fundamental role. The performed analysis is confirmatory, *i.e* a block ordering is suggested by experience and background knowledge, while association and causal structures are inferred. Generally, the postulated chain is the sum of qualitative knowledge and quantitative information provided by previous exploratory data analyses. Moreover, in medical genetic contexts researcher's prior information is fundamental for the choice of the genes to be sequenced. If the analysis were purely

[2] Web site for download: http://www.stat.uni-muenchen.de/~blauth/GraphFitI/graphFitI.html. Author: Angelika Blauth.

exploratory, *i.e.* without any subject-matter knowledge, postulating a dependence chain could be inconsistent with the aim of finding the most parsimonious independence model (Lauritzen and Richardson, 2002).
In this paper two alternative chains, representing two different biological hypotheses, are postulated and compared. Two graphical models are then derived together with the respective sets of association and causal relations. The two initial chains are represented in Figure 5.
Reading the chain in Figure 5(a) from left to right (consistently with the candidate gene-phenotype chain in Figure1) three categories of variables can be identified. A first chain component ("pure response") contains the selected phenotypes (*SBP* and *BMI*). A second chain component includes the indicator variable (*Therapy*) for being on chronic drug treatment for high blood pressure. The third group, corresponding to the biological level, is built using plasma glucose (glycaemia level as indicator, *Glyc*), triglycerides (*Trig*), cholesterol (*Chol*), and uric acid (*Uric*) measurements. The fourth chain component includes environmental and genetic risk factors (pure background variables). The genotypes are those analysed in the factor analysis and described in Appendix 1. The family variables have bee illustrated in Section 3.
The second chain, Figure 5(b), differs from that in Figure 5(a) mainly in the role of *BMI* and *Therapy*. In the first postulated chain, a hypothesis is formulated concerning the identification of the final phenotype in terms of the quantitative traits *BMI* and *SBP*. In the second postulated chain systolic blood pressure is the only final quantitative phenotype and *BMI* is considered as a possible explanatory variable affecting the values of the other quantitative traits. This assumption is more consistent with the "insulin resistance" mechanism described in the previous section. Furthermore, the variable *Therapy* is now placed on the same footing as the quantitative intermediate traits. In this way we study how being at risk for hypertension (*Therapy*=1) is associated with other biological determinants rather than the impact of such indicators on *Therapy* itself.
By comparing the two postulated chains we address the following crucial question*: is a high body mass index an indicator of hypertension or a prognostic factor*? An answer to this question can be found through the information gained from biological, genetic and environmental variables about the determinants of the final phenotype.
In the following subsections the screening and estimation procedures will be illustrated in detail for the first postulated chain (Figure 5(a)). We will then give summarised results for the second chain (Figure 5(b)).

(Figure 5(a) and 5(b) here)

4.2 Screening for interactions and non-linearities

This step of the analysis aims at searching significant interaction terms or non-linear influences to be included in the multiple regression analysis. Only those interactions and non-linear relations showing either statistical relevance or biomedical interest are selected. The screening tests (Cox and Wermuth 1994) are based on testing the systematic departure from multivariate normality. To detect significative cross-product terms, the *t*-values from trivariate linear regressions, such as that of a response variable Y on X_i and X_j and X_i*X_j, are examined. In absence of interactions, for large sample sizes, the studentized *t*-statistics approximately follows a standard normal distribution.
The software GraphFitI produces normal probability plots of the expected value of the normal ordered statistic versus the ordered *t* statistics. Under the assumption of no interaction, the points spread along the diagonal. Therefore points far from the diagonal

line (departure points) denote highly significative interaction terms. “Epistatic interactions” between the candidate genes could be suggested performing this analysis.
The screening for non-linearities proceeds likewise. Quadratic terms only are included since Taylor expansions up to the second order are a good approximation tool of non-linear dependency in a large framework. Normal probability plots are drawn to find out eventual departure points denoting significative quadratic effects.
The screening for interactions and non-linearities is performed separately for each chain and involves all the variables. It produces similar graphical plots and identical statistical results for both chains (Figures 6(a), 6(b), 7(a), 7(b)).

(Figure 6(a), 6(b) here).

(Figure 7(a), 7(b) here).

The following interactions and non-linearities, common to both chains, are included as explicative variables in the system of regressions:
1. Interaction between *age* and genotype *Raas2* in affecting *BMI.*
2. Interaction between *age* and *FamDYS* in affecting the *BMI*.
3. Interaction between *age* and *FamDYS* in affecting the *SBP*.
4. Interaction between *FamHYP and FamDIA* in affecting *SBP*
5. Interaction between genotype *Raas3* and *FamDIA* in affecting cholesterol
6. Interaction between genotype *Raas3* and tryglicerides in affecting cholesterol
7. Interaction between genotype *Snps6* and tryglicerides in affecting plasma glucose (*Glyc*)
8. Interaction between genotype *Snps6* and glycaemia in affecting tryglicerides
9. Quadratic effect of triglycerides on both cholesterol and uric acid
10. Quadratic effect of *age* on *SBP*
11. Quadratic effect of plasma glucose *(Glyc)* on tryglicerides

Note that in the screening phase biological, genetic and environmental components are involved.

4.3 System of univariate regressions

This step consists in investigating the form of the conditional distributions by means of separate regression analyses, as implemented in GraphFitI. A system of univariate regressions is performed for each chain component as clarified in Section 2. The model type is related to the scale of the selected response variables, according to the postulated chain. Therefore, a system of normal and logit regressions is performed. This seems to be consistent with complex disease terminology where quantitative response variables correspond to phenotypic traits and qualitative variables are typically environmental or genotype prognostic factors. The graphical model we are fitting is a block-regression graph (Cox and Wermuth, 1996) where each edge connecting two nodes (i,j) concerns a conditional relation between Y_i and Y_j given all the remaining variables ignoring future responses, *i.e.* variables belonging to blocks on the left of the boxes of Y_i and Y_j. In other words, suppose that Y_i is continuous, in order to find the variables directly influencing Y_i, a regression of Y_i on all the variables belonging to its present and past, i.e. contained in the same box or in boxes on its right, is performed.
A forward and backward selection procedure (Caputo et al. 1999) is then performed alternating a “nesting” step, where variables are added to enlarge a minimal model, and a “reducing” step that follows a typical likelihood ratio reduction procedure. The criteria are

based on the change in the F – statistic (Rao, 1995) derived by comparing the R^2 of the full model with R^2 of the reduced model at a 0.05 significance level.
In Table 3 the results of the system of univariate regressions for the first postulated chain are reported. For the second postulated chain, the results concerning the variables in the first three boxes only (reading Figure 5(b) from the left to the right) are reported in Table 4. The results relative to the background variables are not reported; they are identical to those in Table 3 because the corresponding boxes in Figure 5(a) and in Figure 5(b) are defined in the same way. The level of the discrete variables is reported in the tables with "group = level of the variable". For instance being heterozygous for *Raas1*, *i.e. Rass1*=1, is represented by writing *Raas1*(group=1). From Table 3 we read that *Raas1* results as an indirect protective factor for *SBP* (via *Uric* and *BMI*), because it significantly reduces the probability of having high uricemy.

Table 3. Results of the system of univariate regressions for the first postulated chain

Response variable	**Explanatory variables**			
BMI	***SBP***	***Glyc***	***Uric***	***Chol***
Est. Coeff.	0.049	0.055	0.503	0.014
t-value	2.66	3.71	3.09	2.80
SBP	***BMI***	***Therapy(group=0)***		
Est. Coeff.	0.769	-7.670		
t-value	3.30	4.15		
Therapy	***Glyc***	***Age***	***Raas2*(group=0)**	***Raas2*(group=1)**
Est. Coeff.	-2.006	2.567	-171.103	-169.962
t-value	2.94	2.34	151.10	112.98
Therapy	***Glyc*Raas2*(group=0)**	***Glyc*Raas2*(group=1)**		
Est. Coeff.	1.974	1.964		
t-value	2.89	2.88		
Glyc	***Chol***	***Trig***	***Age***	***Snps6*(group=0)**
Est. Coeff.	0.065	0.232	0.344	27.318
t-value	2.77	6.00	3.46	3.62
Glyc	***Snps6*(group=1)**	***FamSTR*(group=0)**	***Trig*Snps6*(group=0)**	***Trig*Snps6*(group=1)**
Est. Coeff.	30.786	4.800	-0.219	-0.241
t-value	3.92	2.39	5.18	5.49
Chol	***Trig***	***Uric***	***Glyc***	
Est. Coeff.	0.459	6.276	0.532	
t-value	6.60	2.97	2.60	
Trig	***Chol***	***Uric***	***Glyc***	***Snps6*(group=0)**
Est. Coeff.	0.575	14.882	4.077	343.264
t-value	5.39	4.35	5.23	3.87
Trig	***Snps6*(group=1)**	***Glyc*Snps6*(group=0)**	***Glyc*Snps6*(group=1)**	
Est. Coeff.	372.277	-3.861	-4.045	
t-value	3.94	4.36	4.26	
Uric	***Trig***	***Raas1*(group=0)**	***Raas1*(group=1)**	***Raas4*(group=0)**
Est. Coeff.	0.010	-0.565	-0.732	0.550
t-value	4.83	2.41	3.36	2.40
Uric	***Raas4*(group=1)**	***Snps3*(group=0)**	***Snps3*(group=1)**	***Raas4*Snps3***
Est. Coeff.	0.834	-0.320	-0.175	-0.236
t-value	3.03	1.13	0.55	2.82
FamHID	***Snps3*(group=0)**	***Snps3*(group=1)**		
Est. Coeff.	1.146	1.229		
t-value	2.37	2.67		
FamSTR	***FamHYP*(group=0)**			
Est. Coeff.	0.806			
t-value	2.65			
FamDYS	***FamDIA*(group=0)**	***FamHYP*(group=0)**	***Adducine*(group=0)**	***Adducine*(group=1)**
Est. Coeff.	1.723	0.386	-24.307	-25.340
t-value	4.14	0.84	83.89	68.01
FamDYS	***FamHYP*Adducine***			
Est. Coeff.	1.396			
t-value	2.23			
FamHYP	***FamSTR*(group=0)**	***FamDYS*(group=0)**		
Est. Coeff.	0.773	1.051		
t-value	2.46	2.67		
FamDIA	***FamDYS*(group=0)**			
Est. Coeff.	1.601			
t-value	4.11			

Raas4	***Raas2***				
	group=0		**group=1**		
	Odds1	Odds2	Odds1	Odds2	
Est. Coeff.	2.171	25.026	-1.796	24.366	
t-value	101.82		84.30		
Raas3	***Gender***				
	Odds1	Odds2			
Est. Coeff.	-0.580	-28.244			
t-value	202.03				
Raas1	***Snps6***				
	group=0		**group=1**		
	Odds1	Odds2	Odds1	Odds2	
Est. Coeff.	-22.167	-22.724	-21.515	-22.271	
t-value	50.30		45.54		
Snps4	***Raas2***				***Snps3***			
	group=0		**group=1**		**group=0**		**group=1**	
	Odds1	Odds2	Odds1	Odds2	Odds1	Odds2	Odds1	Odds2
Est. Coeff.	22.674	-0.317	23.208	-0.333	-2.884	-0.448	-1.716	-0.804
t-value	78.85		62.05		3.13		2.63	
Snps3	***Raas2***				***FamHID***			
	group=0		**group=1**		**group=0**			
	Odds1	Odds2	Odds1	Odds2	Odds1	Odds2		
Est. Coeff.	1.208	24.444	2.203	25.282	1.202	1.260		
t-value	91.19		68.71		2.63			
Adducine	***Raas2***				***Age*RAAS2***			
	group=0		**group=1**		**group=0**		**group=1**	
	Odds1	Odds2	Odds1	Odds2	Odds1	Odds2	Odds1	Odds2
Est. Coeff.	-300.598	35.928	-330.468	11.066	5.023	-0.547	5.504	-0.149
t-value	27.44		41.96		8.28		9.24	
Adducine	***Age***							
	Odds1	Odds2						
Est. Coeff.	-5.540	0.118						
t-value	9.41							

Table 4. Results of the system of univariate regressions for the second postulated chain

Response variable	**Explanatory variables**			
SBP	***Therapy*(group=0)**	***BMI***		
Est. Coeff.	-7.670	0.769		
t-value	4.15	3.30		
Therapy	***Glyc***	***FamHID*(group=0)**	***Raas2*(group=0)**	***Raas2*(group=1)**
Est. Coeff.	-1.917	0.914	-163.198	-161.6044
t-value	2.73	2.47	137.68	98.28
Therapy	***Glyc*Raas2*(group=0)**	***Glyc*Raas2*(group=1)**		
Est. Coeff.	1.883	1.869		
t-value	2.69	2.66		
Glyc	***Therapy(gropu=0)***	***Trig***	***BMI***	***Chol***
Est. Coeff.	50.575	0.178	3.037	-0.064
t-value	2.29	4.58	3.88	2.86

Glyc	***FamHID*(group=0)**	***Snps6*(group=0)**	***Snps6*(group=1)**	***Trig*Snps6*(group=0)**
Est. Coeff.	4.037	26.674	37.902	-0.243
t-value	2.20	4.21	4.86	4.86
Glyc	***Trig*Snps6*(group=1)**	***BMI*Therapy*(group=0)**	***Therapy*Snps6***	
Est. Coeff.	-0.241	-2.382	-5.673	
t-value	5.55	2.85	2.22	
Chol	***Trig***	***Trig^2***	***Uric***	***Glyc***
Est. Coeff.	0.448	-0.001	-7.186	-0.642
t-value	6.50	4.21	3.38	3.53
Chol	***BMI***			
Est. Coeff.	1.978			
t-value	2.38			
Trig	***Glyc***	***Chol***	***Uric***	***Snps6*(group=0)**
Est. Coeff.	4.077	0.575	14.882	343.264
t-value	5.23	5.39	4.35	3.87
Trig	***Snps6*(group=1)**	***Glyc* Snps6*(group=0)**	***Glyc* Snps6*(group=1)**	
Est. Coeff.	372.277	-3.861	-4.045	
t-value	3.94	4.36	4.26	
Uric	***Trig***	***BMI***	***Gender***	***FamHYP*(group=0)**
Est. Coeff.	0.005	0.059	1.147	-0.485
t-value	3.71	2.55	7.54	3.13
Uric	***Raas1*(group=0)**	***Raas1*(group=1)**	***Trig*BMI***	
Est. Coeff.	-0.276	-0.622	-0.001	
t-value	1.31	3.21	3.19	
BMI	***FamSTR*(group=0)**			
Est. Coeff.	1.285			
t-value	2.66			

5 Results

In this section we discuss how, starting from both statistical as well as clinical considerations it is possible to investigate and visualize through the final chain the determinants of hypertension and their dependence structure. The chain graphs shown in Figure 8 and Figure 9 are associated to the postulated chains in Figure 5(a) and 5(b), respectively.
Initially only the highly significant influences and associations (those with *t-value* ≥ 3) were represented. The results of the regression analyses (Table 3 and Table 4) indicated some important genetic relations that might have been excluded with such a strict criterion. Thus in the final graph we also included some links, which are not highly statistically relevant ($2.2 \leq$ *t-value* < 3) but are of great interest from a biological - genetic viewpoint. In Figure 8 and Figure 9 highly significant and weakly significant links are represented by thick lines (*t-value* ≥ 3) and thin lines ($2.2 \leq$ *t-value* < 3) respectively.
Consider the final chain in Figure 8. Notice that, apart from a weak quadratic effect of *age* on *SBP* (found in the initial screening phase shown in Section 4.2), *age* directly acts on the final phenotypes *SBP* and *BMI* through interaction with a genotype or family history. Nevertheless, the results in Table 3 and the final chain in Figure 8 attribute an important role to *age*. It indirectly influences the phenotypes *SBP* and *BMI*. In particular *age* affects *BMI* only indirectly via *Therapy* and glycaemia. For instance *age* has an increasing effect on both plasma glucose level (*Glyc*) and the probability of being treated, *i.e.* at risk for hypertension. This suggests that older patients are more likely to become at risk for hypertension and to have a high *BMI* since the glycaemia level in older patients is usually higher. This is a very important clinical statement.
The relation between *BMI* and *SBP* could have a substantial biological explanation in **hypertensive metabolic syndrome** (see Castellano M., Di Serio C., *et al.*, 2002). Quantitative trait and genetic variable effects are supposedly mediated by increased insulin levels, related to high BMI, that affects systolic blood pressure. The results in Table 3 and in Figure 8 seem to be consistent with this hypothesis. An interpretation of the missing edges in terms of conditional independencies although appealing in this framework has a meaning in exploratory terms only, since the data-driven strategy adopted does not ensure the equivalence of the Markov properties. However, it provides important directions for further steps in the search for candidate genes. From our analyses the hypertensive metabolic syndrome seems to have genetic ground through *BMI* only. Moreover the phenotype *BMI* is influenced by Adrenergic Receptor genes (*Snps3* and *Snps6*) and by the polyphormisms on the renine gene (*Raas1*, and *Raas4*) through uric acid (*Uric*) triglycerides (*Trig*) and glycaemia (*Glyc*). In general, Figure 8 shows that a crucial role is played by intermediate phenotypes, which behave like “genetic” filters in the chain.
Some further remarks about genetic interactions are still needed. Triglycerides and *Snps6* show an interactive protective effect on glycaemia, whereas separately each of these factors have an increasing risk effect on glycaemia. The distributions of *Glyc* and *Trig* conditionally on *Snps6* (see Figure10) provide a possible explanation to this apparent inconsistency. From Figures 10(a) and 10(b) we notice that in the wild type group (*Snps6*=0) *Glyc* has many outliers and extreme values, very likely corresponding to diabetic patients. Being diabetic (mostly concentrated in wild type allele of *Snps6*) is an increasing risk factor for having a high glycaemia level. Thus, the Glycaemia mean level for patients with *Snps6*=0 is higher due to the extreme values. To understand the

interaction between triglicerides and *Snps6*, it is useful to look at the distribution of triglicerides in the patient group with *Snps6*=0 (Figure 10(c)); patients in this group have the lowest values of triglicerides. Since triglicerides and glycaemia are positively correlated (correlation coeff.= 0.424, $p<0.05$), we can deduce that the patients detected by the extreme values in glycaemia (probably diabetic and with *Snps6*=0) are mainly responsible for the revealed detrimental effect of triglicerides on glycemia. We notice that the considerations above suggest to investigate *Snps6* as a possible candidate for diabetes. The chain itself can help not only in identifying some candidate genes but also in retrieving information about epistatic genetic interactions. The interaction term *Raas4*Snps3* shows a protective effect on *Uric Acid* (coef= -0.236, $p < 0.0025$), whereas *Raas4* itself has a detrimental main effect on uric acid and Snps3 is not significative. This suggests that a combined effect of possessing a pathogenic genotype for both polymorphisms has a significant protective impact on *Uric Acid* as compared to the effect of possessing only *Raas4* or *Snps3*. This may be associated to some evidence of functional relevance in activity modulation of both sympathetic nervous system (through ARF) and Raas (renin angiotensin aldosterone system). Furthermore it is interesting to notice that genetic factors affect the intermediate phenotype cholesterol indirectly only, through *Uric Acid* and tryglicerides; this suggests that there is no direct genetic predisposition to high cholesterol. The associations discussed above are only an example of the amount of information and suggestions that the graphical models methodology may provide to genetics.
The results can be summarised as follows:

I. relevance of *age* and its interactions with some quantitative traits in affecting *BMI* and *SBP*;
II. key role of the quantitative traits (in the third block) in connecting the genotypes and environmental variables with the final phenotypes. Interactions among genotypes and intermediate quantitative traits have significant influence on the level of the other intermediate quantitative traits.
III. Identification of epistatic interactions among the involved genes (ARF and Raas).
IV. Family history (*FamSTR*, *FamHYP*, *FamDYS* and *FamDIA*) is a predisposing factor for high values of intermediate traits. For instance, the stroke occurrence in the family directly predisposes to higher plasma glucose levels. Other family history aspects are indirectly informative for the final phenotypes in interaction with the genotypes.

With one single graph (Figure 8) we are able to represent all these influences, associations and interactions within and between the different biological levels.
The second postulated chain (Figure 5(b)) was strongly consistent with the "insulin resistance" mechanism. However the associated final chain (Figure 9) does not lead to as biologically convincing explanation as does the final chain in Figure 8. This is mainly due to the role of *BMI* and to irrelevance of *age*. Indeed, *BMI* is not affected by any of the genetic variables suggesting that there is no genetic predisposition to a high *BMI*, which is quite inconsistent with biological knowledge (see Cusi and Bianchi,1998). Moreover, *age* does not affect *BMI* except through family history variables. This does not have a straightforward biological explanation. Other inconsistencies in Figure 9 with clinical principles are the absence of a direct impact of cholesterol on *SBP*.

(Figure 8 here)

(Figure 9 here)

6 Conclusions

We have proposed a new statistical approach to investigate the complex structure of a genetic disorder such as hypertension, accounting for its multifactorial nature. We have expressed the genotype-phenotype biological chain as a chain graph. A graphical chain model has then been inferred in order to: i) define the complex phenotype, ii) analyse eventual epistatic interactions affecting the final phenotype iii) identify the filter-like role of some intermediate phenotypes.
Two possible chains, incorporating two different biological hypotheses, are postulated and compared. The major initial difference between them was whether *BMI* was to be treated as a final phenotype rather than as a prognostic factor. As seen in Figure 8, the hypothesis that *BMI* is a final response is more consistent with biological evidence.
In conclusion, even if graphical models cannot assess a causality structure in the data, they can help in excluding some postulated "directions" in the association links. Moreover, both the outstanding associations and the "causal" relations linking the studied variables can be read from the final graphs in Figure 8 and Figure 9. The filter role of the quantitative traits in connecting the genotypes and environmental variables to the final phenotypes was relevant. In addition, several important interactions among genes (epistatic interactions) and among these and quantitative traits have been suggested. All these results may be useful and interesting for further research in this multifactorial genetic disease. Moreover, the graphical chain model approach could be an important tool to study other multifactorial diseases.

Acknowledgements
We gratefully acknowledge Maurizio Castellano for providing the data used in the analyses and for his fundamental support in the discussion of the biological framework and the results. Thanks also to Daniele Cusi for his important clinical suggestions. Important issues for both statistical and computing methods have been discussed with Iris Pigeot and Angelika Blauth. The authors are also grateful to the unknown referees and Julia Mortera for fundamental suggestions and comments. This work was partially supported by Cofin-MIUR.

References
Bengtsson K., Melander O., Orho-Melander M., Lindblad U, Ranstam J., Ranstam L., Groop L., (2001) Polymorphism in the beta(1)-adrenergic receptor gene and hypertension. *Circulation,* **104(2)**, 187-90.
Bickeboeller, H., Clarget-Darpoux, (1995) F., Statistical properties of the allelic and genotypic trasmission disequilirbium test for multiallelic markers. *Genet. Epidemiol*., **12**, 865-870.
Blauth, A., Pigeot, I. and Bry, F. (2000) Interactive analysis of high-dimensional association structures with graphical models. *Metrika,* **51**, 53-65.
Castellano *et al.* (1995) Angiotensin-Converting Enzyme *I/D* Polymorphism and Arterial Wall Thickness in a General Population. The Vobarno Study. *Circulation*, **91**, 2721-2724.
Castellano, M., Di Serio, C., *et al*. (2002) Complex Association Structure in defining a genotype-phenotype causal mechanism in the Vobarno study. 2002, Mimeo.
Caputo, A., Heinicke, A., and Pigeot, I. (1999) A graphical chain model derived from a selection strategy for the sociologists graduates study. *Biometrical Journal*, **41**, 217-234.
Cavalli Sforza L.L., Bodmer W.F. (1973) *The genetics of human populations*. Freeman, New York.

Clarget-Darpoux, F. (1998) Overview on strategies in complex genetic diseases. *Kidney International*, 1441-144.
Cox, D.R., and Wermuth, N. (1994*)* Tests of Linearity, Multivariate normality and Adequacy of Linear Scores. *Appl. Statist.*, **43**, 347-355.
Cox, D.R., and Wermuth, N. (1996) *Multivariate dependencies - models, analysis and interpretation*. Chapman and Hall, London.
Cusi D., Bianchi G (1998) A primer on genetic of hypertension. *Kidney International*, **54**, 328-342.
Dawid, A. P. (1979) Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. B*, **41**, 1-31.
Edwards, D. (2000) *Introduction to graphical modelling*. New York, Springer, 2nd ed.
Edwards, D and Lauritzen, S. L. (2001) The TM algorithm for maximising a conditional likelihood function. *Biometrika*, **88**, 961-72.
Frydenberg, M. (1990) The chain graph Markov property. *Scandinavian J. Statist.*, **17**, 333-353.
Johnson RJ, et al. (2003) "Is there a pathogenic role for uric acid in hypertension and cardiovaclusar and renal disease", *Hypertension*, **41** (6), 1183-90.
Lauritzen, S.L. (1996) *Graphical models*. Oxford University Press, Oxford.
Lauritzen, S. L. and Richardson T. S. (2002) Chain graph models and their causal interpretations, *Journal of Roy. Stat. Soc. B*, **64**, 2, 1-28.
Ott, J. (1991) *Analysis of Human Genetic Linkage*. Johns Hopkins University Press.
Pearl, J. (2000) *Causality*. Cambridge University Press.
Pigeot, I., Heinicke, A., Caputo, A., and Brüderl, J. (2000) The professional career of sociologists: a graphical chain model reflecting early influences and associations. *Allgemeines Statistisches Archiv*, **84**, 3-21.
Rao, C. R. (1995) *Linear Models. Least squares and alternatives*. Springer Verlag, New York
Risch, N. (1990) Linkage strategies for genetically complex traits. *Am. Journal Hum. Genet.*, **46**, 242-253.
Terwillinger, J., Ott, J. (1992) A haplotype based "haplotype relative risks" approach to detecting allelic association, *Hum. Heredity*, **42**, 337-346.
Wermuth, N., and Lauritzen, S.L. (1990) On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). *Journal of the Royal Statistical Soc. B*, **52**, 21-72.

Reference address: Di Serio Clelia
University Hospital S. Raffaele, Università Vita- Salute S. Raffaele
Faculty of Psychology and Medicine
Via Olgettina 58
20132 Milano
email: diserio.clelia@hsr.it

Appendix 1.

The following tables show the genotype distributions.
Figure10(a) shows the box-plots of the conditional distributions of the variable *Glycaemia* given the levels of the *Snps6*. The distributions of the variables *Glyc* and *Trig* in the patient group with *Snps6*=0 are given in Figure 10(b) and Figure 10(c), respectively.

(Figure 10 here)

Table 5 Distribution of the genotypes included in the analysis

Snps1		
	Frequency	Percent
0	245	89.7
1	28	10.3
2	0	0
Total	273	100.0

Snps2		
	Frequency	Percent
0	109	39.9
1	136	49.8
2	28	10.3
Total	273	100.0

Snps3		
	Frequency	Percent
0	93	34.1
1	146	53.5
2	34	12.5
Total	273	100.0

Snps4		
	Frequency	Percent
0	28	10.3
1	101	37.0
2	144	52.7
Total	273	100,0

Snps5		
	Frequency	Percent
0	220	82.4
1	46	17.2
2	1	0.4
Total	267	100.0

Snps6		
	Frequency	Percent
0	175	64.1
1	88	32.2
2	10	3.7
Total	273	100.0

Raas1		
	Frequency	Percent
0	84	30.8
1	137	50.2
2	52	19.0
Total	273	100.0

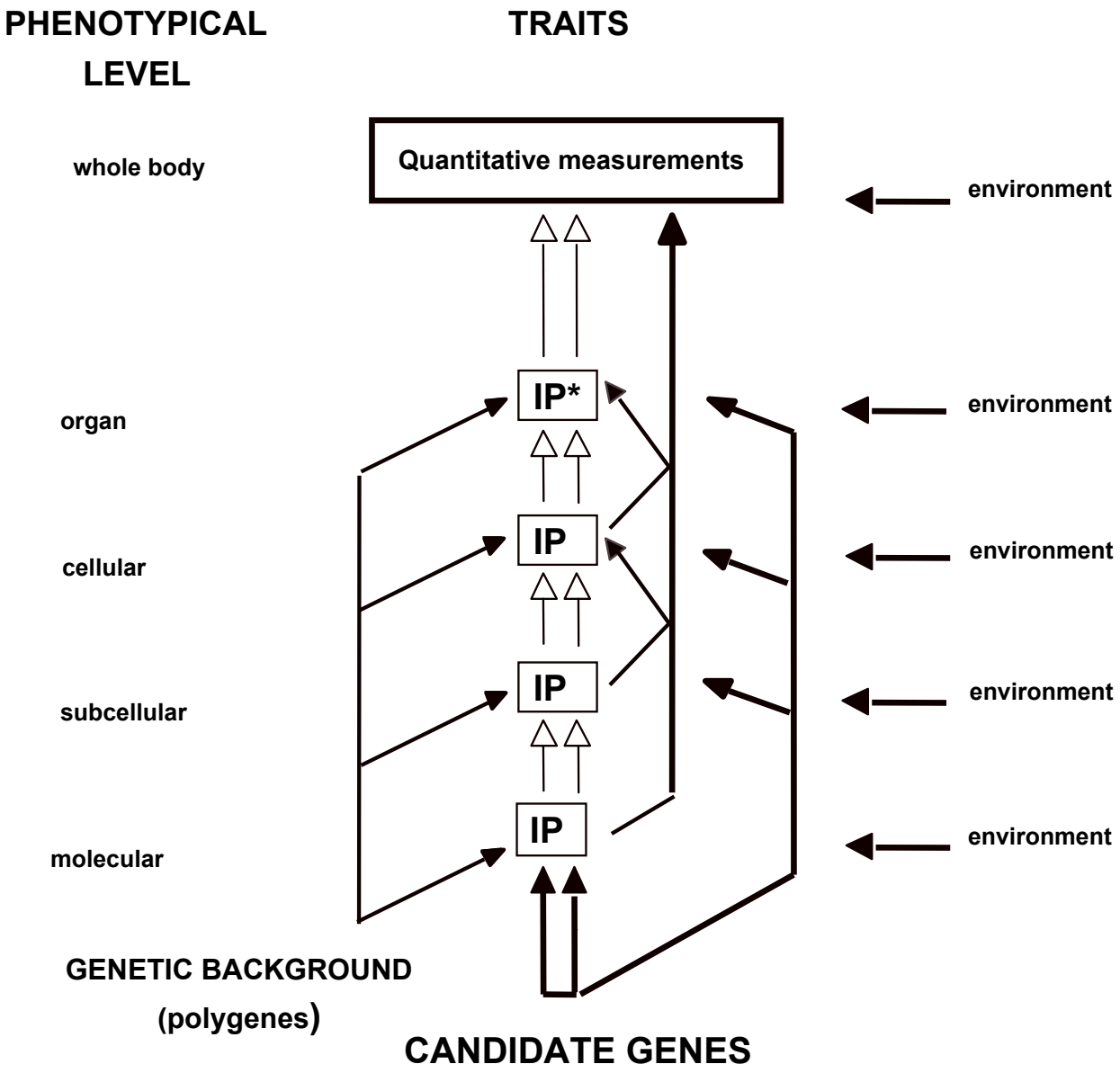
Raas2		
	Frequency	Percent
0	192	70.3
1	75	27.5
2	6	2.2
Total	273	100,0

Raas3		
	Frequency	Percent
0	136	49.8
1	48	17.6
2	89	32.6
Total	273	100.0

Raas4		
	Frequency	Percent
0	99	36.3
1	121	44.3
2	53	19.4
Total	273	100.0

Adducine		
	Frequency	Percent
0	184	74.8
1	51	20.7
2	11	4.5
Total	246	100.0

Figure 1. Genotype-Phenotype chain



* IP = Intermediate Phenotype

Figure 2. First postulated structure in complex diseases

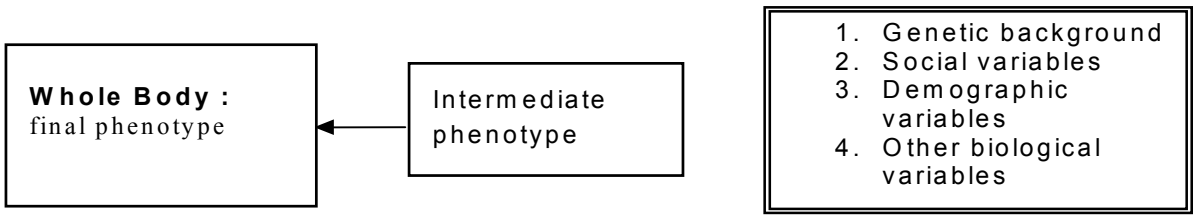


Figure 3. Marked chain graph

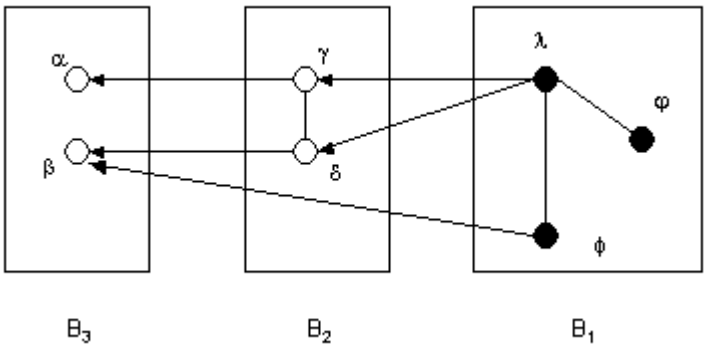


Figure 4. Chain graph for the genotype-phenotype chain

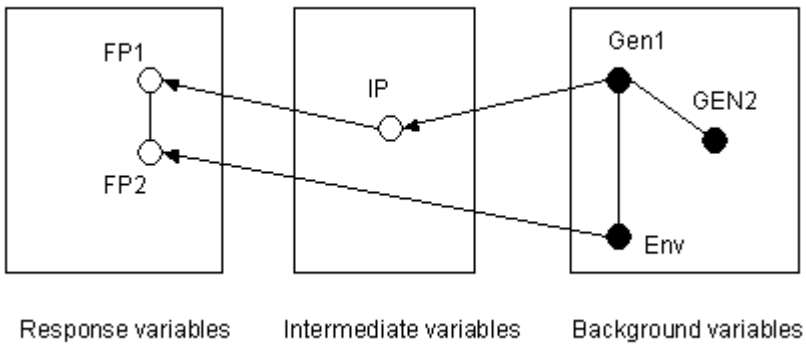


Figure 5. First (a) and second (b) postulated dependence chains

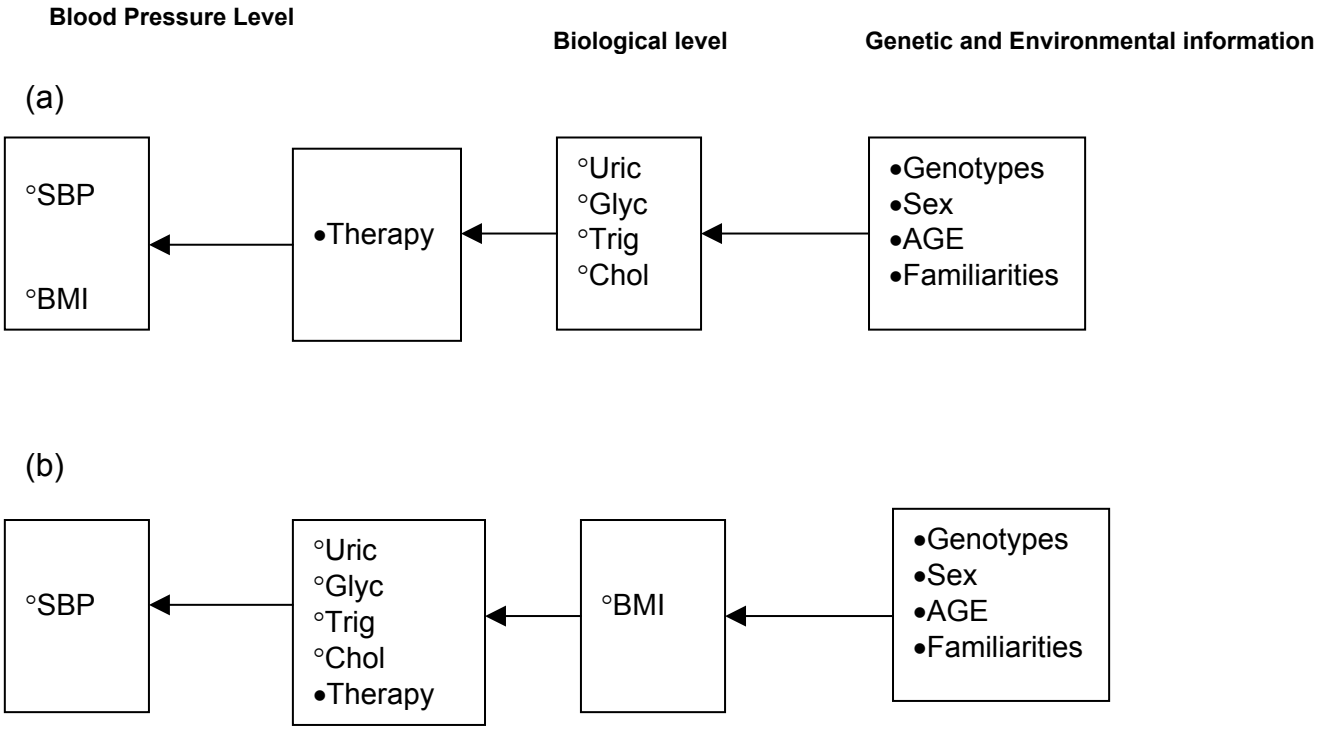
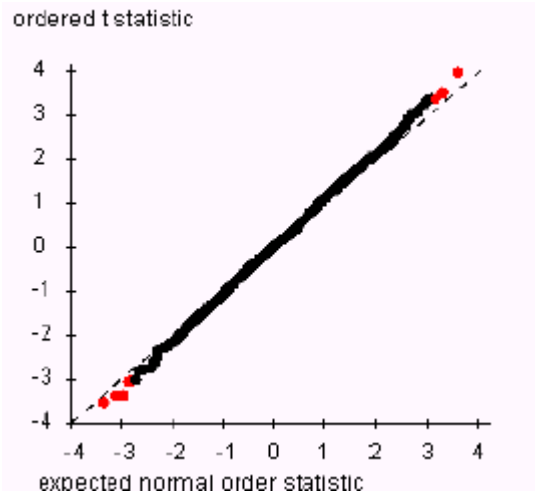
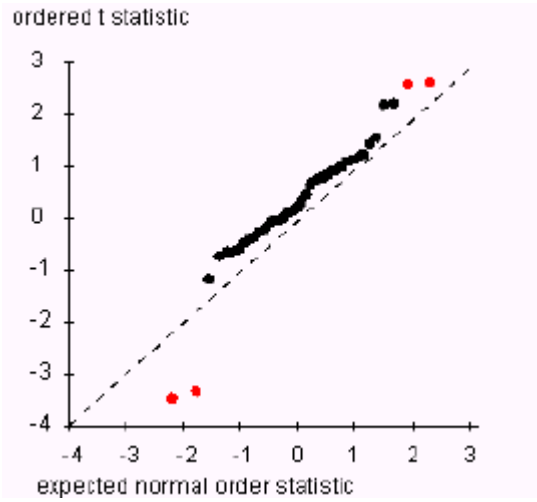


Figure 6a) screening for interaction. First postulated chain.



t = -3.55 y: BMI x: età_familiarDYS
t = -3.41 y: BMI x: età_RAAS2
t = -3.40 y: PAS24H x: età_familiarDYS
t = -3.08 y: COL x: familiarDIA_RAAS3
t = -3.05 y: PAS24H x: familiarHYP_familiarDIA
t = 3.34 y: TRIG x: GLICEMIA_SNPS6
t = 3.79 y: GLICEMIA x: TRIG_SNPS6
t = 3.21 y: COL x: TRIG_RAAS3

Figure 6b) screening for non-linearities.

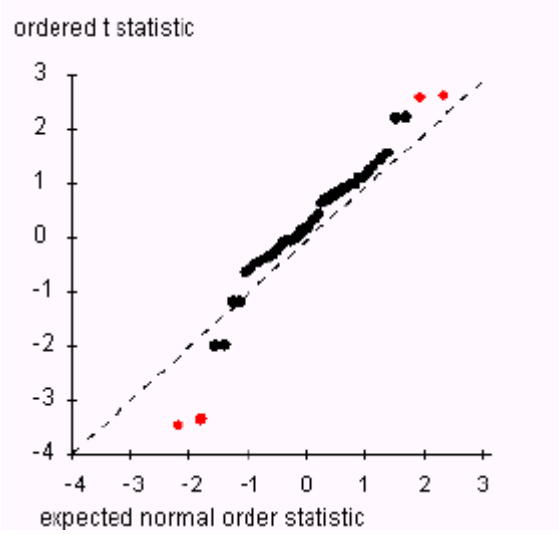


t = -3.48 y: COL x: TRIG^2
t = -3.35 y: URIC x: TRIG^2
t = 2.57 y: TRIG x: GLICEMIA^2
t = 2.61 y: PAS24H x: età^2

Figure 7a) screening for interaction. Second postulated chain.

t = -3.41 y: BMI x: età_RAAS2
t = -3.40 y: PAS24H x: età_familiarDYS
t = -3.08 y: COL x: familiarDIA_RAAS3
t = -3.05 y: PAS24H x: familiarHYP_familiarDIA
t = -3.55 y: BMI x: età_familiarDYS
t = 3.34 y: TRIG x: GLICEMIA_SNPS6
t = 3.79 y: GLICEMIA x: TRIG_SNPS6
t = 3.21 y: COL x: TRIG_RAAS3

Figure 7b) screening for nonlinearities.



t = -3.35 y: URIC x: TRIG^2
t = -3.48 y: COL x: TRIG^2
t = 2.57 y: TRIG x: GLICEMIA^2
t = 2.61 y: PAS24H x: età^2

Figure 8. Final chain related to the first postulated chain

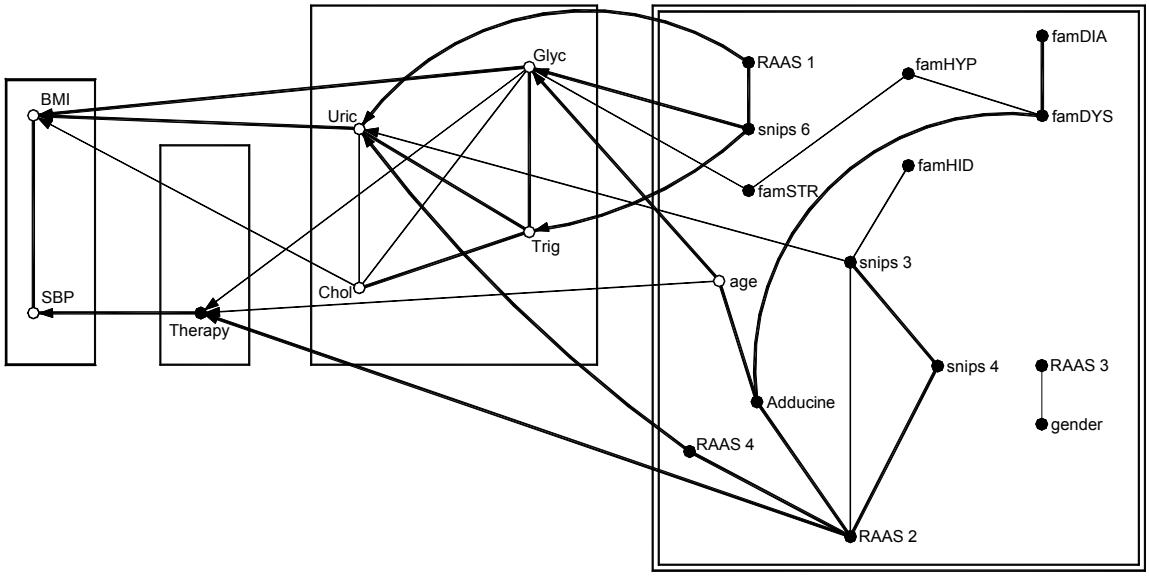


Figure 9. Final chain related to the second postulated chain

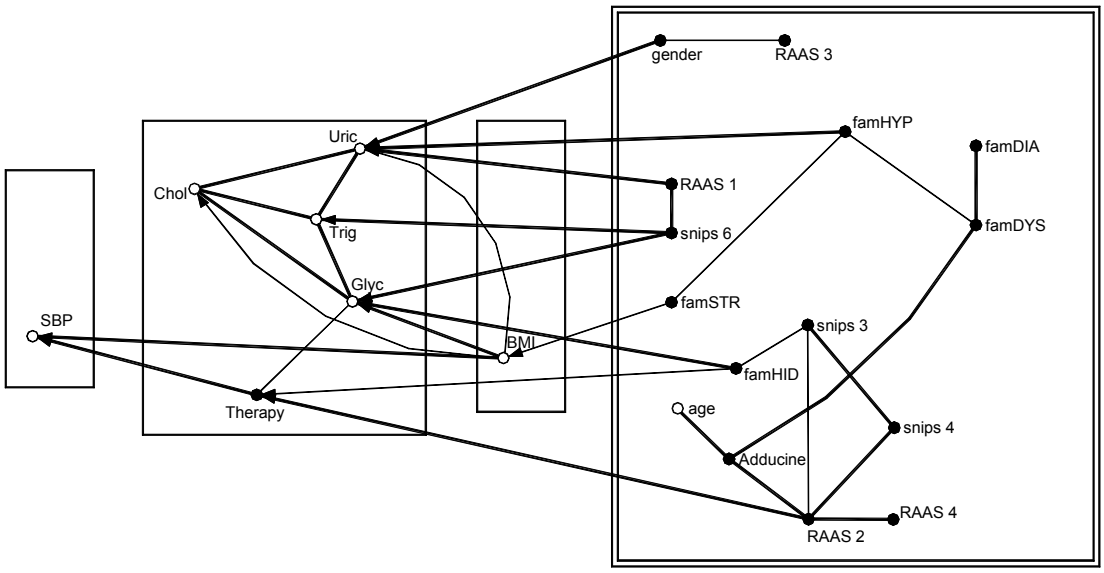


Figure 10. Box-plot of the variable *Glycaemia* conditionally on the levels of *Snps6* (A); distribution of *Glycaemia* conditionally on *Snps6*=0 (B); distribution of *Triglicerides* conditionally on *Snps6*=0 (C).

(a) (b) (c)