Genome Medicine

MINIREVIEW

Simulating gene-environment interactions in complex human diseases

Bo Peng*

Abstract

Because little is currently known about how genes interact with environmental factors in human diseases, and because of the large number of possible interactions between and within genetic and environmental factors, it is difficult to simulate samples for a disease caused by multiple interacting genetic and environmental factors. A recent article by Amato and colleagues in *BMC Bioinformatics* describes a mathematical model to characterize gene-environment interactions and a computer program that simulates them using biologically meaningful inputs. Here, I evaluate the advantages and limitations of the authors' approach in terms of its usefulness for simulating genetic samples for real-world studies of gene-environment interactions in complex human diseases.

Introduction

Simulated datasets with known underlying disease mechanisms have been widely used to develop efficient statistical methods for deciphering the complex interplay between the genetic and environmental factors responsible for complex human diseases, such as hypertension, diabetes and cancer [1-3]. Although genetic and environmental risk factors have been identified for various human diseases, little is currently known about how genes interact with environmental factors in these diseases. Because the number of possible interactions between and within genetic and environmental factors is large, it is difficult to specify and simulate samples for a disease caused by multiple interacting genetic and environmental factors. Consequently, existing studies have focused on simple models with low-order interactions between a few genetic and environmental factors, using specialized simulation programs. Here, I discuss a recent article by Amato and colleagues in *BMC Bioinformatics* [4], which describes a mathematical model to characterize gene-environment interactions (GxE) and a computer program that simulates them using biologically meaningful inputs. I evaluate the usefulness of the authors' method for simulating samples with GxE for future studies.

*Correspondence: bpeng@mdanderson.org
Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

Specifying a GxE model for disease risk

A disease model is needed before a sample can be simulated. If the number of genetic factors that cause a disease is G, we can denote each genetic factor by g_i (where $i = 1,\ldots,G$), and each of these will have three diploid genotypes. Similarly, with E environmental factors, we can denote these x_j and each would have b_j possible discrete values (where $j = 1,\ldots,E$). A complete GxE model would then have $3^G \times \Pi_{j=1}^{E} b_j$ possible items for each combination of genetic and environmental factors. In addition, the model would require this same number of parameters to specify the risk associated with each item. Although such models can be used to specify arbitrary gene-gene and gene-environment interactions, estimating a large number of parameters from empirical data is challenging and usually not feasible.

Amato *et al.* [4] propose a statistical model, called the Multi-Logistic Model (MLM), that is designed to describe disease risk in datasets that simulate case-control samples. MLM, which is a natural extension of logistic models used by others [2,3], allows the specification of disease risks caused by all genetic factors and by interactions between genotype and all environmental factors. It reduces the required number of parameters to $3^G \times (1 + E)$ by making the following assumptions: that the log odds ratios of environmental factors are additive; and that the different environmental factors are independent and additive. The latter assumption means that only $1 + E$ parameters are required for each combination of genotypes because the impact of b_j levels of exposure for each environmental factor is represented by one parameter and no interaction between environmental factors is allowed. These assumptions limit the application of MLM in studies with correlated environmental factors (for example, smoking and drinking [5]). The

simplified model therefore cannot be used to model complex GxE structures, such as the development of lung cancer caused by smoking and genetic factors because the impact of smoking is highly correlated with age, which is a common covariate in such models. Despite this limitation, the MLM approach could be made more generally applicable by making adjustments, such as by applying principal component analysis, to the environmental factors to ensure their independence.

Even with the reduction of parameters afforded by the assumption of independence, the number of parameters in an MLM is still large if multiple genetic and environmental factors are involved. For example, an MLM requires 18 parameters when there are two genetic factors and one environmental factor. This is why the authors [4] focused on a version of MLM with only one genetic factor and one environmental factor (giving only six parameters), which they implemented in Matlab in their program Gene-Environment iNteraction Simulator (GENS). Furthermore, users of GENS can choose from four simpler models of GxE (Figure 1): a genetic model (no environmental factors, three parameters), an environmental model (no genetic factors, two parameters), a gene-environment interaction model (genotypes do not directly affect disease risk, four parameters), and an additive model (environmental factors have the same effect in all genotypes, four parameters). These models, although incomplete, should be sufficient for most theoretical studies of GxE models with one genetic factor and one environmental factor.

Because changing an interaction item might change many properties (such as the marginal effects of a model) in an unpredictable way, it is difficult for users to adjust parameters in a GxE model to control for key epidemiological features of a disease such as population incidence. Amato *et al.* [4] used an innovative system, the Knowledge-Aided Parameterization System (KAPS), to translate user input in familiar epidemiological terminologies, such as model of inheritance, into the parameters used in MLM, which makes it easy for users to specify model parameters that are epidemiologically sensible. Other constraints, such as relative risk between homozygotes and heterozygotes, are added to facilitate the search for suitable parameters. KAPS works well for models with one genetic and one environmental factor because the number of epidemiological variables that users need to input is similar to the number of model parameters. For a general GxE model with multiple genetic and environmental factors, the number of epidemiological features of the disease and individual genetic and environmental factors will be far less than the number of model parameters because of the large number of interaction terms in MLM. Because multiple models with different interaction terms could have the same epidemiological features, additional constraints are required to limit the number of plausible models, and a complex search algorithm might be needed to parameterize MLM with sensible interaction parameters. An example of fitting a more complex model was presented by Moore *et al.* [6], who used a genetic algorithm to discover, among a large number of plausible theoretical models, a special set of high-order gene-gene interaction models in which genes influence disease risk only through interactions with other genes, without any main effects.

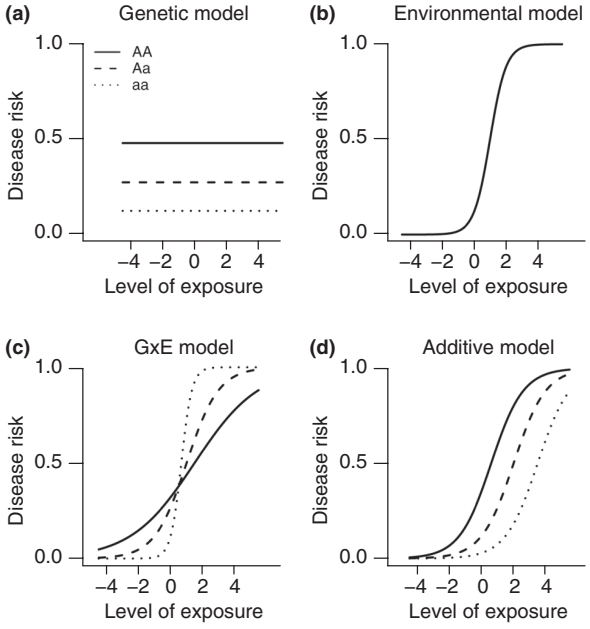

Figure 1. Four GxE interaction models provided by the Knowledge-Aided Parameterization System (KAPS): (a) a genetic model, (b) an environmental model, (c) a gene-environment interaction model and (d) an additive model. Each curve represents the relationship between disease risk (*y*-axis) and an environmental factor (*x*-axis) for individuals with a particular genotype (AA, Aa or aa) at a disease-predisposing locus (DPL). The environmental model has only one curve because the relationship between environmental exposure and disease risk is identical for all genotypes. Adapted from Amato *et al.* [4].

Applicability of the simulation tool

Various different methods have been used to simulate case-control samples based on penetrance models. Before applying their GxE model, Amato *et al.* [4] simulated a population to determine the affection status of each individual (that is, whether or not that individual is affected by the disease). There are two possible ways of doing the simulation. The first method is to simulate a large population and then select case-control or other types of samples, such as pedigrees, from it. This approach allows maximum flexibility in the specification

of a penetrance model and is usually used in a forward-time approach in which a population is simulated by evolving from a founder population forward in time under the influence of multiple genetic and demographic forces [7]. This method can be inefficient if the disease is so rare that a large population needs to be simulated to obtain enough cases. If, alternatively, a disease model is simple enough, $\Pr(g_i, x_j \mid \text{affection status})$ (that is, the probability that the genotype is g_i and the environmental factor value is x_j given the affection status) can be determined from $\Pr(\text{affection status} \mid g_i, x_j)$ together with other parameters, such as frequencies of these factors and disease prevalence. If this is the case, genotype and environmental factors can be simulated directly and only the required number of cases and controls needs to be simulated. This second approach has been used by many simulation programs, such as HapSample [8], hapgen [9] and GWAsimulator [10]. As a compromise between these two approaches, a rejection-sampling algorithm can be used to simulate samples without simulating a large population (for example, genomeSIMLA [11]). This method repeatedly simulates individuals, assigns affection status, and collects cases and controls until enough samples have been simulated. This approach is suitable for situations in which environmental factors can be independently simulated for each individual, and could be used to improve the efficiency of GENS.

Because genotypes at the disease-predisposing loci (DPL) of a genetic disease might not be available, many statistical methods rely on linkage disequilibrium (LD) between DPL and their surrounding markers to indirectly map the DPL. GENS does not consider LD between DPL and surrounding genetic markers, so more sophisticated simulation methods are needed to simulate linked markers using genetically related individuals. Existing approaches include: resampling from existing data (for example, HapSample [8] or hapgen [9]); reconstructing from statistical properties obtained from existing sequences (GWAsimulator [10]); simulating a complete genealogy (coalescent tree) of a sample or population (for example, cosi [12], GENOME [13]); and evolving forward in time from a population [7,11]. The power, flexibility, performance and quality of simulated samples vary greatly from program to program. For example, forward-time methods are most flexible because they can follow the evolution of a real population closely, but they are inefficient because they simulate all ancestors, including those who do not have offspring in the simulated population. GWAsimulator [10] retains the short-range LD structure of the human population (or more specifically, the HapMap sample) but discards long-range LD because the method simulates haplotypes according to short-range LD patterns obtained from the HapMap sample [14] using a sliding-window approach.

Although it is generally possible to simulate large populations using these methods and then apply the GxE disease model proposed by Amato *et al.* [4], several obstacles remain. For example, many coalescent-based simulation methods [13,15] simulate markers with varying location and allele frequency, so it is difficult to apply a fixed-disease model to replicate simulations. If a forward-time approach is used to simulate samples with the same set of markers, sample frequencies of the DPL will vary because of the impact of random genetic drift, unless special algorithms are used to control allele frequencies [7]. Even if samples with the same allele frequencies are simulated, the individuals generated may not have enough genetic variations to allow adequate modeling with a GxE model because of insufficient combinations of genetic and environmental factors. For example, from a sample of 20,000 sequences of 40 tightly linked markers over a 100 kb region on chromosome 17 simulated using hapgen [9], there were only 74 unique haplotypes because all the haplotypes were derived from the 63 unique haplotypes in the HapMap CEU sample [14] using an imputation approach.

Conclusions

Amato *et al.* [4] have provided a mathematical model for the specification of interactions between genetic and environmental risk factors. Their simulation program GENS can be used to generate simple, independent case-control samples with clear epidemiological interpretations and can be used to validate a statistical method or compare the performance of several statistical methods under specific assumptions. However, because real-world studies usually involve a large number of linked markers, a useful statistical method should be able to identify informative variables (DPL and environmental factors) from a large number of markers and covariates [16], or be efficient enough to be used to search for GxE signals exhaustively [17]. The performance of statistical methods that detect GxE in complex human diseases, including sensitivity, specificity and ability to handle linked loci, should be tested against simulated samples of long genome sequences with realistic disease models and LD patterns. Although progress has been made in both the simulation of long genome sequences [10,15] and GxE disease models [4], the combination of these two approaches would produce realistic samples that could greatly aid the study of GxE in complex human diseases.

Abbreviations

DPL, disease-predisposing locus; GENS, Gene-Environment iNteraction Simulator; GxE, gene-environment interaction; KAPS, Knowledge-Aided Parameterization System; LD, linkage disequilibrium; MLM, Multi-Logistic Model.

Competing interests

The author declares that he has no competing interests.

Acknowledgements
This work was supported by grant R01 CA133996 from the National Cancer Institute and by MD Anderson's Cancer Center Support Grant CA016672 from the National Institutes of Health. I thank Christopher Amos for his helpful comments about the manuscript.

Published: 23 March 2010

References

1. Motsinger-Reif AA, Reif DM, Fanelli TJ, Ritchie MD: **A comparison of analytical methods for genetic association studies.** *Genet Epidemiol* 2008, **32:**767-778.
2. Li D, Conti DV: **Detecting gene-environment interactions using a combined case-only and case-control approach.** *Am J Epidemiol* 2009, **169:**497-504.
3. Wang T, Ho G, Ye K, Strickler H, Elston RC: **A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped.** *Genet Epidemiol* 2009, **33:**6-15.
4. Amato R, Pinelli M, D'Andrea D, Miele G, Nicodemi M, Raiconi G, Cocozza S: **A novel approach to simulate gene-environment interactions in complex diseases.** *BMC Bioinformatics* 2010, **11:**8.
5. Isohanni M, Oja H, Moilanen I, Rantakallio P, Koiranen M: **The relation between teenage smoking and drinking, with special reference to non-standard family background.** *Scand J Soc Med* 1993, **21:**24-30.
6. Moore JH, Hahn LW, Ritchie MD, Thornton TA, White BC: **Routine discovery of complex genetic models using genetic algorithms.** *Appl Soft Comput* 2004, **4:**79-86.
7. Peng B, Amos CI, Kimmel M: **Forward-time simulations of human populations with complex diseases.** *PLoS Genet* 2007, **3:**e47.
8. Wright FA, Huang H, Guan X, Gamiel K, Jeffries C, Barry WT, Pardo-Manuel F, Sullivan PF, Wilhelmsen KC, Zou F: **Simulating association studies: a data-based resampling method for candidate regions or whole genome scans.** *Bioinformatics* 2007, **23:**2581-2518.
9. Marchini J, Howie B, Myers S, McVean G, Donnelly P: **A new multipoint method for genome-wide association studies by imputation of genotypes.** *Nat Genet* 2007, **39:**906-913.
10. Li C, Li M: **GWAsimulator: a rapid whole-genome simulation program.** *Bioinformatics* 2008, **24:**140-142.
11. Edwards TL, Bush WS, Turner SD, Dudek SM, Torstenson ES, Schmidt M, Martin E, Ritchie MD: **Generating linkage disequilibrium patterns in data simulations using genomeSIMLA.** *Lect Notes Comput Sci* 2008, **4973:**24-35.
12. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D: **Calibrating a coalescent simulation of human genome sequence variation.** *Genome Res* 2005, **15:**1576-1583.
13. Liang L, Zöllner S, Abecasis GR: **GENOME: a rapid coalescent-based whole genome simulator.** *Bioinformatics* 2007, **23:**1565-1567.
14. The International HapMap Consortium: **A haplotype map of the human genome.** *Nature* 2005, **437:**1299-1320.
15. Chen GK, Marjoram P, Wall JD: **Fast and flexible simulation of DNA sequence data.** *Genome Res* 2009, **19:**136-142.
16. Chanda P, Sucheston L, Zhang A, Brazeau D, Freudenheim JL, Ambrosone C, Ramanathan M: **AMBIENCE: a novel approach and efficient algorithm for identifying informative genetic and environmental associations with complex phenotypes.** *Genetics* 2008, **180:**1191-1210.
17. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81:**559-575.

doi:10.1186/gm142
Cite this article as: Peng B: Simulating gene-environment interactions in complex human diseases. *Genome Medicine* 2010, **2:**21.