Kalpa Publications in Computing

Volume 14, 2023, Pages 46–49

Proceedings of V XoveTIC Conference. XoveTIC 2022

Generating penetrance tables of high-order epistasis models with PyToxo

Borja González-Seoane[1], Christian Ponte-Fernández[2], Jorge González-Domínguez[2], and María J. Martín[2]

[1] Universidade da Coruña, Facultade de Informática, 15071, A Coruña, Spain
borja.gseoane@udc.es

[2] Universidade da Coruña, CITIC, Computer Architecture Group, 15071, A Coruña, Spain
{christian.ponte, jorge.gonzalezd, maria.martin.santamaria}@udc.es

Abstract

The interaction among different genes when expressing a particular phenotype is known as epistasis. High-order epistasis, when more than two loci are involved, is an active research area because it could be the cause of many complex traits. The most common abstraction for specifying an epistasis interaction is through a penetrance table, which captures the probability of expressing the studied phenotype given a particular genotype.

Although it is very common for simulators to use penetrance tables, most of them do not allow the user to generate them directly, or present limitations for high-order interactions and/or realistic prevalence and heritability values. In this work, we present PyToxo, a Python tool for generating penetrance tables from any-order epistasis models. PyToxo allows to work with more appropriate scenarios than other state-of-the-art tools. Additionally, it also improves in terms of accuracy, speed and ease of use, being available as a library, through a CLI or through a cross-platform GUI.

1 Introduction

Genome-Wide Association Studies (GWAS) analyze genetic markers to find associations between diseases and genetic variations. Conventional GWAS are based on the analysis of differences between the genotype frequencies from individual genetic markers of case and control samples. However, recent studies have shown that epistasis interactions need to be considered to establish a relation between genotypes and phenotypes in many traits [6].

Epistasis is the interaction of a genetic variation at two or more loci (specific positions of a gene) to produce a phenotype that is not explained by the additive combination of effects of the individual loci. If epistasis involves more than two loci, it is called high-order epistasis. High-order epistasis is related with diseases such as breast cancer [3] or Alzheimer's [5].

Epistasis relationships are typically represented through a penetrance table, which captures the probability of expressing the phenotype to be studied given a particular allele combination. In the literature there are software simulators, such as EpiSIM [4], that allow obtaining a penetrance table for a previously established prevalence ($P(D)$) and heritability (h^2). However,

as the other tools of the state of the art, it can only work, in practice, with second-order models and low prevalence and heritability values [2], preventing it from simulating realistic scenarios.

To overtake these limitations, in a previous work we introduced Toxo [2], a MATLAB library for calculating penetrance tables of epistasis models with no limitation on the interaction order, thanks to a simplification of the mathematical method. Toxo can calculate penetrance tables with prevalence and heritability values much higher than those observed in the state of the art, but it also has its own disadvantages.

To solve those drawbacks we developed PyToxo, a Python version of Toxo that was firstly presented in a BMC Bioinformatics [1] paper in 2022. PyToxo is distributed as open-source software and, unlike its predecessor, it does not need any comercial license. In addition, PyToxo improves Toxo in terms of the complexity of the epistasis models to handle, the accuracy in the obtained tables, the execution times and the ease of use.

2 Methods

2.1 Mathematical method

There are in the literature simulators that allow obtaining a penetrance table for a given prevalence ($P(D)$) and heritability (h^2) through the solution of the following system of equations:

$$P(D) = \sum_i P(D|g_i)P(g_i) \qquad h^2 = \frac{\sum_i \big(P(D|g_i) - P(D)\big)^2 P(g_i)}{P(D)\big(1 - P(D)\big)} \tag{1}$$

where $P(D|g_i) = f_i(x, y)$ is the proportion of individuals showing trait D when having the genotype g_i, $P(g_i)$ is the population frequency of the genotype g_i and $f_i(x, y)$ is the function of two variables that defines the epistasis model.

As we explain in more detail in our BMC Bioinformatics paper [1], the mathematical approach followed for Toxo, and on which PyToxo is settled, is based on fixing heritability or prevalence and maximizing the other. So, instead of finding a specific combination of heritability and prevalence, like in original Equation 1, the Toxo library maximizes one of the two parameters, when the other is fixed.

Due to above, the likelihood of formulating an incompatible system is significantly reduced and the resolution is simplified. This allows Toxo (and PyToxo) to calculate penetrance tables with prevalence and heritability values much higher than those observed in the state of the art.

2.2 Software implementation

PyToxo is implemented in Python, which is currently one of the most widely used programming languages and one of the most frequently used options in the bioinformatics interdisciplinary field. PyToxo takes as input an epistasis model, a heritability (or prevalence) value and the Minor Allele Frequencies (MAFs) associated with each of the considered loci, and generates as output a penetrance table maximizing the prevalence (or heritability).

In addition to being offered as a Python library for developers, PyToxo provides two user interfaces: a CLI for advanced users which are able to run the program in batch processing, and a GUI, especially oriented for users unfamiliar with command-line execution environments. All the interfaces are cross-platform and can be used in Linux, MacOS and Windows machines.

3 Results

During the development of PyToxo, a large number of tests were run to check its performance. After completing all the tests, we concluded categorically that PyToxo improves on Toxo in all the studied aspects. PyToxo has a better model coverage, being able to solve all the configurations that Toxo solves and more complex cases; better computation accuracy, with an average error of 5.74×10^{-17} versus the 1.78×10^{-4} of Toxo; an average speedup of 1.90 over Toxo, and much better usability, with three different interfaces and no requirement for a commercial license. PyToxo would become the tool of choice between the two for obtaining penetrance tables, regardless of the epistatic model and input parameters considered.

4 Discussion and conclusions

The best way to prove new algorithms to detect high-order epistasis is through simulated data, since they provide a controlled environment where the implied epistatic interactions are known in advance. It is very frequent for simulators to use penetrance tables to capture epistasis interactions, but most of them present limitations for high-order interactions and/or realistic prevalence and heritability values.

PyToxo is a Python utility to calculate penetrance tables for high-order epistasis models, improving its peers in four different aspects: first, in terms of coverage, PyToxo is able to generate penetrance tables of complex epistasis models with realistic configurations; then, in terms of accuracy, because its average error is in the order of 10^{-17}; third, in terms of speed, computing a penetrance table in less than a minute even for order 8 epistasis models; and last, in terms of ease of use, due to PyToxo can be used as a library, through a CLI or through a GUI, which may be especially useful for those users not experts in programming or command-line execution environments.

PyToxo is available to the whole scientific community as an open-source cross-platform software. The source code of PyToxo, detailed user guides, and all the models and code examples used are available in the GitHub repository: `https://github.com/bglezseoane/pytoxo`. In addition, PyToxo is uploaded to the official Python PyPI repository (`https://pypi.org/project/pytoxo/`) and it can be very easily installed by only running the `pip install pytoxo` command.

To know more about PyToxo, we recommend reading our original paper, published in BMC Bioinformatics [1] this year.

References

[1] Borja González-Seoane, Christian Ponte-Fernández, Jorge González-Domínguez, and María J. Martín. PyToxo: a Python tool for calculating penetrance tables of high-order epistasis models. *BMC Bioinformatics*, 23(1):117, April 2022.

[2] Christian Ponte-Fernández, Jorge González-Domínguez, Antonio Carvajal-Rodríguez, and María J Martín. Toxo: a library for calculating penetrance tables of high-order epistasis models. *BMC bioinformatics*, 21(1):1–9, 2020.

[3] Marylyn D Ritchie, Lance W Hahn, Nady Roodi, L Renee Bailey, William D Dupont, Fritz F Parl, and Jason H Moore. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics*, 69(1):138–147, 2001.

[4] J Shang, J Zhang, X Lei, W Zhao, and Y Dong. Episim: Simulation of multiple epistasis, linkage disequilibrium patterns and haplotype blocks for genome-wide interaction analysis. *Genes & Genomics*, 35(3):305–16, 2013.

[5] Jiya Sun, Fuhai Song, Jiajia Wang, Guangchun Han, Zhouxian Bai, Bin Xie, Xuemei Feng, Jianping Jia, Yong Duan, and Hongxing Lei. Hidden risk genes with high-order intragenic epistasis in alzheimer's disease. *Journal of Alzheimer's Disease*, 41(4):1039–1056, 2014.

[6] Matthew B. Taylor and Ian M. Ehrenreich. Higher-order genetic interactions and their contribution to complex traits. *Trends in Genetics*, 31(1):34–40, 2015.