Gene-gene interaction: the curse of dimensionality

Amrita Chattopadhyay, Tzu-Pin Lu

Institute of Epidemiology and Preventive Medicine, Department of Public Health, National Taiwan University, Taipei

Contributions: (I) Conception and design: A Chattopadhyay; (II) Administrative support: TP Lu; (III) Provision of study materials or patients: A Chattopadhyay; (IV) Collection and assembly of data: A Chattopadhyay; (V) Data analysis and interpretation: A Chattopadhyay; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Tzu-Pin Lu, PhD. Room 518, No. 17, Xu-Zhou Road, 100, Taipei. Email: tplu@ntu.edu.tw.

Abstract: Identified genetic variants from genome wide association studies frequently show only modest effects on the disease risk, leading to the "missing heritability" problem. An avenue, to account for a part of this "missingness" is to evaluate gene-gene interactions (epistasis) thereby elucidating their effect on complex diseases. This can potentially help with identifying gene functions, pathways, and drug targets. However, the exhaustive evaluation of all possible genetic interactions among millions of single nucleotide polymorphisms (SNPs) raises several issues, otherwise known as the "curse of dimensionality". The dimensionality involved in the epistatic analysis of such exponentially growing SNPs diminishes the usefulness of traditional, parametric statistical methods. With the immense popularity of multifactor dimensionality reduction (MDR), a non-parametric method, proposed in 2001, that classifies multi-dimensional genotypes into one-dimensional binary approaches, led to the emergence of a fast-growing collection of methods that were based on the MDR approach. Moreover, machine-learning (ML) methods such as random forests and neural networks (NNs), deep-learning (DL) approaches, and hybrid approaches have also been applied profusely, in the recent years, to tackle this dimensionality issue associated with whole genome gene-gene interaction studies. However, exhaustive searching in MDR based approaches or variable selection in ML methods, still pose the risk of missing out on relevant SNPs. Furthermore, interpretability issues are a major hindrance for DL methods. To minimize this loss of information, Python based tools such as PySpark can potentially take advantage of distributed computing resources in the cloud, to bring back smaller subsets of data for further local analysis. Parallel computing can be a powerful resource that stands to fight this "curse". PySpark supports all standard Python libraries and C extensions thus making it convenient to write codes to deliver dramatic improvements in processing speed for extraordinarily large sets of data.

Keywords: Gene-gene interaction; parallel computing; PySpark; deep-learning (DL); machine-learning (ML); multifactor dimensionality reduction (MDR)

Submitted Nov 02, 2019. Accepted for publication Dec 05, 2019.
doi: 10.21037/atm.2019.12.87
View this article at: http://dx.doi.org/10.21037/atm.2019.12.87

The single nucleotide polymorphism (SNP) is the key genetic unit that is utilized in genome-wide association studies (GWASs) to unravel the genetic basis of complex human diseases. Thousands of GWASs have been reported in the GWAS Catalog (1), each of which recounts several genetic variants associated with a disease of interest. However, the identified genetic variants frequently show only modest effects on the disease risk, which is referred to as the "missing heritability" problem (2). Among other factors, this "missingness" is attributed to genetic heterogeneity, epistasis (gene-gene interaction), and gene-environment interaction and is an impediment to accurate prediction of disease risk from genetic information (3,4). Epistasis occurs when 2 or more genes interact to affect the phenotype of an organism (5). Given the complexity of biomolecular interactions (biological epistasis) (6) in gene regulation and metabolic systems, the relationship between DNA variants and clinical endpoints (statistical

© Annals of Translational Medicine. All rights reserved.

epistasis) (6) is likely to involve gene-gene interactions. Moreover it has been predicted that epistasis could drive the evolution of recombination frequencies among genes on the same chromosome, thereby altering gene order. Therefore a negative correlation between epistasis and gene distance potentially evolves leading to them having similar expression profiles (7).

Accounting for the effect of genetic interactions can help with identifying gene functions, pathways, and drug targets. These interactions include synthetic, suppressive, and epistatic types. Synthetic interaction occurs when two genes from parallel pathways produce a phenotype, so the phenotype can still be observed even if one of the genes gets inactivated (knocked out); however, if both genes are knocked out, the phenotype will be altered. Suppressive interaction occurs when the phenotypic defects caused by a mutation in a particular gene are rescued by a mutation in a second gene. Finally, epistatic interaction occurs between two genes if one allele of the first gene masks the expression of an allele of the second gene. In a nutshell, epistatic genes mask each other's presence or combine to produce an entirely new trait (8). In the last 2 decades, numerous studies have been attempted to better understand genetic interactions and their contribution to missing heritability (9).

A lot of challenges are associated with conducting gene-gene interaction analysis. It becomes challenging to conduct whole genome gene-gene interaction analysis, as the number of potential interactions exponentially increases with the increasing number of SNPs. Multi-level analysis has been implemented in prior studies to address this issue of dimensionality. For example, in numerous studies, many SNPs are excluded through quality control measures. Also, most GWASs are limited to common variants (10), focusing on only SNPs with a minor allele frequency (MAF) (>5%). It has also been shown that low-frequency (1%≤ MAF <5%) and/or rare (MAF <1%) variants account for part of the missing heritability (11). In rare Mendelian disorders, causal rare variants tend to show high penetrance, whereas in complex disorders, the penetrance levels of rare variants are mostly moderate to low. Also, within the extensive worldwide efforts aimed at creating reference genomes with millions of SNPs (12-14), epistasis studies are overwhelmingly difficult because of the problem of dimensionality. High dimensionality, together with multiple polymorphisms, exponentially increases the computational complexity of traditional statistical approaches, not to mention the time.

One of the most famous strategies for reducing the dimension of high order variables, which has been used profusely, is a non-parametric approach, first proposed by Ritchie *et al.*, known as the multifactor dimensionality reduction (MDR) method (15). The fundamental idea of MDR is to classify multi-dimensional genotypes into one-dimensional binary approaches by pooling genotypes of multiple SNPs using the ratio of cases and controls. Many variations of MDR have been proposed over the years. A roadmap of MDR showing the temporal development of MDR and MDR-based approaches has been developed by Gola *et al.* (16). For instance, generalized MDR, proposed by Lou *et al.*, uses a score-based residual on both binary and continuous phenotypes to classify multi-level genotypes into high and low risk (17). Other variations include model-based MDR (18), odds ratio-based MDR (19), and robust MDR (20). Two variations of MDR survival prediction models, surv-MDR (21) and Cox-MDR (22), have also been proposed. The former uses a log-rank test statistic with survival time instead of case-control ratios to identify survival associated with multi-way SNP interactions; however, it cannot adjust for covariate effects, whereas the latter uses a martingale residual classifier to adjust for the covariate effects. Another modification of such a survival-based MDR method is AFT-MDR (23), which utilizes the normalized difference between observed and expected log survival times as the standardized residual for classifying gene-gene interactions into high and low risk. AFT-MDR has also been combined with unified model MDR, which uses a non-central chi-square test to find the significance of gene-gene interactions for all possible pairs of SNPs without any intensive permutations (23). In spite of the popularity of MDR, its basic structure eliminates useful SNPs due to exhaustive searching, which might exclude important SNPs.

Machine learning (ML) methods are a powerful alternative to traditional methods for the analysis of gene-gene interactions. They are usually model-free and able to detect nonlinear interactions in high-dimensional datasets through supervised learning. Random Forests (24) is one such method that captures interactions between SNPs based on decision tree modeling on non-linear associations; however, it fails if neither of the SNPs have a marginal effect on the disease of interest. Therefore, it was followed by other methods that addressed this limitation, such as SNPInterforest (25) and EpiForest (26). Another method, the support vector machine, has been applied to separately interacting and non-interacting SNP pairs

© Annals of Translational Medicine. All rights reserved.

using a hyperplane (27), but it suffers from very high type-I-errors. Some variations of neural networks (NNs), such as the grammatical evolution NN (28), conduct both variable selection and statistical modeling to detect genetic interactions. Genetic programming NNs (29) and other methods, where graphs consist of nodes and arcs, have been proposed, the nodes and arcs denoting the SNPs and SNP interactions, respectively. The NN layers are a series of nonlinear statistical models, similar to regression models. NNs can be expressed as a weighted linear combination of inputs. However, no single computational or statistical method is optimal for every dataset. Moreover, the efficiency of ML methods is enhanced by limiting the number of input features, which is why it is very important to perform variable selection before searching for epistasis. Hence, there still exists the risk of missing important SNPs in the variable selection step.

In the deep-learning (DL) field, deep structured learning models are applied to high-throughput genetic data to detect and classify multi-locus SNPs. Such models provide stability, generalization, and scalability to big data through high prediction accuracy. One such DL method that has been applied to gene-gene interactions was proposed by Uppu *et al.*, where a deep feed-forward NN was trained by three hidden layers, thus displaying an improvement in the prediction accuracy from previously reported popular models (Random Forest, Logistic Regression, naive Bayes, and Gradient Boosted Machines), when applied to multiple simulated datasets (30). Each layer from the hidden layers was trained with 50 computational units and the method processed 1,000 epochs per 1,000 iterations on 10 compute nodes. The entire data was processed, by default, on every node locally by shuffling the training samples in each iteration. The model took 17.658 seconds to run 320,000 samples, which were used for training. The training speed of the model was estimated as 8,122.098 samples/second. The validation error of the model was 0.294. The error of a subsequent test set of the model was estimated as 0.661. The best DL model that was chosen by n-fold cross-validation was evaluated on a published breast cancer dataset, which predicted two-locus SNPs and SNPs with main effects that were highly associated with breast cancer. However, a major drawback of DL methods is that they are highly specialized to a specific domain, and reassessment is needed to solve issues that do not pertain to that identical domain. Also, such models are unable to understand the context of the data that they are trained with, which could be an issue while interpreting the results. A different route could be to use DL techniques by unifying them with traditional statistical methods or other ML methods to maximize the predictive accuracy. Incorporating prior information about regulatory genetic elements enabled the identification of a majority of the variants associated with amyotrophic lateral sclerosis (ALS); this strategy used a two-step hybrid of the Promoter-CNN and ALS-Net methods and yielded good classification results and high accuracy (0.67) on genome-sized data by focusing on the promoter regions (31). In this case, the dimensionality issue disappeared, as the identified regions of the genome were those that are relevant to classification of ALS patients versus healthy controls.

In the future, handling dimensionality issues will be synonymous with switching to parallel computing, which would allow the distribution of the data on several processors, thus minimizing the loss of useful and meaningful loci. PySpark, based on Python is a great language that could be used to create more scalable analyses and pipelines. The Spark data frame can be analogized with a table distributed across a cluster that has meaningful functionality. One of the major strengths of Spark is its compatibility with multiple different programming languages. For data scientists and analysts familiar with Python, the PySpark application programming interface makes it easy to write code that takes advantage of Spark to deliver dramatic improvements in processing speed for large sets of data. The future solution to dimensionality issues is tools such as PySpark, which makes it easy to take advantage of distributed computing resources in the cloud, and then bring back a smaller subset of data for further local analysis. Furthermore, because PySpark supports all standard Python libraries and even C extensions, existing code in other computing languages can take advantage of the power of Spark with only minimal modifications.

Acknowledgments

We thank Melissa Stauffer, PhD, for editing the manuscript.
Funding: This manuscript was partly supported by grants from the Ministry of Science and Technology, Taiwan (MOST-106-2314-B-002-134-MY2 and MOST-104-2314-B-002-107-MY2).

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Choi S, Bae S, Park T. Risk Prediction Using Genome-Wide Association Studies on Type 2 Diabetes. Genomics Inform 2016;14:138-48.
2. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747-53.
3. Bateson W, Mendel G. Mendel's principles of heredity. Courier Corporation; 2013.
4. Fisher RA. XV.—The correlation between relatives on the supposition of Mendelian inheritance. Earth and Environmental Science Transactions of the Royal Society of Edinburgh 1919;52:399-433.
5. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 2002;11:2463-8.
6. Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. Bioessays 2005;27:637-46.
7. Yang YF, Cao W, Wu S, et al. Genetic Interaction Network as an Important Determinant of Gene Order in Genome Evolution. Mol Biol Evol 2017;34:3254-66.
8. Phillips PC. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet 2008;9:855-67.
9. Zuk O, Hechter E, Sunyaev SR, et al. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A 2012;109:1193-8.
10. Sengupta Chattopadhyay A, Hsiao CL, Chang CC, et al. Summarizing techniques that combine three non-parametric scores to detect disease-associated 2-way SNP-SNP interactions. Gene 2014;533:304-12.
11. Bandyopadhyay B, Chanda V, Wang Y. Finding the Sources of Missing Heritability within Rare Variants Through Simulation. Bioinform Biol Insights 2017;11:1177932217735096.
12. Fan CT, Lin JC, Lee C. Taiwan Biobank: a project aiming to aid Taiwan's transition into a biomedical island. Pharmacogenomics 2008;9:235-46.
13. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, et al. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061-73.
14. 1000 Genomes Project Consortium, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56-65.
15. Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 2003;24:150-7.
16. Gola D, Mahachie John JM, van Steen K, et al. A roadmap to multifactor dimensionality reduction methods. Brief Bioinform 2016;17:293-308.
17. Lou XY, Chen GB, Yan L, et al. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. Am J Hum Genet 2007;80:1125-37.
18. Cattaert T, Calle ML, Dudek SM, et al. Model-based multifactor dimensionality reduction for detecting epistasis in case-control data in the presence of noise. Ann Hum Genet 2011;75:78-89.
19. Chung Y, Lee SY, Elston RC, et al. Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions. Bioinformatics 2007;23:71-6.
20. Gui J, Andrew AS, Andrews P, et al. A robust multifactor dimensionality reduction method for detecting gene-gene interactions with application to the genetic analysis of bladder cancer susceptibility. Ann Hum Genet 2011;75:20-8.
21. Gui J, Moore JH, Kelsey KT, et al. A novel survival multifactor dimensionality reduction method for detecting gene-gene interactions with application to bladder cancer prognosis. Hum Genet 2011;129:101-10.
22. Lee S, Kwon MS, Oh JM, et al. Gene-gene interaction analysis for the survival phenotype based on the Cox model. Bioinformatics 2012;28:i582-8.
23. Lee S, Son D, Yu W, et al. Gene-Gene Interaction Analysis for the Accelerated Failure Time Model Using a Unified Model-Based Multifactor Dimensionality Reduction Method. Genomics Inform 2016;14:166-72.
24. Botta V, Louppe G, Geurts P, et al. Exploiting SNP correlations within random forest for genome-wide association studies. PLoS One 2014;9:e93379.
25. Yoshida M, Koike A. SNPInterForest: a new method for detecting epistatic interactions. BMC Bioinformatics 2011;12:469.
26. Jiang R, Tang W, Wu X, et al. A random forest approach to the detection of epistatic interactions in case-control studies. BMC Bioinformatics 2009;10 Suppl 1:S65.
27. Chen SH, Sun J, Dimitrov L, et al. A support vector

© Annals of Translational Medicine. All rights reserved.

machine approach for detecting gene-gene interaction. Genet Epidemiol 2008;32:152-67.

28. Motsinger-Reif AA, Fanelli TJ, Davis AC, et al. Power of grammatical evolution neural networks to detect gene-gene interactions in the presence of error. BMC Res Notes 2008;1:65.
29. Ritchie MD, Motsinger AA, Bush WS, et al. Genetic Programming Neural Networks: A Powerful Bioinformatics Tool for Human Genetics. Appl Soft Comput 2007;7:471-9.
30. Uppu S, Krishna A, Gopalan RP. A deep learning approach to detect SNP interactions. JSW 2016;11:965-75.
31. Yin B, Balvert M, van der Spek RAA, et al. Using the structure of genome data in the design of deep neural networks for predicting amyotrophic lateral sclerosis from genotype. Bioinformatics 2019;35:i538-47.

Cite this article as: Chattopadhyay A, Lu TP. Gene-gene interaction: the curse of dimensionality. Ann Transl Med 2019;7(24):813. doi: 10.21037/atm.2019.12.87

© Annals of Translational Medicine. All rights reserved.