Gene-Gene and Gene-Environment Interactions Underlying Complex Traits and their Detection

Abbreviations

MDR: Multifactor Dimensionality Reduction; CART: Classification and Regression Trees; MARS: Multivariate Adaptive Regression Spline; GPNN: Genetic Programming Optimized Neural Network

Introduction

No genes or environmental factors are isolated from the interactive genomic and epigenomic networks in shaping a biological phenotype [1-3]. Non intuitivity and nonlinearity are a natural property of the network’s architecture [4] (also see an illustrative example in Box 1). Consequently, the existence of interactions among genes, called gene-gene (also known as epistatic) interactions, and between genes and environmental factors (broadly defined as all non-genetic exposures), called gene-environment (GE) interactions, is the normal rather than an exception [5-8]. Several converging lines of evidence have pointed to the dominant role of interactions in the inherited traits [6-9]; in particular, epistatic and GE interactions are considered as one of the primary culprits for missing heritability [10,11]. Identification of background-specific factors among genes in combination with lifestyles and environmental exposures is an important scientific topic in genetics, breeding, and genetic epidemiology.

A high degree of context dependence of genetic architecture likely results in a relatively weak marginal genotype-phenotype correlations for complex traits, making traditional univariate approaches that test for association one factor at a time futile [5,11]. The multi factorial strategies are thus critical in hunting highly mutually dependent factors underlying a trait. However, such a search has to face a significant obstacle called “the curse of dimensionality”, a problem caused by the exponential increase in volume of possible interactions with the number of factors to consider [15]. The conventional regression methods, established by the extension under the concept of single factor-based approaches, are hardly appropriate for tackling ubiquitous yet elusive interactions because of several problems: heavy computational burden (usually computationally intractable), increased Type I and II errors, and reduced robustness and potential bias as a result of highly sparse data in a multi factorial model [16]. Diverse novel approaches such as data mining and machine learning have been explored recently for various kinds of phenotypes [17-19], namely, Bayesian belief network [20,21], tree-based algorithms including multivariate adaptive regression spline (MARS) [22], classification and regression trees (CART) or recursive partitioning methods [23-25] and random forests approach [26,27], pattern recognition approaches including neural network strategies such as the parameter decreasing method (PDM) [28] and genetic programming optimized neural network (GPNN) [29], genetic algorithm strategies [30], and cellular automata (CA) approach [31], support vector machine (SVM) [32], penalized regression [33], and Bayesian methods [34,35].

Among these methods emerged recently, data reduction approaches (a constructive induction strategy) such as the multifactor dimensionality reduction method (MDR) [36,37], the combinatorial partitioning method [38], and the restricted partition method [39], are promising to address the multidimensionality problems. Rather than modeling the interaction term per se as with regression methods, a data reduction strategy seeks for a pattern in a combination of factors/attributes of interest that maximizes the phenotypic variation it explains. It treats the joint action as a whole, coinciding to the very original epilasis coined by Bateson [40], offering a solution that avoids decomposition as in regression methods where the number of interaction parameters grows exponentially as each new variable is added. It also has a straightforward correspondence to the concept of the phenotypic landscape that unifies biological, statistical genetics and evolutionary theories [41-45]. Notably the pioneering MDR method has sustained its popularity in detection of interactions since its launch [46].

Several extensions of the MDR have been made for analyzing different traits, e.g., binary, count, continuous, polytomous, ordinal, time-to-onset, multivariate and others, as well as combinations of those, and also entertaining various study designs including homogeneous and admixed unrelated-subject and family as well as mixtures of them [47]. Such extensions include to inclusion of covariates [48,49], to continuous traits [49], to survival data [50,51], to multivariate phenotypes [52,53], to multi-categorical or ordinal phenotypes [47,54], to case-control study in structured populations [55,56], to family study [57,58], and to unified analysis of both unrelated and related samples [59]. With these extensions, the MDR-type methods offer a powerful tool for handling the breadth of data types and
addressing statistical issues associated with study design and sampling scheme.

Despite the methodological progresses in detection of multifactor interactions, there are still difficult computational challenges and multiple hypothesis testing problems in practice, especially high-order interactions for the large-scale such as whole genome data. Both the computational time and the number of hypotheses to test increase exponentially with the number of factors to consider. The implementation may quickly become prohibitively costly when considering more than 15 factors simultaneously. Further theoretical and computational work is required for effective identification of interacting factors underlying the complex traits. Specifically, it will be worth exploring the application of the sophisticated efficient algorithms such as the branch-and-bound algorithm [60-62] and the depth-first search algorithm [63] to this field. The heuristic searches among the huge combinatorial search space such as TABU [64, 65] are also encouraged for a much reduced computational burden while getting a solution approximating but good enough for practical purposes. On the other hand, the effective correction procedures for multiple testing, rather than the rectangle-like Bonferroni-type corrections, including those for controlling false discovery rate [66-68], will play a pivotal role in avoiding a flood of false-positive claims and true hits being missed.

Box 1: An illustrative example for the nonintuitivity in a network system

The following real-life experiment on a simple series circuit in electrics, as shown in Figure 1, is used to demonstrate the natural property of a network or pathway. The series circuit contains a light bulb, a pencil, and a battery. As shown in Figure 1A, a half part of the pencil is shaved off along the long way so that the graphite center is exposed for most of the length of the pencil. Two wire ends connect with the graphite part of the pencil that will function as a resistor. One of the ends may move along the graphite, changing the resistance, and, correspondingly, the brightness of the light bulb will change. Assume the battery has a voltage of 120 volt and a resistance of 0 ohm. Suppose there are two light bulbs, 20 watt/120 volt (i.e., having a resistance of 720 ohm) and 40 watt/120 volt (i.e., having a resistance of 360 ohm), respectively.

Consider two scenarios:

1. Moving the sliding end makes two ends coincide so that the resistance of the resistor is 0 ohm.
2. Moving the sliding end to some place makes the resistance to be 14,400 ohm.

In the first scenario, the light of 40 watt will be brighter than the light of 20 watt and the output power of the former (40 watt) will be twice as much as that of the latter (20 watt). However, in the second case, the light of 40 watt will be darker than the light of 20 watt and the output power of the former (40/412 watt) is nearly a half of that of the latter (20/212 watt). This illustrative experiment supports that the existence of context-specific effects is widespread even in the simple electric pathway.

Figure 1: An experiment on a simple series circuit in electrics. A. The pictorial drawing of the circuit that consists of a light bulb, a pencil acting as a variable resistor, and a battery. B. The schematic diagram of circuit corresponding to Figure 1A, in which the resistor is a graphite rheostat.

Acknowledgement

The author thanks Guo-Bo Chen, Hai-Ming Xu, Xi-Wei Sun, and Lei Yan for their contributions to the development of GMDR. This project was supported in part by NIH Grant DA025095 to X.Y.L.

Conflict of Interest

None

References

1. Barry P (2008) No gene is an Island: Even as biologists catalog the discrete parts of life forms, an emerging picture reveals that life’s functions arise from interconnectedness. Science News 174(12): p. 22-26.
2. Szathmary E, Jordan F, Pal C (2001) Molecular biology and evolution. Can genes explain biological complexity? Science 292(5520): 1315-1316.
3. Hartwell L (2004) Genetics. Robust interactions. Science 303(5659): 774-775.
4. Nijhout HF (2003) On the association between genes and complex traits. Journal ofInvestigative Dermatology Symposium Proceedings 8: 162-163.
5. Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. Nature 429(6990): 446-452.
6. Shaoa H, Lindsay CB, David SS, Annie EH, Sheila RE, et al. (2008) Genetic architecture of complex traits: Large phenotypic effects and pervasive epistasis. PNAS 105(50): 19910-19914.
7. Huang W, Richards S, Carbone MA, Zhu D, Anholt RR, et al. (2012) Epistasis dominates the genetic architecture of Drosophila quantitative traits. Proc Natl Acad Sci USA 109(39): 15553-15559.
8. Zuka O, Eliana H, Shamil RS, Eric SL (2011) The mystery of missing heritability: Genetic interactions create phantom heritability. PNAS 109(4): 1193-1198.
9. Stratton MR, Rahman N (2008) The emerging landscape of breast cancer susceptibility. Nat Genet 40(1): 17-22.
10. Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10(4): 241-251.
11. Phillips PC (2008) Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet 9(11): 855-867.
12. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461(7265): 747-773.

13. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11(6): 446-450.

14. Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456(7218): 18-21.

15. Moore JH, Ritchie MD (2004) STUDENT JAMA, The challenges of whole-genome approaches to common diseases. JAMA 291(13): 1642-1643.

16. Carlborg O, Haley CS (2004) Epistasis: too often neglected in complex trait studies? Nat Rev Genet 5(8): 618-625.

17. Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10(6): 392-404.

18. Heidema AG, Boer JM, Nagelkerke NJ, Mariman EC, van der A DL, et al. (2006) The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. BMC Genet 7: 23.

19. Motsinger AA, Ritchie MD, Reif DM (2007) Novel methods for detecting epistasis in pharmacogenomics studies. Pharmacogenomics 8(9): 1229-1241.

20. Sebastiani P, Ramoni MF, Nolan V, Baldwin CT, Steinberg MH (2005) Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. Nat Genet 37(4): 435-440.

21. Horng JT, Hu KC, Wu LC, Huang HD, Lin FM, et al. (2004) Identifying the combination of genetic factors that determine susceptibility to cervical cancer. IEEE Trans Inf Technol Biomed 8(1): 59-66.

22. Cook NR, Zee FY, Ridker PM (2004) Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. Stat Med 23(9): 1439-1453.

23. Tracy JC, Michael DS, Mahyar S, Xiangjun G, Rishika S, et al. (2003) Use of tree-based models to identify subgroups and increase power to detect linkage to cardiovascular disease traits. BMC Genet 4(Suppl 1): S66.

24. Province MA, Shannon WD, Rao DC (2001) Classification methods for confronting heterogeneity. Adv Genet 42: 273-286.

25. Shannon WD, Province MA, Rao DC (2001) Tree-based recursive partitioning methods for subdividing sibpairs into relatively more homogeneous subgroups. Genet Epidemiol 20(3): 293-306.

26. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P (2004) Screening large-scale association study data: exploiting interactions using random forests. BMC Genet 5: 32.

27. Xiang C, Ching-Ti Liu, Meizhao Z, Heping Z (2007) A forest-based approach to identifying gene and gene-gene interactions. Proc Natl Acad Sci USA 104(49): 19199-19203.

28. Tomita Y, Tomida S, Hasegawa Y, Suzuki Y, Shirakawa T, et al. (2004) Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma. BMC Bioinformatics 5: 120.

29. Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH (2003) Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. BMC Bioinformatics 4: 28.

30. Jason HM, Lance WH, Maryln DR, Tricia AT, Bill CW (2004) Routine discovery of complex genetic models using genetic algorithms. Appl Soft Comput 4(1): 79-86.

31. Moore JH, Hahn LW (2002) A cellular automata approach to detecting interactions among single-nucleotide polymorphisms in complex multifactorial diseases. Pac Symp Biocomput 53-64.

32. Chen SH, Sun J, Dimitrov L, Turner AR, Adams TS, et al. (2008) A support vector machine approach for detecting gene-gene interaction. Genet Epidemiol 32(2): 152-167.

33. Park MY, Hastie T (2009) Penalized logistic regression for detecting gene interactions. Biostatistics 9(3): 30-50.

34. Zhang Y, Liu J (2007) Bayesian inference of epistatic interactions in case-control studies. Nat Genet 39(9): 1167-1173.

35. Nengjun Y, Brian SY, Gary AC, David BA, Eugene JE, et al. (2005) Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. Genetics 170(3): 1333-1344.

36. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, et al. (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69(1): 138-147.

37. Hahn LW, Ritchie MD, Moore JH (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19(3): 376-382.

38. Nelson MR, Kardia SLR, Ferrell RE, Sing CF (2001) A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res 11(3): 450-470.

39. Culverhouse R, Klein T, Shannon W (2004) Detecting epistatic interactions contributing to quantitative traits. Genet Epidemiol 27(2): 141-152.

40. Bateson W (1909) Mendel’s principles of heredity. Cambridge University Press, UK.

41. Nijhout HF (2008) Developmental phenotypic landscapes. Evol Biol 35(2): 100-103.

42. Wright S (1932) The roles of mutation, inbreeding, crossbreeding, and selection in evolution. Proceedings of the Sixth International Congress on Genetics 1: 356-366.

43. Rice SH (2002) A general population genetic theory for the evolution of developmental interactions. Proc Natl Acad Sci U S A 99(24): 15518-15523.

44. Wolf JB (2002) The geometry of phenotypic evolution in developmental hyperspace. Proc Natl Acad Sci U S A 99(25): 15849-15851.

45. Nijhout HF (2002) The nature of robustness in development. Bioessays 24(6): 553-563.

46. Motsinger AA, Ritchie MD (2006) Multifactor dimensionality reduction: an analysis strategy for modelling and detecting gene-gene interactions in human genetics and pharmacogenomics studies. Hum Genomics 2(5): 318-328.

47. Lou XY (2014) UGMDR: A unified conceptual framework for detection of multifactor interactions underlying complex traits. Heredity 116(2): 2589-2595.
combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. Am J Hum Genet 80(6): 1125-1137.

50. Gui J, Moore JH, Kelsey KT, Marsit CJ, Karagas MR, et al. (2011) A novel survival multifactor dimensionality reduction method for detecting gene-gene interactions with application to bladder cancer prognosis. Hum Genet 129(1): 101-110.

51. Lee S, Kwon MS, Oh JM, Park T (2012) Gene-gene interaction analysis for the survival phenotype based on the Cox model. Bioinformatics 28(18): i582-i588.

52. Choi J, Park T (2013) Multivariate generalized multifactor dimensionality reduction to detect gene-gene interactions. BMC Syst Biol 7(Suppl 6): S15.

53. Xu HM, Sun XW, Qi T, Lin WY, Liu N, et al. (2014) Multivariate dimensionality reduction approaches to identify gene-gene and gene-environment interactions underlying multiple complex traits. PLoS One 9(9): e108103.

54. Kim K, Kwon MS, Oh S, Park T (2013) Identification of multiple gene-gene interactions for ordinal phenotypes. BMC Med Genomics 6(Suppl 2): S9.

55. Niu A, Zhang S, Sha Q (2011) A Novel Method to Detect Gene-Gene Interactions in Structured Populations: MDR-SP. Ann Hum Genet 75(6): 742-754.

56. Lou XY (2012) A PCA-based generalized multifactor reduction method for correcting population stratification. Genetic Epidemiology 36(7): 753.

57. Martin ER, Ritchie MD, Hahn L, Kang S, Moore JH (2006) A novel method to identify gene-gene effects in nuclear families: the MDR-PDT. Genet Epidemiol 30(2): 111-123.

58. Lou XY, Chen GB, Yan L, Ma JZ, Mangold JE, et al. (2008) A combinatorial approach to detecting gene-gene and gene-environment interactions in family studies. Am J Hum Genet 83(4): 457-467.

59. Chen GB, Liu N, Klimentidis YC, Zhu X, Zhi D, et al. (2014) A unified GMDR method for detecting gene-gene interactions in family and unrelated samples with application to nicotine dependence. Hum Genet 133(2): 139-150.

60. Land AH, Doig AG (1960) An Automatic Method of Solving Discrete Programming Problems. Econometrica 28(3): 497-520.

61. Narendra PM, K Fukunaga (1977) A Branch and Bound Algorithm for Feature Subset Selection. Computers, IEEE Transactions on C-26(9): 917-922.

62. Lawler EL, Wood DE (1966) Branch-And-Bound Methods: A Survey. Operations Research 14(4): 699-719.

63. Cormen TH, Charles EL, Ronald LR, Clifford S (2009) Introduction to Algorithms, third Edition. 2009: The MIT Press. Pp. 1312.

64. Glover F (1986) Future paths for integer programming and links to artificial intelligence. Computers & Operations Research 13(5): 533-549.

65. Sun M, Nemati HR (2003) Tabu Machine: A new neural network solution approach for combinatorial optimization problems. Journal of Heuristics 9: 5-27.

66. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) 57(1): 289-300.

67. Storey JD (2002) A direct approach to false discovery rates. J R Statist Soc B 64, part 3: 479-498.

68. Efron B (2008) Microarrays, empirical Bayes and the two-groups model. Statistical Science 23(1): p. 1-22.