Bases, Bits and Disease

Bases, bits and disease: a mathematical theory of human genetics

Jason H Moore

European Journal of Human Genetics (2008) **16**, 143–144; doi:10.1038/sj.ejhg.5201936; published online 31 October 2007

The field of human genetics has recently become overwhelmed with the need for information management and analysis. Genome-wide association studies, transcriptional profiling and proteomics have all contributed greatly to the increased need for expertise from the information sciences. One significant challenge for human genetics is to identify those genetic, genomic, proteomic and environmental factors that increase or decrease susceptibility to disease. This is a difficult challenge due to the enormous volume of information that is often noisy and the complexity of the genotype-to-phenotype mapping relationship that arises from nonlinear phenomena such as epistasis or gene–gene interaction. It is increasingly clear that an analytical retooling to confront both the volume and complexity of the data is critical for moving the field forward.[1] The parametric statistical paradigm has served us well over the years but now is the time to explore a wide range of different analytical methods for complex problem solving from a diversity of fields such as computer science, mathematics, etc. The paper by Dong *et al*[2] on page 229 of this issue explores the use of information theory for the genetic analysis of complex human diseases.

Information theory was launched as a formal discipline in 1948 with the publication of Shannon's[3,4] two papers on 'A Mathematical Theory of Communication'. Shannon at that time worked for Bell Labs and was interested in mathematical methods for encoding information for transmission through electronic signals. The basic problem is to encode a message, transmit it as a signal, receive it and decode it with minimal noise such that the original message is not lost. It was in these seminal papers that Shannon introduced and defined entropy as a measure of uncertainty to help with maximizing the efficiency and accuracy of encoding, sending, receiving and decoding a message. Consider the following simple genetic example as a means for explaining entropy. Assume that 100 cases with a particular disease and 100 healthy controls were sampled from a population for an association study and a bi-allelic single-nucleotide polymorphism (SNP) in a particular candidate gene is genotyped. A fundamental question is whether any information about who is a case and who is a control can be obtained from knowledge about genotype. Let us assume that the *AA* and *aa* genotypes are each represented in 25 cases and 25 controls, while the *Aa* genotype is represented in 50 cases and 50 controls. With this set of data it is clear that no information about case–control status can be gained by looking at the genotypes and thus there is maximum uncertainty. With a binary outcome like case–control status entropy is simply $[-p \times \log_2 p]-[(1-p) \times \log_2(1-p)]$, where p is the probability of being a case (p) within a genotype and $(1-p)$ is the probability of being a control within a genotype. Thus, for our simple example, the entropy for genotype *AA* would be $-[(0.5*-1)+(0.5*-1)]=1$. Consider the example where the *AA* genotype is represented in 50 cases and 0 controls. In this case the entropy would be 0, indicating maximum knowledge about who is a case and who is a control. Another important concept is information, which is simply 1−entropy. In our first example, information is 0 (the minimum) and in our second example information = 1 (the maximum). Using the communications analogy, the case–control status (1s and 0s) of each individual in the data set forms a message that could be sent to a collaborator, for example. The 1s and 0s in the message could be coded as genotypes *AA*, *Aa* and *aa*, and transmitted via e-mail, for example, to the collaborator along with a genetic model that assigns high- and low risk to each genotype. The collaborator could then decode each individual's genotype and re-assign case–control status using the genetic risk model as a key. The entropy or uncertainty of these assignments could be used as a measure of the quality of the information encoding and the quality of the transmission, which might be susceptible to the introduction of noise from random events or systematic errors.

Parameterization of genetic association studies in this manner is useful because there is a nice foundation of mathematical theory for information sciences and there are many useful data analysis tools that make use of entropy and other information-based measures. In addition, casting the problem as an information theory problem provides a parallax view that is often very useful when confronting a complex problem. The paper by Dong *et al*[2] introduces the Entropy-Based SNP–SNP Interaction Method (ESNP2) for detecting and modeling epistasis in genetic association studies. This approach consists of three steps. First, the entropy of the data set (cases and controls) is computed as a baseline. Second, the information gained about case–control status from knowledge of the genotypes at each single SNP is estimated. Third, the information gained about case–control status from knowledge about genotypes gained by combining two SNPs as a Cartesian product is estimated. This final step provides a measure of the interaction information for any pair of SNPs and thus can be used to assess the presence of epistasis in a genetic association study. The concept of interaction information was described by McGill[5] and has been rediscovered every few years since. Jakulin and Bratko[6] provide the most recent comprehensive evaluation of information theory to detect, visualize and interpret nonadditive interactions between variables. Moore *et al*[7] have adapted recently these entropy-based interaction analysis methods as a way to detect and interpret epistasis in case–control studies. The novelty of the study by Dong *et al*[2] is the comparison of interaction results with two-locus epistasis models such as those described by Li and Reich.[8] The paper includes a nice application of the ESNP2 approach to sickle cell anemia and malaria where they find significant evidence for interaction. A nice feature of this study is the availability of open-source software that can be used to implement the method with other studies.

As the analysis of epistasis and its role in the genetic architecture of common diseases becomes more popular, there are several important challenges that need to be addressed. First, we need an analytical retooling to embrace the complexity of the genotype–phenotype mapping relationship.[1] Second, epistasis analysis on a genome-wide scale is difficult due to the combinatorial magnitude of the problem.[7] Finally, making biological inferences about cellular processes from statistical models of epistasis derived from population data is extremely difficult.[9,10] The studies by Dong *et al*[2] and others[7] suggest that information theory has an important role to play in developing mathematical theories of human genetics■

JH Moore is at the Computation Genetics Laboratory, Dartmouth-Hitchcock Medical Center, One Medical Center Drive, 706 Rubin Building, HB7937, Lebanon, NH 03756, USA.
Tel: +1 001603 653 9939;
Fax: +1 001603 653 9900;
E-mail: jason.h.moore@dartmouth.edu

References

1 Thornton-Wells TA, Moore JH, Haines JL: Genetics, statistics and human disease: analytical retooling for complexity. *Trends Genet* 2004; **20**: 640–647.

2 Dong C, Chu X, Wang Y *et al*: Exploration of gene–gene interaction effects using entropy-based methods. *Eur J Hum Genet* 2008; **16**: 229–235.

3 Shannon CE: A mathematical theory of communication. *Bell Sys Tech J* 1948; **27**: 379–423.

4 Shannon CE: A mathematical theory of communication. *Bell Sys Tech J* 1948; **27**: 623–656.

5 McGill WJ: Multivariate information transmission. *Psychometrika* 1954; **19**: 97–116.

6 Jakulin A, Bratko I: Analyzing attribute interactions. *Lect Notes Artif Intell* 2003; **2838**: 229–240.

7 Moore JH *et al*: A flexible computational framework for detecting, characterizing and interpreting patterns of epistasis in genetic studies of disease susceptibility. *J Theor Biol* 2006; **241**: 252–261.

8 Li W, Reich J: A complete enumeration and classification of two-locus disease models. *Hum Hered* 2000; **50**: 334–349.

9 Moore JH, Williams SM: Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays* 2005; **27**: 637–646.

10 Moore JH: A global view of epistasis. *Nat Genet* 2005; **37**: 13–14.