Does our DNA keep us awake?

Cüneyt
<cuneyt.guzey@ntnu.no>

Daniela
<daniela.bragantini@ntnu.no>

Luca
<piero.mana@ntnu.no>

Yasser
<yasser.roudi@ntnu.no>

Draft of 11 November 2018 (first drafted 22 August 2018)

abstract

1 Problem setup

Every single-nucleotide polymorphism (SNP), together with the huge variety of external factors, can in principle affect the appearance of a disease's symptoms. It is extremely complicated to identify and untangle these causal mechanisms and interactions and to ascertain their degrees, although their causal graph (Pearl 2009) is easy to draw (fig.**). The interacting mechanisms represented by the arrows are difficult to study, and the external factors X are innumerable and largely unknown.

An indication of the causal strength of one or more SNPs on one or more symptoms can be obtained by replacing the causal graph with a corresponding simplified Bayesian network (Pearl 2009) of *conditional probabilities* (fig.**). The external factors X disappear from the graph but their presence is implicit in the probabilistic relation between the nodes; the latter also accounts for the complexity of the causal mechanisms and our uncertainty about them. Consider a particular SNP and a symptom S. In an arbitrarily large population, if the individuals having one allele and those having the other allele show markedly different *conditional frequencies* of incidence of the symptom, then we can conclude that the SNP must have some causal relevance, however indirect, for the symptom.

These conditional frequencies are far easier to study than causal relations, given the conditional frequencies in a population sample.

In this study we show how to quantify our degrees of belief about such conditional frequencies in an arbitrarily large population, given (1) the conditional frequencies in a population sample, (2) a representation

of our initial information about such frequencies. The idea behind the calculation is simple: using Bayes's theorem we have

$$\mathrm{p}(\text{frequencies}|\,\text{sample data}, \text{initial info}) \propto$$
$$\mathrm{p}(\text{data}|\,\text{frequencies}, \text{initial info}) \times \mathrm{p}(\text{frequencies}|\,\text{initial info}). \quad (1)$$

The first degree of belief in the product above is given by a simple sampling formula; the second can be modelled in several ways, but if the sample data are enough it will lead to basically the same final degrees of belief about the conditional frequencies.

Once we have a quantified distribution of belief about the conditional frequencies in an arbitrarily large population, it is easy to see whether we expect the latter to be significantly different for different alleles. See for example the distributions in fig. 1: from the sample data we expect conditional frequencies 0.259 and 0.284 for the symptom conditional on the two alleles, with standard deviations 0.008 and 0.006. From the two belief distributions we can even calculate our belief that the two conditional frequencies are equal to within some range; in the case of the figure our belief that the two frequencies are the same within 0.01 is 3.2%. Many other quantifications are possible; for example our belief that a conditional frequency lies between two particular values, and so on.

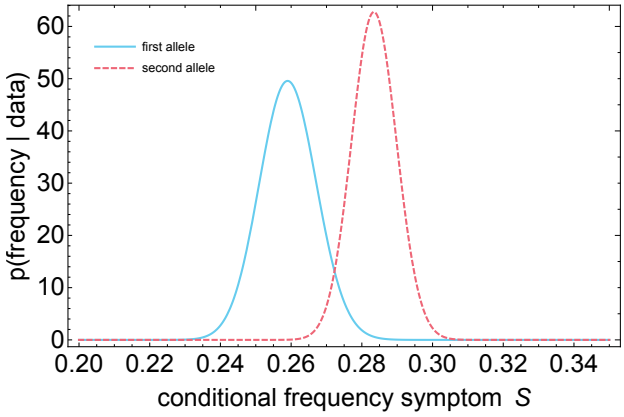

Figure 1 Example of distributions of belief

✠ add something about dependence of broadness on sample size, and 'smoothing' as discussed by MacKay & Bauman Peto (1995 § 2.6).

This approach is not dichotomous, unlike a 'significance' test. Rather, we will find a graduation of cases: from frequencies predicted to be clearly distinct, to frequencies with uncertainties too large for drawing definite conclusions. These cases can be sorted, obtaining a sequence of SNPs with a decreasing belief of causal association with the symptom. How many of these SNPs are to be selected for further study depends on one's experimental and computational resources.

1.1 Summary of the main formulae

We have a sample of size n. We check the subsample of individuals that have a particular allele, say Bx, for a particular gene, say rs697680_A. Suppose that in this subsample n_0 individuals *don't* show symptom A and n_1 *do* show symptom A. This also means that the size of our subsample (individuals with allele Bx) is $n \coloneqq n_0 + n_1$.

Our degree of belief about the frequency f_1 of symptom A among the individuals with allele Bx in an *infinite* population is a Beta distribution with parameters $n_0 + \theta_0$, $n_1 + \theta_1$, with $\theta \coloneqq \theta_0 + \theta_1$:

$$\mathrm{p}(f_1 | \, n_0, n_1, \theta_0, \theta_1) \, \mathrm{d}f_1 = \frac{\Gamma(n+\theta)}{\Gamma(n_0+\theta_0)\,\Gamma(n_1+\theta_1)} \, (1-f_1)^{n_0+\theta_0-1} \, {f_1}^{n_1+\theta_1-1} \, \mathrm{d}f_1 \quad (2)$$

This distribution has expected value and variance

$$\begin{aligned} \mathrm{E}(f_1 | \, n_0, n_1, \theta_0, \theta_1) &= \frac{n_1+\theta_1}{n+\theta}, \\ \mathrm{var}(f_1 | \, n_0, n_1, \theta_0, \theta_1) &= \frac{(n_0+\theta_0)\,(n_1+\theta_1)}{(n+\theta)^2\,(n+\theta+1)}. \end{aligned} \quad (3)$$

✠ Possible further developments: use of hyper-Dirichlet priors, use of graphical models to infer causal relationships (Pearl 2009)

Thanks

PGLPM thanks Mari & Miri for continuous encouragement and affection, and to Buster Keaton and Saitama for filling life with awe and inspiration.

To the developers and maintainers of LaTeX, Emacs, AUCTeX, Open Science Framework, Python, Inkscape, Sci-Hub for making a free and unfiltered scientific exchange possible.

Bibliography

('de *X*' is listed under D, 'van *X*' under V, and so on, regardless of national conventions.)

MacKay, D. J. C., Bauman Peto, L. C. (1995): *A hierarchical Dirichlet language model*. Nat. Lang. Eng. **1**[3], 289–307.

Pearl, J. (2009): *Causality: Models, Reasoning, and Inference*, 2nd ed. (Cambridge University Press, Cambridge). First publ. 2000.