The Causal Mediation Analysis in Genomic Data

Giulia Malaguti, Séverine Affeldt, and Hervé Isambert

UMR 168 CNRS-UPMC, Institut Curie, 26 rue d'Ulm, 75248 Paris, France
giulia.malaguti@curie.fr

Keywords: Causal Inference Methods, Mediation Analysis, Direct *versus* Indirect Effects, Genomic Data, Whole Genome Duplication

1 Introduction

The increasing amount of high throughput data has allowed to reveal that many genomic properties are to some extent correlated, suggesting direct statistical associations. Yet, these correlations could in fact be mediated by the indirect effect of third properties and only provide partial insights on complex biological systems. To quantify direct and indirect effects in genomic data, we performed Mediation analyses following the approach of Pearl[1], a method typically used in social sciences and epidemiology. We present here the extension of the Mediation framework to the case of more than one mediating variable, and its application to the genomic properties implicated in the biased retention of whole genome duplicated genes, so-called ohnologs, in the course of vertebrate evolution.

2 The Mediation Framework

The Mediation analysis assesses the importance of a mediator M in transmitting the indirect effect of a variable X on the response variable Y. In this framework, the average direct effect $DE_{xx'}$ is defined as the change on Y due to a change of X from the initial value x to x', while keeping M fixed at the value $m(x)$. Similarly, the average indirect effect $IE_{xx'}$ is defined as the change on Y if M could be changed from the value $m(x)$ to $m(x')$ while keeping X to its original value x[1]. The Mediation formulae can be rewritten in terms of the expectation value $E(Y|x, m)$ of Y given X and M, and $P(M|x)$, that denotes a value of M drawn from a distribution P conditionally on X,

$$DE_{xx'} = Y(x', m(x)) - Y(x, m(x)) = \sum\nolimits_m [E(Y|x', m) - E(Y|x, m)]P(M|x)$$

$$IE_{xx'} = Y(x, m(x')) - Y(x, m(x)) = \sum\nolimits_m E(Y|x, m)[P(M|x') - P(M|x)]$$

$$TE_{xx'} = Y(x', m(x')) - Y(x, m(x)) = \sum\nolimits_m E(Y|x') - E(Y|x)$$

If X, M, and Y are binary variables, expectation values of Y and values of M drawn from a conditional distribution P can be estimated from the counts in the data of the triplets corresponding to each possible combination of values for (X, M, Y). This means that all the effects can be directly derived from the number n_{xmy} of triplets where $(X, M, Y) = (x, m, y)$, for each value of (x, m, y)[1].

2.1 Extension to Many Mediators

In order to study several correlated properties in genomic data, we extended the single mediator formulae to the case of more than one mediator. The simplest generalization consists in gathering all the k mediators for the ordered pair of variables (X, Y) in a single super mediator $M = \{M_1, M_2, \ldots M_k\}$. This approach allows to discriminate the direct effect $DE_{xx'}$ from the effects through any other indirect path connecting X and Y.

However, it is fundamental to identify the correct set of mediators for an ordered pair of variables (X, Y) in a causal graph. The mediators are the variables that contribute to the correlation between X and Y, and belong to indirect paths between X and Y that do not include any 'v-structure' $(X... \rightarrow M \leftarrow ...Y)$. The erroneous inclusion of non mediators in the super mediator M creates artificial correlations between X and Y. By iteratively collecting the mediators of each edge of a graph, we developed an efficient algorithm to recover in polynomial time the correct set of mediators for each pair of variables in a graph.

3 The Mediation Analysis Applied to Genomic Data

We considered the 20,506 protein-coding genes in human. Based on data mining analyses[2], we identified several gene properties, such as the essentiality or the implication in cancers, and we inferred the underlying causal graph linking these properties. We applied the extended Mediation analysis formulae to specifically disentangle the direct from the indirect effects of properties favoring the specific retention of ohnologs[2] during vertebrate evolution[3, 4]. Our results demonstrate that the retention of human ohnologs is surprisingly more strongly caused by their susceptibility to dominant deleterious mutations than their interactions within multi-protein complexes, unlike frequently invoked in the field.

3.1 Conclusion and Perspectives

Our study highlights the need to go beyond statistical correlations in the analysis of genomic data and to rely on more advanced inference methods to disentangle indirect from direct interaction pathways. The application of this approach to other biological properties emerging from genomic data analysis will cast new light on the function and evolution of biological properties and networks.

References

1. Pearl, J.: The mediation formula: A guide to the assessment of causal pathways in nonlinear models. Causality: Statistical Perspectives and Applications. John Wiley and Sons, Ltd, Chichester, UK (2012): 151-179.
2. Singh, P-P. *et. al*: On the Expansion of Dangerous Gene Repertoires by Whole-Genome Duplications in Early Vertebrates. Cell Reports, Vol. 2, Issue 5 (2012)
3. Ohno, S.: Evolution by gene duplication. Springer, New York, (1970).
4. Putnam, N.H. *et. al*: The amphioxus genome and the evolution of the chordate karyotype. Nature 453, 1064-1071, (2008).