Peertechz

Life Sciences Group

Open Journal of Bioinformatics and Biostatistics

DOI CC By

Ahmed Rebai*

Centre of Biotechnology of Sfax, Post Box 177, 3018 Sfax, Tunisia

Dates: Received: 09 June, 2017; **Accepted:** 26 June, 2017; **Published:** 27 June, 2017

***Corresponding author:** Ahmed Rebai, Centre of Biotechnology of Sfax, Post Box 177, 3018 Sfax, Tunisia, E-mail: ahmed.rebai@gmail.com; Ahmed.Rebai@cbs.rnrt.tn

https://www.peertechz.com

Review Article

Causality in Genomics Studies: Time is ripe for a new Paradigm

Introduction

Biological sciences have been for so long dominated by observational approaches as shown by the scarcity of studies trying to infer causality from association, starting from the historical and very instructive studies of Pierre Louis (1835) on the efficacy of the standard treatments for pneumonia and John Snow (1855) on the causal relationship between cholera and water contamination [1].

Several paradigms have occurred in successive waves during the last two centuries and they have moved life sciences from empiric to experimental (typically hypothesis-driven) and then to data-driven, concretized by the recent entry in the "big data" era. The value of large, complex and linkable information generated by genomics and other -omics fields and large scale epidemiologic studies has resulted in the domination of associational reasoning, where scientists look for association/correlation between a large number of diverse and heterogeneous attributes (genes, proteins, polymorphisms, phenotypes). This associational approach based on observational data has dominated genomics studies because of the scarcity of experimental data, where some variables are controlled (fixed to some values) or manipulated and their effects on other variables are measured. The lack of 'experimental' interventional data in genomics is mainly due to the logistical feasibility and high cost of their production, which requires genes and proteins experimental perturbations (gene knockout for example).

The analyses of data from observational genomics studies are associational by nature, when one looks to make some kind of inference and knowledge discovery from patterns of association using statistical and data mining techniques. Most of these techniques, even those that model relationships between dependent and independent variables (such linear/logistic regression), have limited interest in inferring causality which means establishing a cause-effect relationship among variables [2], for a general and wide review on the subject). In fact, to make causal inference, we must assume that equations of correlations linking variables are invariant under proposed interventions and it is problematic to verify such assumptions without making interventions. Moreover, if the model changes when variables are subject to intervention (rather than being simply measured) then this model is of poor utility for predicting results of interventions. Although this conceptual hurdle is well understood by most investigators in life sciences, the use of such models is still widespread despite the high risks of misinterpretation [1].

Although huge efforts have been made to separate 'signal' from 'noise' in associational analyses of genomics data by requiring replication and imposing more stringent criteria for statistical significance, the major weakness of this approach is that it has not allowed yet to show which of these associations have meaning, and particularly a causal meaning. Observational data are sensitive to many biases, such as selection, confounding, latent unobserved variables and lack of generalizability. To overcome these problems a new reasoning framework was needed, providing an iterative process of interpreting what we know and what we need to know. This management, synthesis and translation of data to knowledge need methods and algorithms that can use observational data to extract some information on causal relationships that can be thereafter confirmed by experimental approaches [3,4].

A outlined by Pearl [2], there are two mental barriers in causal inference that have slowed the development of an appropriate reasoning framework for this purpose; the first is that causality cannot be tested from observational data alone and the second is that it cannot be inferred based on classical probability calculus (and the statistical approaches based on it). To overcome these barriers, he developed the *do-calculus* framework for the identification of causal effects in non-parametric models.

Using this framework, Maathuis et al. [3], in a seminal paper showed that when making some assumptions on the distribution of the data, one can estimate bounds on causal effects of variables on a target variable using graphical models.

In this short review paper I present the rationale behind this method and give examples of its application for inferring

Citation: Rebai A (2017) Causality in Genomics Studies: Time is ripe for a new Paradigm. Open J Bioinform Biostat 1(1): 010-014.

causality in gene expression data. I then discuss its possible application to microbiome data.

Theoretical Issues

What is meant by causality in biology?

Mayr [5], gave the following pragmatic definition of cause in functional biology: "a member of a set of jointly sufficient reasons without which the event would not happen". A typical causal question in genomics is whether a given gene acts causally on the expression of another gene or whether a given gene mutation or a genotype is a cause of a phenotype. One way to check this causal relationship is to make the gene 'disappear' (this can be done by what is called knock-out experiments) or fix it to some genotype and see how the phenotype changes. However, since most of these experiments cannot be done on humans and are very heavy in models organisms, establishment of causality in biology is a big challenge. Even if these experiments could be done, causality is always difficult to demonstrate due to the high complexity of living systems and as Mayr [5] noted "in view of the high number of multiple pathways possible for most biological processes and in view of randomness of many of the biological processes particularly on the molecular level, causality in biological systems is not predictive or at best is only statistically predictive".

What is the difference between observational and experimental data?

We call observational data , data that are gathered by observing/measuring variables of a biological system (tissues, cells, bacteria..) under some conditions, for example expression or methylation levels of genes, genotypes for a set of markers in diseased and control individuals.

Experimental data, also called interventional data, arise when the researchers apply some treatments to a biological system in a randomized controlled experiment and look to the effect of such treatments on a response variable.

Why estimating causal effect from observations is very useful?

In genomics studies, experimental data are sometimes infeasible, unethical and often time consuming and expensive while observational data are cheaper, more abundant and available. Thus developing methods that can answer causal questions, at least partially, about biological systems from observational data only is very useful. However, moving from the observational world to the experimental world, i.e. from pre-interventional to post-interventional distribution of data, is impossible without some assumptions.

One of these assumptions is to consider that data were generated from a causal structure that can be modeled by a directed acyclic graph (DAG) either it is known or to be learned from the data. In the following paragraphs, I gave brief presentation of the methods that can be used to achieve this task.

What is DAG?

A graph is a set of nodes (variables) connected by a set of, directed or indirected, edges; in expression data nodes are genes whose levels of expression are measured while edges might mean regulation relationships (regulation means that a gene inhibits, stimulate or moderate the expression of another gene). A graph is called complete if every pair of nodes is connected by an edge and we call the *parents* of any node X (denoted $Pa(X)$) all those nodes that are connected to X with a directed edge pointing to X. A Directed Acyclic Graph (DAG) is a graph where there exist no path (a set of edges that go from one node to another) linking any node X to any node Y that can go from X to Y by a directed set of edges and then back from Y to X by another set of directed edges.

What is a structural equation model (SEM) and what are its properties?

A SEM is a model that describes how any variable X (observed) is generated from the set of all or part of the other variables i.e. how changes in these variables lead to a change in X, added to some random noise. The SEM can be represented as a DAG, where the edges are drawn such that a set of variables other than X is the parent set of X. This graph is called the causal graph since an edge that goes from Y to X means that Y is a direct cause of X. Assuming that the causal structure is a DAG means that we do not allow feedback loops or unmeasured confounding variables. We also assume that the errors are jointly independent.

Given these assumptions one can show[2] that the joint density function of all nodes $f(x_1,x_2,..x_p)$ can be factorized as the product over all nodes of the conditional densities of each node X_i given its parents : $f_i(x_i/Pa(x_i))$.

How the SEM framework allows to move from observational to experimental worlds?

Given the previous, an intervention at some variable X_i is simply equivalent to changing the generating mechanism of X_i that is to change the corresponding structural equation relative to X_i without changing the others. More clearly, setting the variable X_i to some value x'_i (expressed as $do(X_i=x'_i)$ in the language of the do-calculus is interpreted as an outside intervention [2]. For example if X_i and X_j denote the expression level of genes i and j then the expectation of X_j given that X_i is fixed to some value x'_i ($E(X_j/do(Xi=x'_i))$ represents the average expression level of gene j after setting the expression level of gene i to the value x'_i by an outside intervention.

So a do-intervention on X_i means that X_i no longer depends on its former parents in the DAG, so that incoming edges into X_i can be removed and we get what is called a truncated DAG, leading to a truncated factorization of the post-intervention distribution:

$$F(x_1,..x_p/do(X_i=x'_i))=\Pi_{j\neq i} f_j(x_j/pa(x_j)) \text{ if } x_i=x'_i \text{ and o otherwise}$$

Citation: Rebai A (2017) Causality in Genomics Studies: Time is ripe for a new Paradigm. Open J Bioinform Biostat 1(1): 010-014.

How can we define a causal effect?

If we have discrete variables, like an expression status of a gene (over expressed versus under expressed), we can define the causal effect of gene i on gene j by $P(X_j = 1/do(X_i=1)) - P(X_j = 1/do(X_i=0))$. If the variables are measured on a quantitative scale then the total causal effect of X_i on X_j can be defined as:

$$\frac{\partial}{\partial x} E(X_j / do(X_i = x)|_{x=x'_i}$$

Based on the factorization formula and assuming that the joint distribution of all variables is Gaussian means that $E(X_j/x'_i, pa(X_i))$ is linear in x'_i and $pa(X_i)$ and thus the causal effect of X_i on X_j is the regression coefficient of X_i in the linear regression of X_j on X_i and $pa(X_i)$ when X_j is not a parent of X_i and 0 otherwise. Note that the causal effect computed by the formula above does not depend on x'_i and can be interpreted (as any linear regression coefficient) as the average increase of the variable X_j when X_i increases by one unit: $E(X_j/do(X_i = x'_i+1)) - P(X_j/do(X_i = x'_i))$.

How can we compute practically the causal effect when the causal DAG is not known?

In the previous developments we assumed that the DAG was known. However in most applications this is not the case, so that we need to learn the DAG from the observational data. This can be done by several structure learning methods [6, 7]. Since the causal DAG cannot be uniquely determined, what we can learn is in fact what is called a completed partially DAG or a CPDAG. A CPDAG is in fact a single representation of a family of many possible and equally likely (we call these Markov equivalent) DAGs; if an edge between two nodes is always directed in the same direction in all DAGs then it is directed in the CPDAG, and if not it is undirected in the CPDAG. One of the most used and fast algorithms to find the CPDAG is the PC algorithm [8].

Given a CPDAG one can perform simple linear regression analysis of each response node of interest over any other node and its parents to estimate its causal effect on the response node for all DAGs contained in the CPDAG. This gives a set of causal effect estimates, so by taking the minimal absolute value of those estimates we obtain a consistent estimate of the lower bound of the causal effect. This approach, developed by Maathuis et al. [3], was named IDA (Intervention–calculus when de DAG is absent).

Thus IDA proceeds in four step to estimate the causal effect of every node X on a response node of interest, Y:

(1) First estimate the CPDAG of the underlying causal DAG using the PC algorithm [8].

(2) List all the DAGs in the family depicted by the CPDAG

(3) Estimate the causal effect of X on Y for each DAG as previously explained

(4) Collect all the values of the causal effect and obtain a lower bound of the causal effect as the minimal absolute value of effects estimated over all possible DAGs.

Since step (2) is computationally intensive and that the estimation of the effect of X on Y only needs to know the parents of X in each DAG, we can only extract this information quite easily from the CPDAG.

The IDA method assumes, besides the two previously stated assumptions (no feedback loops in the causal model and that the variables have a joint multivariate Gaussian distribution), that there is no hidden variables.

Is there any computer program to use IDA?

IDA has been implemented in an R package named *pcalg* [9].

Can we use IDA to estimate joint causal effects of many variables at a time?

A generalization of the IDA method to estimate the effect of multiple simultaneous interventions (such as knocking out many genes in expression analyses) has been developed and named JointIDA [10]. There are also some other methods based on Gaussian Bayesian networks that can jointly estimate causal effect from observational and interventional data [11].

Is there any extensions to IDA?

Stekhoven et al. [12], presented an extension of IDA that they called Causal Stability Ranking (CStaR), where they estimate the stability of estimates of causal effects using resampling from the data and ranking of the variables according to their effects. They showed that their method improves the true positive rate as compared to IDA and largely outperforms classical high-dimensional regression methods. Teramoto et al. [13], proposed an extension of IDA to deal with non-Gaussian distribution of the data that they call the non-paranormal method (NPN-IDA). In this method a non-paranormal correlation coefficient is used within the PC algorithm (instead of the Pearson correlation coefficient) to learn the CPDAG. They showed, based on application to real data, that this improves learning of the DAG, thus leading to more accurate estimation of causal effects.

Several other extensions of the IDA method have been proposed to deal with the presence of hidden variables and feedback loops using appropriate algorithms for the learning of the causal graphical [14-16]. Some of these algorithms are available in the *pcalg* R package. There are also extensions of IDA to deal with time series data [17], heteregeous data, like when one is using a mix of observational and experimental data sets [18], different datasets with overlapping sets of variables [19] or a combination of both [20].

Selected Applications

Gene expression

In their seminal paper Maathuis et al. [3], gave an application of IDA on evaluating the causal effect of the expression level of 4088 genes (continuous variables) on riboflavin production by *Bacillus subtilis*. Since the number of observations (71 strains) is very small compared to the number of variables (what is

Citation: Rebai A (2017) Causality in Genomics Studies: Time is ripe for a new Paradigm. Open J Bioinform Biostat 1(1): 010-014.

called small *n*, big *p* problem), the basic IDA method for causal DAG learning cannot be applied so a local variant of IDA was proposed [3], based on many bootstrapping rounds from the original dataset. They showed that within the top ten most causal genes identified by IDA only one was identified by standard regression techniques in high-dimensional problems, like Lasso [21], which are not able to estimate causal effect but only statistical association effects. In their second paper [4] applied IDA to the classical dataset of Hughes et al. [22], on gene expression in yeast which contain both observational and interventional data (observational: 5361 genes for 63 wild-type cultures; interventional: 5361 genes for 234 single gene deletion mutant strains). They showed that IDA largely outperformed high-dimensional regression techniques (like Lasso and Elastic-net [23], and identified causal effects of knock-out genes that are consistent with the predictions of interventional experiments. Stekhoven et al. [12], described two applications of the modified IDA method (CStaR); the first is the yeast dataset of Hughes et al. [22], where they showed improvement of CStaR over standard IDA. The second dataset is to estimate the causal effect of the expression profiles of genes measured in 47 natural accessions from diverse geographic origins on time to flowering in the model plant *Arabidopsis thaliana* [24]. They showed that the gene identified as having the highest causal effect (SOC1) was one of the major genes that have been shown by interventional experiment to regulate flowering time and two other major flowering time regulator genes were in the top 20 causal genes (FRI and FLC).

Le TD et al. [25], applied IDA to the epithelial-to-mesenchymal transition (EMT) datasets in order to estimate causal effects of 43 microRNA on RNA expression of 1635 genes in 60 cancer cell lines [26]. The causal effects were calculated for each pair of miRNA and mRNA and genes regulated by each miRNA (with non-zero causal effect of this miRNA) were ranked based on their absolute value. The authors then validated, by further experiments, the genes that were predicted based on causal effects to be regulated by miR-200 family (miR-200a and miR200-b). They found a significant overlap of the top genes regulated by miR-200 and those predicted by IDA. The target genes predicted by causal analyses were found to be enriched in the five functional classes known to be critical to EMT: cellular movement, cell-to-cell signaling and interaction, cellular assembly and organization, cellular growth and proliferation and cell morphology.

Microbiome data

Recent studies have suggested the role of some bacterial species in the gut microbiota in colon cancer. Two recent studies have reported a highly significant association between the rate of some bacterial species (evaluated by 16sRNA next generation sequencing) and colon cancer risk [27,28]. In a recent work (submitted), we used IDA to calculate causal effects of the bacterial species rate on cancer risk. Since, the response variable here is the patients status (cancer versus control), which is a binary variable, we used linear regression as an approximation considering the response as a dummy variable. Another solution (currently under consideration) is to use binary logistic regression (a generalized linear regression model) but we need to demonstrate that, in this case, the good properties of IDA estimate still hold.

We found that *Fusobacterium nucleatum* species have the highest causal effect on cancer risk, which is in good agreement with several reports that described the possible mechanisms by which these species can enhance cancer. However, differently from gene expression studies where knockout experiments can allow to validate causal effects, experiments with microbiota, even in model organisms, are very difficult since the microbiota itself is an extremely complex and dynamic system and its enrichment with some specific species might not be possible to validate their effect on some response variable.

Conclusion

Estimating causal effects is a very important issue in modern genomics since it will allow us to move from the current paradigm of associational studies to a new paradigm, where we can predict with high accuracy and confidence, without the need for interventional studies (which are not ethically possible in Humans and very expensive in models organism) the function of genes and proteins and their effect on phenotypes.

The IDA method (or family of methods) , based on the *do-calculus* and causal graphs, offer a first step on this long road that will allow us to infer, with high precision and significant, biological meaning, causal effects from observational data. Further efforts are needed to develop new conceptual frameworks and efficient algorithms for causality inference from genomics data because without this the clinical translation of the findings will still remain far from reach.

References

1. Friedman D (1999) from association to causation: some remarks on the history of the statistics. Statistical Science 14: 243-258. Link: https://goo.gl/jUXYsv
2. Pearl J (2009) Causality: Models, Reasoning and Inference. Cambridge University Press Cambridge second edition. Link: https://goo.gl/ynyFD4
3. Maathuis MH, Kalisch M, Buhlmann P (2009) Estimating high-dimensional intervention effects from observational data. Ann Statist 37: 3133-3164. Link: https://goo.gl/Wj7kYe
4. Maathuis MH, Colombo D, Kalisch M, Buhlmann P (2010) Predicting causal effects in large-scale systems from observational data. Nature Methods 7: 247-248. Link: https://goo.gl/R3sBeN
5. Mayr E (1961) Cause and effect in biology. Science 134: 1501-1506. Link: https://goo.gl/wmjrqS
6. Neapolitan RE (2004) Learning Bayesians networks. Pearson 674. Link: https://goo.gl/ioYYBx
7. Rebai A (2010) Bayesians Networks. Intech 442. Link: https://goo.gl/AV77fq
8. Spirtes P, Glymour C, Scheines R (2000) Causation, prediction, and search. MIT Press Cambridge second edition. Link: https://goo.gl/a6cVZ6
9. Kalisch K, Machler M, Colombo D, Maathuis MH, Buhlmann P (2012) Causal inference using graphical models with the R package pcalg. J Stat Soft 47: 1-26. Link: https://goo.gl/GBEUkx

Citation: Rebai A (2017) Causality in Genomics Studies: Time is ripe for a new Paradigm. Open J Bioinform Biostat 1(1): 010-014.

10. Nandy P, Maathuis MH, Richardson TS (2014) Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. Link: https://goo.gl/hza9zg

11. Rau A, Jaffrézic F, Nuel G (2013) Joint estimation of causal effects from observational and intervention gene expression data. BMC Systems Biology 7: 111. Link: https://goo.gl/fs48bm

12. Stekhoven DJ, Moraes I, Sveinbjörnsson G, Hennig L, Maathuis MH, et al. (2012) Causal stability ranking. Bioinformatics 28: 2819-2823. Link: https://goo.gl/1X9a1J

13. Teramoto R, Saito C, Funahashi S (2014) Estimating causal effects with a non-paranormal method for the design of efficient intervention experiments. BMC Bioinformatics 15: 228. Link: https://goo.gl/echLTP

14. Richardson TS, Spirtes P (2002) Ancestral graph Markov models. Ann Statist 30: 962-1030. Link: https://goo.gl/nMEzFN

15. Richardson TS (1996) A discovery algorithm for directed cyclic graphs. In Proc UAI. Link: https://goo.gl/9qp5FD

16. Malinsky D, Spirtes P (2016) Estimating Causal Effects with Ancestral Graph Markov Models. JMLR workshop and conference proceedings 52: 299-309. Link: https://goo.gl/iqDGAM

17. Brodersen KH, Gallusser F, Koehler J, Remy N, Scott SL (2015) Inferring causal impact using bayesian structural time-series models. Ann Appl Statist 9: 247-274. Link: https://goo.gl/1NZq3K

18. Hauser A, Buhlmann P (2012) Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. J Mach Learn Res 13: 2409-2464. Link: https://goo.gl/ZpVWwa

19. Tillman RE, Danks D, Glymour C (2008) Integrating locally learned causal structures with overlapping variables. Adv Neural Inf Process Syst 21: 1665-1672. Link: https://goo.gl/XWxVYH

20. Tsamardinos I, Triantafillou S, Lagani V (2012) towards integrative causal analysis of heterogeneous data sets and studies. J Mach Learn Res 13: 1097-1157. Link: https://goo.gl/AU7wBE

21. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc Series B 58: 267-288. Link: https://goo.gl/nA3k2r

22. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, et al. (2000) Functional discovery via a compendium of expression profiles. Cell 102: 109-126. Link: https://goo.gl/QR1VkK

23. Zou H, Hastie T (2005) Regularization and Variable Selection via the Elastic Net. J Roy Stat Soc Series B 67: 301–320. Link: https://goo.gl/Dn12hx

24. Lempe J, Balasubramanian S, Sureshkumar S, Singh A, Schmid M, et al. (2005) Diversity of flowering responses in wild Arabidopsis thaliana strains. PLoS Genet 1: 109-118. Link: https://goo.gl/astZ21

25. Le TD, Liu L, Tsykin A, Goodall GJ, Liu B, et al. (2013) Inferring microRNA-mRNA causal regulatory relationships from expression data. Bioinformatics 29: 765-771. Link: https://goo.gl/rcUkK3

26. Søkilde R, Kaczkowski B, Podolska A, Cirera S, Gorodkin J, et al. (2011) Global microRNA analysis of the NCI-60 cancer cell panel. Mol Cancer Ther 10: 375-384. Link: https://goo.gl/SbTRNv

27. Marchesi JR, Dutilh BE, Hall N, Peters WH, Roelofs R, et al. (2011) Towards the human colorectal cancer microbiome. PLoS One 6: e20447. Link: https://goo.gl/kfsDYv

28. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, et al. (2014) Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol 10: 766. Link: https://goo.gl/hJNSQs

Citation: Rebai A (2017) Causality in Genomics Studies: Time is ripe for a new Paradigm. Open J Bioinform Biostat 1(1): 010-014.