My thesis

Lluís Revilla Sancho

2018-04-30

Contents

Preface

This is the preface of my thesis

Chapter 1

Introduction

Integration is usually defined as:

> the process of combining two or more things into one — https://dictionary.cambridge.org/dictionary/english/integration

We can classify and explore integration in several ways, here I outline some classifications in the bioscience field, with relationships to concrete methodology and usages.

1.1 Data type: numeric vs categorical

Sometimes doctors want to understand the relationship between some phenotype they observe as a defined state from the underlying mechanistic point of view. Usually this involves looking how the metabolites, the gene expression, the methylation, the number of variants a gene has, and other quantified variables are related to the observed phenotype. Which means relating numeric and categorical data types.

1.2 Goal

The purpose of the integration can be divided in two: - Knowing a relation explore a particular case. - Explore two sets of variables to find how do they relate

1.2.1 Explore known relationships

When checking a particular case or condition for a given relationship it is easier because one already knows what kind of relationship do the variables have. If we already know that X is related to Y, we need to check if this holds or not.

1.2.2 New relationships

However on the second case is difficult because there is no prior information about the type of relationship between the variables, which makes it harder to find it. When there isn't any information about how two sets of variables relate, for instance how the methylation affects the microorganism of the gastrointestinal tract, it is difficult to select a tool to analyze it.

1.3 Relationship between variables and samples

Traditionally for each sample few variable has been measured, for instance for a tree only the height and width are measured, however with the new omics techniques (transcriptomics, metabolomics, methylomics, genomics), thousands of variables are measured for the same sample

1.3.1 More variables than samples

For a single sample of RNA around 50k genome identifers (genes, long non coding RNAs, iRNA, pseudogenes,. . .) can be measured. Which lead to the case where there are (many) more variables than samples, generally "old" statistics don't consider this case, as it has its own complications like covariance between the variables. When two variables are tighly correlated, discerning which is the lead and which is following is near to impossible.

1.3.2 More samples than variables

This is the typical example when from a cohort of patients the temperature is measured along the stage of a disease, two variables for each sample.

1.4 Relationship between samples

Each sample has the correspondent sample on the other data-set or not, so if for the same patient we can measure the transcriptomics, the metabolomics and the methylation at the same time it is XXXX. Sometimes because the piece of the patient obtained through surgery is not big enough to be divided into the different experiments needed. In some cases we have transcriptomics data from some patients and microbiom data of other patients. This results in a new source of variation that has to be dealt with. Which complicates the conclusion one can draw from the studies of these kind of data.

1.5 Time

Time is one of the factors that we can't usually control, despit having programmed visits every two weeks some patients might come early due to calendar reasons (holidays), family reasons, or disease state. At the same time is very important to consider it because two variables can seem correlated if we don't take time into consideration

1.6 Relationship between the variables

Since the lactose operon we know how some genes regulated between them, but for other variables we don't know how they are related. For instance how does the increase in expression of a gene affects the growth of a microorganism?

It is important to note that some integrations may lead to interactions while others don't, that is, the relationship of two different type of variables might be due to a physical relationship or through a more obscure relationship with no known direct contact.

1.7 Interpretation

If one does a correlation between two variables, the interpretation of the analysis is clear, if one variable increase, the other too. However as more complicated methodologies are developed the interpretation is not so clear, for instance how can one interpret the result of a canonical correlation analysis?

1.7.1 Individually

How each variable relates to another, like in the correlation analysis, the relationship between two variables under study.

Or by patient: how do interpret that in these patient variable A and B is X and Y?

1.7.2 Globally

In a PCA for instance how do we interpret that some variables have the same loading? What happens in a more difficult method like canonical correlation analysis.

What does all these variables say about each patient?

1.8 Conclusion

The field of integration is large and complex, with high interest in the recent days, specially in the pshycology and omics field, which lacks of a large study.