Data integration on inflammatory bowel disease

Lluís Revilla Sancho

2021-03-30

Contents

Preface

The main topic of the thesis is data integration applied in the inflammatory bowel disease (IBD) research. The data I will be integrating are different omics data and the phenotype of the patients. This disease is complex and there are hypothesis pointing that the microbiome is a major factor in the disease. In the precision medicine framework data integration is important to consider all the relevant variables that influence a disease.

The thesis is performed on the IDIBAPS research institute. My colleges are biologist, microbiologists, veterinaries… and we have weekly meetings with the doctors visiting at the nearby hospital.

The thesis program allows one to defend the thesis after 3 years, and up to 5 in total. My timeline is to finish in 2022 after 4 years (I'll add it when it is finished to see how it went) with the guide of my thesis directors Juanjo Lozano and Azucena Sala, who help as bioinformatician and disease expert respectively.

Chapter 1

Introduction

The inflammatory bowel disease (IBD) involves Crohn's disease(CD) and ulcerative colitis(UC). It generally affects the terminal ileum and the colon but it can affect any segment of the gastrointestinal tract. UC is a recurrent, chronic and continuous inflammation of the colon and rectum while the CD is not a continuous inflammation and affects the whole gastrointestinal tract.

IBD etiology is unknown. However, once it has initiated the most prevalent hypothesis of its chronicity suggests an aberrant immunological response to antigens of the commensal microbiome.

Treatments provided for the IBD are palliative. Those treatments include, noninflammatory drugs, suppressors and biologic. The therapeutic options can induce remission in some patients, however they often need continuous treatment to avoid recurrence. Nevertheless, many patients are refractory or intolerant to those therapies and need to undergo surgery.

The disease present unique characteristics that require the usage of integration methods in order to find the specific relationship of the microbiome and the intestine on the disease.

1.1 Integration

Integration is defined as:

> "the process of combining two or more things into one" — Cambridge Dictionary

Other words that are used integrati(-on, -ve), multi-omics, pluri-omics. However, here integration will be used as it is the more general one and not restricted to omics. Since the beginning of the integration methods there have been many methods proposed. Some of these methods are specific for one application or data while others are more general. On the recent years with the increase of more datasets with more variables than previously has seen an increase on the number of methodologies available on several disciplines but mainly on the biological science. This in turn has increased the importance of classification, review and comparison of the tools available, as well as, benchmarking these tools against the same dataset (C. Wu et al. 2019). Other times only the strategies used are used to classify the methods (Cavill et al. 2016; Chong and Xia 2017).

1.1.1 Classification of integration method's

We can characterize and classify integration methods in several ways. Here I outline some classifications in the bioscience field, with relationships (and references) to concrete methodology and usages.

1.1.1.1 Data type: numeric vs categorical

The most important distinction in integration methods is what kind of data are combined. In general data can be divided between categorical and numeric variables, which are usually found in several fields. Sometimes doctors want to understand the relationship between a phenotype they observe and the underlying mechanism. Usually this involves looking how the metabolites, the gene expression, the methylation, the number of variants a gene has, and other numeric variables are related to the observed (categorical) phenotype (like pain).

Depending on what is a method aimed for it handles both data types or just one, often they are used differently. The most common way to handle different type of data is converting the categorical values to a mock or dummy variable. For each categorical factor there is a new numerical value with 1 if that sample had this value or 0 otherwise. Sometimes Often the number of variables created is one less than the number of factors that existed. For instance, if the categorical variable has three values (A, B, C) it would be converted to A (1, 0, 0) B (0, 1, 0) and C (0, 0, 1) and usually one of these three variables is omitted. This transformation allows to use the categorical values in methods though for numeric variables.

1.1.1.2 Aim

The results and methods of data integration depend on the objective of the method and on the (biological) question. Most of the times one (or all) of the following results are expected:

- An overview of the role of each 'omics in a biological system

Sometimes several omics are used and the question is which omic method is the best one to describe the disease. Or several methods are used to understand better how two omics interact.

- A better understanding of the relationships between the 'omics types

When the relationship between omics is known there are methods to find those relationships and improve upon previous knowledge. For instance, checking that in a particular case or condition there is a given relationship, and that this relationships follows our model or not.

- A molecular signature[1] leading to more insight into molecular mechanisms

The usage of multifactor analysis might lead to the need to redefine the features that are essential for identifying a phase or a cell line.

- A predictive model

If a good model is known it can be used as a prediction if enough information is gathered, which might improve the treatment.

1.1.1.3 Relationship between variables and samples

Depending on the amount of variables and in which samples they have been measures we can classify in two types of integration. 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.

- More variables than samples:

For a single sample of RNA around 50k genome identifiers (genes, long non coding RNAs, iRNA, pseudogenes,…) can be measured. High-throughput data analysis typically falls into the category of $p \gg n$ problems, where the number of genes or proteins, p, is considerably larger than the number of samples, n. Which leads 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 co-variance between the variables. When two variables are tightly correlated, discerning which is the lead and which is following is near to impossible.

- More samples than variables:

An example would be when from a cohort of patients the temperature is measured along the stage of a disease: two variables for each sample. If in the cohort there are more than 2 patients, then the number of samples is greater than the number of variables studied. This is described in the literature as $n \gg p$.

[1] A signature is usually a group of features that describe/are representative of a cell line, a process or a stage.

1.1.1.4 Relationship between samples

Depending on the relationship between the samples, the questions answerable and the methods available differ. If each sample has all the expected data we wanted to measure it is a complete case.

Sometimes because the sample is not enough, or there are some technical or organizational problems we might lose a source of data for a sample (which is known as an incomplete case). 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.

Even when all the cases of a sample are complete the samples can come from several sites of the same individual or with different combinations of variables, which makes is relevant to understand.

1.1.1.4.1 Time

Time is one of the factors that sometimes cannot be controlled, despite having programmed visits every two weeks some patients might come early or later due to calendar reasons (holidays), family reasons, or disease state. Sometimes is precisely the object of the study, to see the relationships at different time, or see how the relationships change with time. Simultaneously is very important to consider it because two variables can seem correlated if we don't take time into consideration. Also, to discover causality between two variables the cause must be before the consequence, which highlights the importance of time. Being aware of the time differences and time scales is crucial in most cases.

1.1.1.5 Relationship between variables

Since the lactose operon we know how some genes regulate each other. 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? Usually the relationships between variables are mediated by many factors or interactions.

Network approaches relate the variables between them (such as (Koh et al. 2019)). And are fairly new and despite being used in rare occasions they are growing in popularity.

In partial correlations some or all of the other variables and considered on how much do they affect and deduced.

1.1.1.6 Input data

Methods can be classified by the kind of input data required. Some of them need data from the same patients on each data set used to integrate while other do not.

- Data from the same samples:

These methods do not handle well or at all missing data. They need complete cases/data of the samples in order to be able to integrate the results. These methods include Regularized Generalized Canonical Correlation Analysis (RGCCA) (A. Tenenhaus and Tenenhaus 2011; A. Tenenhaus et al. 2014), Multiple co-inertia analysis (MCIA) (A. C. Culhane, Perrière, and Higgins 2003), Multi-Study Factor Analysis (MSFA) (Vito et al. 2019), Multi-Omics Factor Analysis (MOFA) (Argelaguet et al. 2018), STATegRa (Gomez-Cabrero et al. 2019).

- Data from different samples:

These methods do not need data from the same sample. They draw their conclusions generalizing from the the data available. Some of them handle missing data, while others do use the data at face value. These method includes MetaPhlAn2, HUMAnN, LEfSe (E. A. Franzosa et al. 2018; Truong et al. 2015; Segata et al. 2011).

- Data types:

HCG, 16S, RNA-seq, metabolomics do not share the same data distribution, and are different between them. Also even with the same data depending on the processing of the data they can have very different properties: OTUs (operational taxonomic unit) do not behave equally as ASV (amplicon sequence variants)

1.1.1.7 Mathematical framework

Depending on the input and the objective methods use different mathematical framework to process the data. The most important ones are listed below:

- Network

Multilayer, including the multiplex, Molti-C-DREAM, RWR-MH, RWR-M. Network embedding MultiVERSE

- Dimensional Reduction Momix, RGCCA,
- Active module identification Multiomic objective genetic algorithm (scores based in two metrics, node score and density of interactions score). MOGA-MUN

1.1.1.8 Output results

According to the output the integration methods can be classified in several groups: For those from dimensional reduction methods there are three: Shared factor across the data, specific factors for each data or mixed factors.

- Shared factors:

The integration results in a vector of the samples in a lower dimensional space that is shared by all the data used to integrate. Such methods include iCluster, Multi-Omics Factor Analysis (MOFA) (Argelaguet et al. 2018).

- Specific factors:

The integration results in several vectors of the samples in a lower dimensional space of each data used to integrate. Such methods include Regularized Generalized Canonical Correlation Analysis (RGCCA) (A. Tenenhaus and Tenenhaus 2011; A. Tenenhaus et al. 2014), Multiple co-inertia analysis (MCIA) (A. C. Culhane, Perrière, and Higgins 2003), Multi-Study Factor Analysis (MSFA) (Vito et al. 2019).

- Mixed factors:

The integration results in both previous factors, specific of each data and common to all the data. Such methods include Joint and Individual Variation Explained (JIVE), integrative Non-negative Matrix Factorization (intNMF).

1.1.1.9 Interpretation

Understanding how to interpret the results of the methods is highly tight to understanding the method. If one does a correlation between two variables, the interpretation of the analysis is clear, if one variable increase, the other one too. However as more complicated methodologies are developed the interpretation becomes less clear, for instance how can one interpret the result of a canonical correlation analysis?

- 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?

- 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?

There have been some articles about how to interpret those methods on real datasets (Sherry and Henson 1981). Others, to benchmark and to learn how to interpret propose analyzing a simulated dataset (Chung and Kang 2019; Martínez-Mira, Conesa, and Tarazona 2018).

To help interpretation frequently synthetic datasets are used to compare the results of the integration with the dataset of interest and to compare different tools. This datasets are created with some relationships that the tools are expected to find.

There exists several methods to create synthetic datasets like MOSim, metaSPARSim, CAMISIM, ballgown, polyester and even edgeR can be used.

1.1.1.10 Conclusion

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

1.1.2 Reviews and benchmarking

The comparison and review of methods independently from original authors have become a crucial step for selecting the right tool for a research (Cantini et al. 2020).

Some of these reviews are focused on a field: metabolomics (ref), genomics, microbiomics… or on a specific characteristic: ?.

1.1.3 Summary

Methods to integrate have many characteristics that allow to classify them which explains the diverse results one can have using one of them.

1.2 Inflammatory bowel disease

Inflammatory bowel disease (IBD) includes the chronic diseases Crohn's disease (CD) and ulcerative colitis (UC) which are characterized by alternating periods of remission and clinical relapse that mainly affect the gastrointestinal tract. CD is a progressive reincident disease that can affect all the gastrointestinal tract but shows mostly on both terminal ilium and colon with a discontinous inflammation. The UC is a colonic reincident disease characterized by a continuous inflammation of the colon. Both of them have different risk factors, clinical, endoscopic an histological characteristics.

On CD the disease has a granulome that can appear on any intestinal layer that consists on big multinucleous cells. while on UC we observe a different damage consisting in many neutrophil in the cript lumen.

In addition to the location(s) inflamed to distinguish between the diseases mosaic zones (patches of inflamed and non-inflamed areas) are more characteristic of CD.

Sometimes blood analysis are also used to if there are some factors altered such as leucocits, a high globular sedimentation speed, increas of protein C reactive and anemia. Also if there is an increase of calprotectine in fecal samples is indication of CD disease.

Around 4,2 million individuals suffer from IBD in Europe and North America combined. The dysregulation of the inflammatory response observed in IBD requires interplay between host genetic factors and the intestinal microbiome. Several studies support the concept that IBD arise from an exacerbate immune response against commensal gut microorganisms. Nonetheless, the disease could result from an imbalanced microbial composition leading to generalized or localized dysbiosis[2].

The role of the gut microbiome in IBD is an ongoing field of research. Several authors are currently studying the alterations reported in IBD of the intestinal microbiome. However, it is still unclear the cause-effect relation between dysbiosis and IBD. Partly due to the multiple variables that might contribute to the disease progression; for instance, age, diet, usage of antibiotic, tobacco, environment, and eventually socioeconomic status. This could be due to both the genetic predisposition and environmental factors; for instance, bacterial or viral infection, diet, usage of antibiotic, and eventually the socioeconomic status.

(see (Human Microbiome Project Consortium et al. 2012))

The relationship between host and microbiome has been proposed to play a fundamental role to maintain the disease. Little is known of the influence of the gastrointestinal microbiome in the expression of the gastrointestinal tract.

1.2.1 Disease etiology

As noted on the introduction, the origin of the disease is unknown. There are several articles that point out to a relationship between the immune system and the microbiome.

As seen the Crohn's disease and the ulcerative colitis are considered two different disease due to their differences despite their similarities.

Although so far we do not know what starts the disease there are differences on the disease when appears at different age. There is a rough classification between very early, early or adult on-set disease. The main differences are [ask Isa?] .

1.2.2 Crohn's disease

As previously introduced, Crohn's disease is a chronic inflammatory disorder characterized by alternating periods of remission and clinical relapse it is

[2] A signature is usually a group of features that describe/are representative of a cell line or a process or a stage.

frequently associated with extraintestinal manifestation and/or concomitant immuno-mediated diseases. Inflammation on the gastrointestinal tract is transmural and can affect from the mouth to the anus, but mainly it manifests on the ileum and colon.

1.2.2.1 Clinical presentation

The disease itself manifest an heterogeneous symptoms that can involve, diarrhea, weight loss, abdominal pain, fever, anorexia, malaise. Other co-occurring manifestations are arthritis, primary sclerosing cholangitis, skin disorders venous or arterial thromboembolism and/or pulmonary involvement. All these symptoms make it hard to provide a diagnosis by non-specialists and can lead to delays on correct diagnosis of the disease as there isn't a single gold standard test for its diagnosis.

The detection of parasites or bacteria, such as Clostridium difficile, would admit an infectious disease. The detection of fecal calprotectin, is a good marker of endocopic activity with sensitivity above 70% and specificity above 80% generally. Usually the diagnosis is performed with a colonoscopy, whether there is inflammation on the gastroinestinal tract on discontinuous regions then is Crohn's disease. This inflammation could also present as an ulceration and rectal sparing. Histological lesions also help to diagnose the patients.

To address this difficulty the Montreal classification aims to classify patients according to their age of disease onset, standardized anatomical disease location an disease behaviour. This classification assumes that the location of Crohn's disease remains stable over time after diagnosis but behavioral phenotypes change with most patients progressing from an inflammatory phenotype to a stricturing or penetrating one.

There is to some extent a disassociated between patient's (clinic) well-being report and the colonoscopy observation. Often patients report feeling better but their gut is as inflamed as previously. This has lead to several scores and thresholds used on research and treatment of the patients.

1.2.2.2 Disease course

In the early stages of the disease the relapsing and remitting course is more frequent. Often relapses are accompanied by clinical symptoms, and very few have prolonged clinical remission. When there is clinical remission, as mentioned above there can still remain some other lesions and sublinical inflammation often persists. Often the damage caused by the disease evolves to fibrostenotic stricture or penetrating lesions (fistula and abscess).

Although Crohn's disease and colon disease are similar the drugs and treatments for each is different. The order of the treatment/drugs changes depending on the disease.

1.2.3 Ulcerative colitis

Around a third of the patients with ulcerative colitis suffer proctitis, followed by left colitis, when the affected area starts on the rectum and extends to the left colon, distal colitis, defined as disease limited to rectum and sigmoid colon, extensive colitis as involvement of at least the descending colon, and pancolitis, when the disease affects all the colon. The extension and severity of the disease does correlate with the clinical observation, where longer extension is a worse prognosis (Etchevers et al. 2009). Extensive colitis is also associated with extensive care (Etchevers et al. 2009).

1.2.4 Clinical care

The goal of treatment is not limited to clinical remission[3] but rather to mucosal healing, which means the reconstitution of the structure and function of the intestinal epithelial barrier. Mucosal healing is a first step towards the healing of deeper layers of the inflamed bowel wall on the Crohn's disease.

Clinical remission is not enough to avoid relapse and the bowel may not be free of lesions, evolving to other phenotypes such as fibrostenotic stricture or penetrating lesions increasing the structural bowel damage.

Colectomy[4] is a surgery procedure done when there is colorectal cancer or the damage on the colon has been too big. That sometimes IBD patients have to endure.

Patients that undergo a colectomy need to have their bowel reconnected with a procedure called ileoanal anastomosis (J-pouch) surgery. This often ends up in a pouchitis[5]. Which might extend to healthy areas.

Dysplasia[6] is often associated with crhonic IBD patients, which is considered a precendence before colorectal cancer develops (Mark-Christensen, Laurberg, and Haboubi 2018).

Fistula[7] is often found on the phenetrating phenotype of IBD patients.

Herpes and citomegalovirus virus.

Imaging techniques

[3] clinical remission is the when the symptoms of IBD have lessened to the point that they're mostly absent, gone, or barely noticeable.

[4] A surgical procedure to remove part or all of your colon.

[5] inflammation that occurs in the lining of a pouch created during surgery to treat ulcerative colitis or certain other diseases.

[6] Abnormal development of cells within tissues or organs.

[7] An abnormal connection between two body parts.

Many scores have been proposed for several purposes, from quality of life to disease severity or patient status. Among the scores most used are the following:

- Mayo: A score designed to be simple to calculate at the bedside based on stool frequency, bleeding, mucosal apperance at endocopy and disease activity. (Schroeder, Tremaine, and Ilstrup 1987)
- SES-CD: simple endoscopic score for Crohn's disease (Daperno et al. 2004). Score based on size of ulcers, ulcerate surface percentage, affected surface and presence of narrowings on the bowel.
- IBDQ: A 32 questionnaire used to assess the quality of life grouped into four categories: bowel, systemic, social and emotional. (Irvine 1999)

Other measured parameters:

- Weight
- Calprotectin
- PCR (Proteinc C reactive)
- Hemoglobine

1.2.4.1 Drugs and treaments

The ones given on the hospital more or less on the order of administration, although this varies between patients and time according to recurrent meetings doctors have and current recomendations:

- anti-TNFα
- Vedolizumab
- Ustekinumab
- Risankizumab
- Tofacitinib
- surgery
- TRIM/HSCT
- 5-asa
- corticoesteroids
- Azathiprine/Mercaptopurine 6MCP
- Methotrexate MTX
- Tacrolimus FK
- Cyclosporin A CyA
- Infliximab
- Adalimumab
- Antibiotics

All these drugs and procedures are available at the time of writing, and patients were taking one or more of those drugs as standard of care.

Often a reduction in dose results in losing the response to the drug. Thus, after an increase of the dose to the previous levels the patients not recover the initial response they had.

A separate procedure that is done on the Hospital is the Hematopoietic Stem Cell Transplant (HSCT). This is a new procedure given only to the most extreme cases for which several publications have checked that it resets the immunological state of the patient.

Last there are several proposal of fecal microbiota transplantation, which are not done on the Hospital Clínic yet. But are proposed to reset or change the intestinal microbiome of the patients to help them on the disease. This has been doing experimentally on mice and mouse for some time and some experiments on patients resulted have been performed already.

1.2.5 Summary

Inflammatory bowel disease is a complex disease that impacts the health of many people in different ways.

Current clinical care in some cases is enough to have a sustained clinical and endoscopic remission but most often is not enough and relapse is expected. Several factors, such as becoming refractory to drugs, intermittent discourse of the disease, make the treatment complex.

The lack of knowledge of what are the factors of the ethiology and causes of the disease make those treatments and drugs to be addressed to block further inflammation and damage, but cannot prevent it and often they do not stop it completely.

1.3 Integration on IBD

One of the hypothesis behind the maintenance of the inflammation involves the microbiome. Several studies have been carried out to discover links between microbiome and the inflammation. Some of these studies used metabolites, DNA-seq sequencing of the microbiome content, or targeted 16S sequencing.

Also the technical method used can differ between extracted from stools or from biopsied samples at colonoscopy or from surgical samples.

Some articles use correlation like (Häsler et al. 2016). There are others that use a combination of methods

Very rarely there is an experimental confirmation. This is due to how complicated it is to test an interaction and the simulate the conditions.

One of the few methods published were the interaction is measured is to expose the ex-vivo sample or cell line with supernatant of a microbiome culture.

Previous methods used include the use of RGCCA, (Tang et al. 2017)

(De Souza, Fiocchi, and Iliopoulos 2017)

Maaslin2 (Hu et al. 2021)

1.3.1 Genetics

Genome wide association studies (GWAS) is one of the most common genetic studies, together with methylation studies.

With GWAS it has been seen that *NOD2* is one highly relevant for the disease (Momozawa et al. 2018). However, there hasn't been a proposed mechanism for how would this work.

1.3.2 Microbiome

Microbiome, is often studies, both with 16S DNA studies or with whole metagenome study on the fecal or intestinal samples.

Faecalibacterium prausnitzii is the microorganism that has been more associated with IBD. But several other microorganisms have been also proposed. The mechanism for them has been associated to chemical compounds such as methane, which is later associated with mucosal healing.

1.3.3 Multiple data

Some studies work with many dataset from IBD patients.

Most of them are based on correlation analysis and microbiome.

1.3.4 Summary

Multiple methods and multiple studies have been done

Few focus on discovering the relationships on the disease (Hu et al. 2021)

The integration of data might help to improve the medicine and reveal links in difficult diseases like IBD. So far it has been applied in IBD with partial success.

Chapter 2

Multi-omic modelling of inflammatory bowel disease with regularized canonical correlation analysis

The canonical correlation is a method that using data from the same sample but from different datasets.

Characteristics

Input Data type: numeric

More variables than samples: It is appropriate when the number of variables is higher than the number of samples. Need to be a complete case. Time is not considered as a specially variable

Relationship between variables

Output:

Specific factors Interpretation:

Depending on the model and options used.

Over several years of progress (Tenenhaus 2008; Tenenhaus and Hanafi 2010; A. Tenenhaus and Tenenhaus 2011; A. Tenenhaus and Tenenhaus 2014; A. Tenenhaus et al. 2014; A. Tenenhaus, Philippe, and Frouin 2015) on the field of canonical correlations it provides a robust method and there is an implementation (**???**).

The regularized generalized canonical correlation is a method that combines several datasets, using data from the same sample. It is a good choice when the number of variables is higher than samples. There have been some improvements to generalize the method when $p \gg n$. Current practices include using a pre selected model of relations between blocks. However this model might not be accurate. To help find the fitting model for the data I created an R package. The package, which is named inteRmodel helps finding the right model and how fit it is for your data.

This method applied to an existing dataset of an autologous haematopoietic stem cell transplantation (Corraliza et al., n.d.). From this dataset there is data about the human transcriptome and the 16S DNA present at biopsies from colonoscopy. For several patients we have samples at different time point and at different locations. This allows us to see both location and time differences.

We looked to several models and searched for the model that fit better with the data. The first model just accounted the transcriptomics and the microbiome data, then in another family of models we added the information we know about those samples. In a further family of models we split the information we know about these samples into three different blocks grouping them according to how are they related between them. We fitted the best model for all the three family of models and we found that the most fitting model was from the family that had the data split on several.

Additionally I introduced some changes to RGCCA package. The modified package can be found here. This version return the same results as the version on CRAN but provides also some code optimization for a lower computation time. Some of these changes are to to be able to change the design between different dimensions. The idea behind this change is that if the first dimensions correctly fits the data the second dimension might need a different model of relationships. For instance, the first dimensions are dominated by two blocks of data while the second is dominated by another block of data.

2.1 Relationships' models

Most multi-omics and integration tools assume one block for each type of data. However, RGCCA uses linear relationships between blocks of data. But doesn't say what constitutes a block of data. Usually a block of data is just a source of data, such as an essay a survey or an experiment. We decided to split the block with data about the samples to separate independent variables from the same block. The hypothesis we made was that with independent blocks the combined linear model would fit better the data.

These news blocks can also help make more interpretable the relationships, because instead of a big metadata block we have a block for time related variables, another one for location and so on. This allows to design a model with an expected relationships between these dimensions.

However, one limitation of the generalized canonical correlation analysis an in general on the dimension reduction methods is that the design matrix must be symmetric. Which implies that we can't infer causal relationship using these methods.

2.1.1 Designing models

What defines a block: Blocks of variables are treated independently but variables within a block are used in a linear combination.

Create models with random links between blocks.

Create models considering the knowledge of the system or disease.

2.1.2 Evaluating models

To evaluate a model RGCCA provides the average variance explained (AVE), inner and outer. Inner AVE is for how well do all the canonical dimensions correlate with the design of the experiment, so it a measure of how good the model is. While outer AVE is a measure of how well do the variables of each block correlate with the canonical dimension, so it measures the agreement between the variables and the canonical dimensions.

To evaluate a design a bootstraping method can be used to know how well the design does apply to a variety of data. Another option is to use an external cohort to validate the same model, or using an external model to see if you reach the same model, which is what we did.

Next it must be evaluated compared to other methods to see if it brings something new or not. Of the multiple methods available we used MCIA (Meng et al. 2014). Which was compared by looking at the area under the curve for classifying the samples according to their location.

Besides a way to compare methods, these models do need to be evaluated by the insights they provide on the biological system they are being applied to, in our case the Crohn's disease. In this article we didn't look in depth to the biological relevance of the microorganisms an genes found.

2.2 Conclusions

The procedure of separating independent variables in their own block of data and later search the best model that fits the data provides a good strategy that should be consider for integration efforts.

The procedural method of searching a model and testing them is implemented on inteRmodel. But the most important thing is to consider which variables are independent of which and if they can be separated into a block for later usage on the modeling.

This was published as a preprint and after peer review published on Plus One (Revilla et al. 2021).

Chapter 3

Biological relevance of results

Provide more information that can be later used by researchers on the wet lab.

3.1 Comparing different dataset

3.2 Shared selected variables

3.3 Conclusions

References

::: {#refs} :::

Appendix A

Online resources

Some links that I found useful on the thesis and could be useful if you are interested on the multi-omics field.

- Awesome multi-omics: An online repository of references to multi-omics methods.
- Bookdown: The book about how to write this type of books.

Appendix B

Software

Along the years of this thesis several pieces of software have been generated as well as packages. Here they are listed for easier retrieval. They are listed on two ways, one with a brief explanation and another one ordered by what software piece is used on each analysis.

B.1 Listed

An improved/tested version of RGCCA, some modifications on the internal functions to ease the maintenance as well as adding tests and sometimes improving the documentation. Also modified so that it is possible to provide a vector of models so that the model of the first dimension is not the same as the model on the second dimension (not sure if mathematically speaking makes sense but from a biological one I think it might be interesting to have it).

Designed to be used with RGCCA I wrote the package inteRmodel to ease the bootstrapping and model selection.

A package to design batches to avoid batch effect experDesign and its website on GitHub.

Explore the effects of the hyperparameters on RGCCA on the provided dataset of gliomaData (Originally provided here) there is this repository sgcca_hyperparameters.

We used a pouchitis cohort published in this article(Morgan et al. 2015) that was used to compare how performs our method in other's dataset. The code used can be found in this repository.

Some functions used to explore the TRIM dataset ended up in the integration package.This include functions for correlation, network analysis, enrichment, normalization of metadata…

I developed a package to analyze sets and fuzzy sets BaseSet (based on what I learned from a previous iteration of the package). This package was meant to be used with the probabilities that arise from bootstrapping the models. However due to the long times of calculation that it would require it was not used.

To analyze the antiTNF cohort (also named BARCELONA) a different repository was created to analyze the data using the previously developed packages.

B.2 By publication

All code of the analysis of the publications is available (in his messed state and complicated history) and a brief description as to why they were used:

Multi-omic modelling of inflammatory bowel disease with regularized canonical correlation analysis:

- TRIM: Mangle with the sample, dataset, explore several methods…
- sgcca_hyperparameters: Explore the effects of the hyperparameters on RGCCA on the provided dataset.
- inteRmodel: Package for easy repeating the methodology developed on TRIM.
- pouchitis: Work with the pouchitis cohort used in this article.
- integration: Package with functions that I wrote or used on different parts of exploring the TRIM dataset ended up here.

Paper 2:

- Barcelona: Mangle with the sample, dataset …
- inteRmodel: To repeat the same procedure as in the first analysis. Further modifications were done to ease the pain to look for a model when millions of models are possible.
- integration: To not repeat myself for the same process.

Argelaguet, Ricard, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. 2018. "Multi-Omics Factor Analysisa Framework for Unsupervised Integration of Multi-Omics Data Sets." *Molecular Systems Biology* 14 (6): e8124. doi:10.15252/msb.20178124.

Cantini, Laura, Pooya Zakeri, Celine Hernandez, Aurelien Naldi, Denis Thieffry, Elisabeth Remy, and Anaïs Baudot. 2020. "Benchmarking Joint Multi-Omics Dimensionality Reduction Approaches for Cancer Study." *bioRxiv*, January. Cold Spring Harbor Laboratory, 2020.01.14.905760. doi:10.1101/2020.01.14.905760.

Cavill, Rachel, Danyel Jennen, Jos Kleinjans, and Jacob Jan Briedé. 2016. "Transcriptomic and Metabolomic Data Integration." *Briefings in Bioinformatics*

17 (5): 891–901. doi:10.1093/bib/bbv090.

Chong, Jasmine, and Jianguo Xia. 2017. "Computational Approaches for Integrative Analysis of the Metabolome and Microbiome." *Metabolites* 7 (4): 62. doi:10.3390/metabo7040062.

Chung, Ren-Hua, and Chen-Yu Kang. 2019. "A Multi-Omics Data Simulator for Complex Disease Studies and Its Application to Evaluate Multi-Omics Data Analysis Methods for Disease Classification." *GigaScience* 8 (5). doi:10.1093/gigascience/giz045.

Corraliza, Ana M., Elena Ricart, Alicia López-García, Maria Carme Masamunt, Marisol Veny, Miriam Esteller, Aida Mayorgas, et al. n.d. "Differences in Peripheral and Tissue Immune Cell Populations Following Haematopoietic Stem Cell Transplantation in Crohn's Disease Patients." *Journal of Crohn's and Colitis*. doi:10.1093/ecco-jcc/jjy203.

Culhane, Aedín C., Guy Perrière, and Desmond G. Higgins. 2003. "Cross-Platform Comparison and Visualisation of Gene Expression Data Using Co-Inertia Analysis." *BMC Bioinformatics* 4 (1): 59. doi:10.1186/1471-2105-4-59.

Daperno, Marco, Geert D'Haens, Gert Van Assche, Filip Baert, Philippe Bulois, Vincent Maunoury, Raffaello Sostegni, et al. 2004. "Development and Validation of a New, Simplified Endoscopic Activity Score for Crohn's Disease: The Ses-Cd." *Gastrointestinal Endoscopy* 60 (4): 505–12. doi:10.1016/S0016-5107(04)01878-4.

De Souza, Heitor S.P., Claudio Fiocchi, and Dimitrios Iliopoulos. 2017. "The IBD Interactome: An Integrated View of Aetiology, Pathogenesis and Therapy" 14 (August). doi:10.1038/nrgastro.2017.110.

Etchevers, María Josefina, Montserrat Aceituno, Orlando García-Bosch, Ingrid Ordás, Miquel Sans, Elena Ricart, and Julián Panés. 2009. "Risk Factors and Characteristics of Extent Progression in Ulcerative Colitis." *Inflammatory Bowel Diseases* 15 (9): 1320–5. doi:10.1002/ibd.20897.

Franzosa, Eric A., Lauren J. McIver, Gholamali Rahnavard, Luke R. Thompson, Melanie Schirmer, George Weingart, Karen Schwarzberg Lipson, et al. 2018. "Species-Level Functional Profiling of Metagenomes and Metatranscriptomes." *Nature Methods* 15 (11): 962. doi:10.1038/s41592-018-0176-y.

Gomez-Cabrero, David, Sonia Tarazona, Isabel Ferreirós-Vidal, Ricardo N. Ramirez, Carlos Company, Andreas Schmidt, Theo Reijmers, et al. 2019. "STATegra, a Comprehensive Multi-Omics Dataset of B-Cell Differentiation in Mouse." *Scientific Data* 6 (1): 1–15. doi:10.1038/s41597-019-0202-7.

Häsler, Robert, Raheleh Sheibani-Tezerji, Anupam Sinha, Matthias Barann, Ateequr Rehman, Daniela Esser, Konrad Aden, et al. 2016. "Uncoupling of Mucosal Gene Regulation, mRNA Splicing and Adherent Microbiota Signatures in Inflammatory Bowel Disease." *Gut*, gutjnl–2016–311651. doi:10.1136/gutjnl-

2016-311651.

Hu, Shixian, Arnau Vich Vila, Ranko Gacesa, Valerie Collij, Christine Stevens, Jack M. Fu, Isaac Wong, et al. 2021. “Whole Exome Sequencing Analyses Reveal Gene–microbiota Interactions in the Context of Ibd.” *Gut* 70 (2): 285–96. doi:10.1136/gutjnl-2019-319706.

Human Microbiome Project Consortium, Barbara A., Karen E. Nelson, Mihai Pop, Heather H. Creasy, Michelle G. Giglio, Curtis Huttenhower, Dirk Gevers, et al. 2012. “A Framework for Human Microbiome Research.” *Nature* 486 (7402): 215–21. doi:10.1038/nature11209.

Irvine, E. Jan. 1999. “Development and Subsequent Refinement of the Inflammatory Bowel Disease Questionnaire: A Quality-of-Life Instrument for Adult Patients with Inflammatory Bowel Disease.” *Journal of Pediatric Gastroenterology & Nutrition* 28 (Supplement): S23–S27. doi:10.1097/00005176-199904001-00003.

Koh, Hiromi W. L., Damian Fermin, Christine Vogel, Kwok Pui Choi, Rob M. Ewing, and Hyungwon Choi. 2019. “iOmicsPASS: Network-Based Integration of Multiomics Data for Predictive Subnetwork Discovery.” *Npj Systems Biology and Applications* 5 (1). Nature Publishing Group: 1–10. doi:10.1038/s41540-019-0099-y.

Mark-Christensen, Anders, Søren Laurberg, and Najib Haboubi. 2018. “Dysplasia in Inflammatory Bowel Disease: Historical Review, Critical Histopathological Analysis, and Clinical Implications.” *Inflammatory Bowel Diseases* 24 (9): 1895–1903. doi:10.1093/ibd/izy075.

Martínez-Mira, Carlos, Ana Conesa, and Sonia Tarazona. 2018. “MOSim: Multi-Omics Simulation in R.” Preprint. Bioinformatics. doi:10.1101/421834.

Meng, Chen, Bernhard Kuster, Aedín C Culhane, and Amin Moghaddas Gholami. 2014. “A Multivariate Approach to the Integration of Multi-Omics Datasets.” *BMC Bioinformatics* 15 (May): 162. doi:10.1186/1471-2105-15-162.

Momozawa, YukihideJulia Dmitrieva, Emilie Théâtre, Valérie Deffontaine, Souad Rahmouni, Benoît Charloteaux, et al. 2018. “IBD Risk Loci Are Enriched in Multigenic Regulatory Modules Encompassing Putative Causative Genes.” *Nature Communications* 9 (1). doi:10.1038/s41467-018-04365-8.

Morgan, Xochitl C, Boyko Kabakchiev, Levi Waldron, Andrea D Tyler, Timothy L Tickle, Raquel Milgrom, Joanne M Stempak, et al. 2015. “Associations Between Host Gene Expression, the Mucosal Microbiome, and Clinical Outcome in the Pelvic Pouch of Patients with Inflammatory Bowel Disease.” *Genome Biology* 16 (1): 67. doi:10.1186/s13059-015-0637-x.

Revilla, Lluís, Aida Mayorgas, Ana M. Corraliza, Maria C. Masamunt, Amira Metwaly, Dirk Haller, Eva Tristán, et al. 2021. “Multi-Omic Modelling of Inflammatory Bowel Disease with Regularized Canonical Correlation Analysis.”

PLOS ONE 16 (2): e0246367. doi:10.1371/journal.pone.0246367.

Schroeder, Kenneth W., William J. Tremaine, and Duane M. Ilstrup. 1987. “Coated Oral 5-Aminosalicylic Acid Therapy for Mildly to Moderately Active Ulcerative Colitis.” *New England Journal of Medicine* 317 (26): 1625–9. doi:10.1056/nejm198712243172603.

Segata, Nicola, Jacques Izard, Levi Waldron, Dirk Gevers, Larisa Miropolsky, Wendy S. Garrett, and Curtis Huttenhower. 2011. “Metagenomic Biomarker Discovery and Explanation.” *Genome Biology* 12 (6): R60. doi:10.1186/gb-2011-12-6-r60.

Sherry, Alissa, and Robin K Henson. 1981. “Conducting and Interpreting Canonical Correlation Analysis in Personality Research: A User-Friendly Primer.”

Tang, Mei San, Rowann Bowcutt, Jacqueline M. Leung, Martin J. Wolff, Uma M. Gundra, David Hudesman, Lisa B. Malter, et al. 2017. “Integrated Analysis of Biopsies from Inflammatory Bowel Disease Patients Identifies Saa1 as a Link Between Mucosal Microbes with Th17 and Th22 Cells.” *Inflammatory Bowel Diseases* 23 (9): 1544–54. doi:10.1097/MIB.0000000000001208.

Tenenhaus, Arthur, and Michel Tenenhaus. 2011. “Regularized Generalized Canonical Correlation Analysis.” *Psychometrika* 76 (2): 257–84. doi:10.1007/s11336-011-9206-8.

———. 2014. “Regularized Generalized Canonical Correlation Analysis for Multiblock or Multigroup Data Analysis.” *European Journal of Operational Research* 238 (2): 391–403. doi:10.1016/j.ejor.2014.01.008.

Tenenhaus, Arthur, Cathy Philippe, and Vincent Frouin. 2015. “Kernel Generalized Canonical Correlation Analysis.” *Computational Statistics & Data Analysis* 90 (October): 114–31. doi:10.1016/j.csda.2015.04.004.

Tenenhaus, Arthur, Cathy Philippe, Vincent Guillemot, Kim-Anh Le Cao, Jacques Grill, and Vincent Frouin. 2014. “Variable Selection for Generalized Canonical Correlation Analysis.” *Biostatistics* 15 (3): 569–83. doi:10.1093/biostatistics/kxu001.

Tenenhaus, Michel. 2008. “Component-Based Structural Equation Modelling.” *Total Quality Management & Business Excellence* 19 (7-8): 871–86. doi:10.1080/14783360802159543.

Tenenhaus, Michel, and Mohamed Hanafi. 2010. “A Bridge Between PLS Path Modeling and Multi-Block Data Analysis.” In *Handbook of Partial Least Squares*, edited by Vincenzo Esposito Vinzi, Wynne W. Chin, Jörg Henseler, and Huiwen Wang, 99–123. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-32827-8_5.

Truong, Duy Tin, Eric A Franzosa, Timothy L Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata.

2015. “MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling.” *Nature Methods* 12 (10): 902–3. doi:10.1038/nmeth.3589.

Vito, Roberta De, Ruggero Bellio, Lorenzo Trippa, and Giovanni Parmigiani. 2019. “Multi-Study Factor Analysis.” *Biometrics* 75 (1): 337–46. doi:10.1111/biom.12974.

Wu, Cen, Fei Zhou, Jie Ren, Xiaoxi Li, Yu Jiang, and Shuangge Ma. 2019. “A Selective Review of Multi-Level Omics Data Integration Using Variable Selection.” *High-Throughput* 8 (1): 4. doi:10.3390/ht8010004.