My thesis

Lluís Revilla Sancho

2020-01-21

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 3 years (I'll add it when it is finished to see how it went :) with the guide of my thesis directors are 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 and chronic of colon characterized by a continuous inflammation of the colon and rectum while the CD is not a continuous inflammation and affects the whole epithelium.

IBD etiology is unknown. The most prevalent hypothesis of its origin suggests that starts with an aberrant immunological response to antigens of the comensal microbiome. However, the factors that initiate IBD are unknown.

Treatments provided for the IBD are palliative and include, mainly, 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.

1.1 Integration

Integration is usually defined as:

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

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

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 values, which are usually found in several fields. 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 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 variable, where each number represents a category, then the methods use these values to correlate and relate the variables. 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).

1.1.2 Aim

The results and methods of data integration depend on the question they seek to answer 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 relationship (correlation) 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 analytical model towards personalised medicine

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

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

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 integrations. 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 covariance 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:

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. This is described in the literature as $n \gg p$.

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 cases 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 case for a samples. 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 be from several sites of the same individual or from several sites or with different combinations of variables

Time

Time is one of the factors that sometimes cannot be controled, despite having programmed visits every two weeks some patients might come early 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.

1.1.5 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.

1.1.6 Input data

Methods can be classified by the kind of input data required. Some of them need data from the same patients while other do not.

- Same samples data:

These methods do not handle well or at all missing data. They need complete data of the samples in order to be able to integrate the results.

- Different sample data:

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.

1.1.7 Output factors

According to the output the integration methods can be classified in three groups: 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)

- 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), Multiple co-inertia analysis (MCIA), Multi-Study Factor Analysis (MSFA)

- 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).

1.1.8 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 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?

- 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?

1.1.9 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.

1.2 Inflammatory bowel disease

Inflammatory bowel disease (IBD) includes both Crohn's disease (CD) and ulcerative colitis (UC). CD is a chronic, progressive reincident disease that affects both terminal ilium and colon, but it can develop in any site of the gastrointestinal tract. The UC is a colonic reincident and chronic disease characterized by a continuous inflammation of the colon and rectum.

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 microbiota. Several studies support the concept that IBD arise from an exacerbate immune response against commensal gut microorganisms. Nontheless, the disease could result from an imbalanced microbial composition leading to generalized or localized ysbiosis.

The role of the gut microbiota in IBD is an ongoing field of research. Several authors are currently studying the alterations reported in IBD of the intestinal microbiota. However, it is stillunclear the cause-effect relation between dysbiosis and IBD. This is 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 predisposistoin 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.

1.3 Conclusion

The integration of data might help to improve the medicine and it is interesting in difficult diseases like IBD. Few tools handle the case when lot of information is known of a specific person but each person is highly different from each other.

Chapter 2

RGCCA

The Regularized generalized canonical correlation is a method developed by Tenenhaus It is a method that combines several datasets, what is called a multi-omics method, from the same sample, that is designed for the case where there are more variables than samples.

(Tenenhaus and Hanafi, 2010; Tenenhaus, 2008; Tenenhaus et al., 2015; Tenenhaus and Tenenhaus, 2011, 2014; Tenenhaus et al., 2014)

(**?**)

Bibliography

Human Microbiome Project Consortium, B. A., Nelson, K. E., Pop, M., Creasy, H. H., Giglio, M. G., Huttenhower, C., Gevers, D., Petrosino, J. F., Abubucker, S., Badger, J. H., Chinwalla, A. T., Earl, A. M., FitzGerald, M. G., Fulton, R. S., Hallsworth-Pepin, K., Lobos, E. A., Madupu, R., Magrini, V., Martin, J. C., Mitreva, M., Muzny, D. M., Sodergren, E. J., Versalovic, J., Wollam, A. M., Worley, K. C., Wortman, J. R., Young, S. K., Zeng, Q., Aagaard, K. M., Abolude, O. O., Allen-Vercoe, E., Alm, E. J., Alvarado, L., Andersen, G. L., Anderson, S., Appelbaum, E., Arachchi, H. M., Armitage, G., Arze, C. A., Ayvaz, T., Baker, C. C., Begg, L., Belachew, T., Bhonagiri, V., Bihan, M., Blaser, M. J., Bloom, T., Bonazzi, J. V., Brooks, P., Buck, G. A., Buhay, C. J., Busam, D. A., Campbell, J. L., Canon, S. R., Cantarel, B. L., Chain, P. S., Chen, I.-M. A., Chen, L., Chhibba, S., Chu, K., Ciulla, D. M., Clemente, J. C., Clifton, S. W., Conlan, S., Crabtree, J., Cutting, M. A., Davidovics, N. J., Davis, C. C., DeSantis, T. Z., Deal, C., Delehaunty, K. D., Dewhirst, F. E., Deych, E., Ding, Y., Dooling, D. J., Dugan, S. P., Dunne, W. M., Jr., Durkin, A. S., Edgar, R. C., Erlich, R. L., Farmer, C. N., Farrell, R. M., Faust, K., Feldgarden, M., Felix, V. M., Fisher, S., Fodor, A. A., Forney, L., Foster, L., Francesco, V. D., Friedman, J., Friedrich, D. C., Fronick, C. C., Fulton, L. L., Gao, H., Garcia, N., Giannoukos, G., Giblin, C., Giovanni, M. Y., Goldberg, J. M., Goll, J., Gonzalez, A., Griggs, A., Gujja, S., Haas, B. J., Hamilton, H. A., Harris, E. L., Hepburn, T. A., Herter, B., Hoffmann, D. E., Holder, M. E., Howarth, C., Huang, K. H., Huse, S. M., Izard, J., Jansson, J. K., Jiang, H., Jordan, C., Joshi, V., Katancik, J. A., Keitel, W. A., Kelley, S. T., Kells, C., Kinder-Haake, S., King, N. B., Knight, R., Knights, D., Kong, H. H., Koren, O., Koren, S., Kota, K. C., Kovar, C. L., Kyrpides, N. C., Rosa, P. S. L., Lee, S. L., Lemon, K. P., Lennon, N., Lewis, C. M., Lewis, L., Ley, R. E., Li, K., Liolios, K., Liu, B., Liu, Y., Lo, C.-C., Lozupone, C. A., Lunsford, R. D., Madden, T., Mahurkar, A. A., Mannon, P. J., Mardis, E. R., Markowitz, V. M., Mavrommatis, K., McCorrison, J. M., McDonald, D., McEwen, J., McGuire, A. L., McInnes, P., Mehta, T., Mihindukulasuriya, K. A., Miller, J. R., Minx, P. J., Newsham, I., Nusbaum, C., O'Laughlin, M., Orvis, J., Pagani, I., Palaniappan, K., Patel, S. M., Pearson, M., Peterson, J., Podar, M., Pohl, C., Pollard, K. S., Priest, M. E., Proctor, L. M., Qin, X., Raes, J., Ravel, J., Reid, J. G., Rho, M., Rhodes, R., Riehle, K. P., Rivera,

M. C., Rodriguez-Mueller, B., Rogers, Y.-H., Ross, M. C., Russ, C., Sanka, R. K., Sankar, J. P., Sathirapongsasuti, F., Schloss, J. A., Schloss, P. D., Schmidt, T. M., Scholz, M., Schriml, L., Schubert, A. M., Segata, N., Segre, J. A., Shannon, W. D., Sharp, R. R., Sharpton, T. J., Shenoy, N., Sheth, N. U., Simone, G. A., Singh, I., Smillie, C. S., Sobel, J. D., Sommer, D. D., Spicer, P., Sutton, G. G., Sykes, S. M., Tabbaa, D. G., Thiagarajan, M., Tomlinson, C. M., Torralba, M., Treangen, T. J., Truty, R. M., Vishnivetskaya, T. A., Walker, J., Wang, L., Wang, Z., Ward, D. V., Warren, W., Watson, M. A., Wellington, C., Wetterstrand, K. A., White, J. R., Wilczek-Boney, K., Wu, Y. Q., Wylie, K. M., Wylie, T., Yandava, C., Ye, L., Ye, Y., Yooseph, S., Youmans, B. P., Zhang, L., Zhou, Y., Zhu, Y., Zoloth, L., Zucker, J. D., Birren, B. W., Gibbs, R. A., Highlander, S. K., Weinstock, G. M., Wilson, R. K., and White, O. (2012). A framework for human microbiome research. *Nature*, 486(7402):215–21.

Tenenhaus, A., Philippe, C., and Frouin, V. (2015). Kernel Generalized Canonical Correlation Analysis. *Computational Statistics & Data Analysis*, 90:114–131.

Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K.-A., Grill, J., and Frouin, V. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3):569–583.

Tenenhaus, A. and Tenenhaus, M. (2011). Regularized Generalized Canonical Correlation Analysis. *Psychometrika*, 76(2):257–284.

Tenenhaus, A. and Tenenhaus, M. (2014). Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *European Journal of Operational Research*, 238(2):391–403.

Tenenhaus, M. (2008). Component-based Structural Equation Modelling. *Total Quality Management & Business Excellence*, 19(7-8):871–886.

Tenenhaus, M. and Hanafi, M. (2010). A Bridge Between PLS Path Modeling and Multi-Block Data Analysis. In Esposito Vinzi, V., Chin, W. W., Henseler, J., and Wang, H., editors, *Handbook of Partial Least Squares*, pages 99–123. Springer Berlin Heidelberg, Berlin, Heidelberg.