Systems biology RESYS projects - 2021

vincent.cabeli@curie.fr marcel.ribeiro-dantas@curie.fr
herve.isambert@curie.fr

October 25, 2021

1 Introduction

The aim of this project is to use the various reconstruction and analysis tools of networks that you have seen during the course to study complex systems from data that is available online.

Some of the important steps in the project include:

- Study the reference articles : familiarize yourself with the scientific background and understand the question it tries to answer. If it comes with data, understand how it was produced and how it is meant to be used.
- Set a goal in line with the project's instructions and in the spirit of the reference papers. Lay out the method for achieving this goal, step by step, using the methods seen in class. You can use other methods or other datasets if they are described in your report.
- Analyze your results in light of the litterature. Test your conclusions against other methods, other studies etc.

You will write a short report (<10 pages) presenting the scientific background, the methods and your results and give a presentation (15min + 5min of questions).

2 Projets

2.1 Project 1: Differentiation of hematopoietic precursors in embryons

Hematopoiesis is the process of differentiating hematopoietic stem cells (HSCs) into every blood cell lines: red blood cells, immunity cells, megakaryocytes which will create the platelets etc... It turns out that the first blood cells specific to the embryo appear in an extra-embryonic membrane, the yolk sac. They emerge from a primitive cell lineage that also gives rise to the endothelial cells (the inner layer of blood vessel walls).

The goal of this project is to try to rebuild the network of regulations governing the expression of key transcriptional factors for the differentiation of this primitive lineage into two distinct lines: hematopoietic and endothelial cells. The proposed dataset is comprised of binarized expression data of different genes that may either have a transcriptional regulatory role, other marker genes or "housekeeper" genes.

You will have to identify the genes that seem relevant, and study the relationships between them by means of different network reconstruction methods. In the end you will propose a graphical model to explain the mechanisms underlying the differentiation of primitive cells in two distinct lines.

- Decoding the regulatory network of early blood development from single-cell gene expression measurements (Nature Biotechnology, Moignard et al)
- Learning causal networks with latent variables from multivariate information in genomic data. (PLoS computational biology 2017, Verny et al.)

The dataset will be provided.

Main steps:

- Get familiar with the hematopoiesis process, understand the original study and the experimental setup
- Classify genes into broad functional categories : when are they expressed ? Do they have functional annotation ?
- Choose and run an appropriate network inference method, label the nodes using your classification
- Interpret the results in the light of the differentiation process, explain why some interactions are incorrect / probably correct

2.2 Project 2: Network inference as a feature selection problem

A common theme in machine learning is finding good predictors for the response variable from a mixed bag of useful and useless measures. Feature selection methods vary in details but share the common goal of trying to find the minimal and most relevant set of variables to the variable of interest, which we can think of as finding the direct neighbours of a node in a graph.

GENIE3 is a method for building gene regulatory networks that explicitly treats the prediction of a regulatory network between p genes as the aggregation of p different regression problems. In each of the regression problems, the expression pattern of one of the genes (target gene) is predicted from the expression patterns of all the other genes (input genes) using random forests. The importance of an input gene in the prediction of the target gene expression pattern is taken as an indication of a putative regulatory link. Putative regulatory links are then aggregated over all genes to provide a ranking of interactions from which the whole network is reconstructed.

In this project, you will have to implement the same idea and use a combination of feature selection and feature ranking to infer networks. You can use the method of your choice and compare the results to GENIE3.

- Inferring Regulatory Networks from Expression Data Using Tree-Based Methods
- SCENIC: single-cell regulatory network inference and clustering
- GENIE3 R package
- GENIE3 source

Main steps:

- Understand the GENIE 3 approach, try to re-implement it for yourself
- Propose your modifications and give your reasons why they could improve the results (e.g. gradient boosting instead of random forests, different hyper parameters etc...)
- Test your approach against methods presented during class, either with one of the other projects' datasets or benchmark simulations (see BN repository and tetrad)

2.3 Project 3: Contact map prediction

The similarity of three-dimensional structure between homologous proteins imposes strong constraints on the variability of their sequences. This results in correlated substitution models between amino acid residues at different sequence positions of a family of proteins. It has long been suggested that these correlations can be used to infer spatial contacts in the structure tertiary protein. In recent years, several methods have been proposed to discern direct and indirect

correlations, which is one of the main determinants of the success of the approach, among them PSICOV and DCA. In this project, you will have to rebuild the internal contact network for a widely studied protein family: the Response regulator receiver domain (Pfam code PF00072). This extremely abundant family of proteins is involved in the transduction of the bacterial signal and acts as a transcription factor interacting with domains of specific DNA binding.

This family is particularly suitable for evaluating performance inference methods for the protein contact network because:

1. it contains a large number of sequenced proteins (63,624)
2. several protein structures belonging to this family have been experimentally resolved
3. it is a classic example that has already been studied in depth in the literature.

You will need to use the dataset provided to evaluate the reconstruction given by PSICOV, DCA and MIIC by comparing the network with the real contacts that are contained in the PDB (1NXS signaling protein). For network reconstruction with MIIC, pay attention to the fact that the samples (sequences) show significant autocorrelation, and the dataset must be filtered to retain "unique" sequences (try different similarity thresholds). In this project you can also use tools for visualizing contacts in structures proteins, such as Pymol and CMView.

- PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments
- Identification of direct residue contacts in protein–protein interaction by message passing.
- Direct-coupling analysis of residue coevolution captures native contacts across many protein families

The dataset will be provided

Main steps:

- Get familiar with contact map prediction approaches, some are in fact very similar to network inference
- Pre-process the dataset paying particular attention to the similarity between sequences
- Run miic and interpret the results as contact predictions
- Compare your results to at least one other contact map prediction approach, and try to explain the differences

2.4 Project 4: Epigenetics marks and targets

Epigenetics is one of the most promising fields of modern molecular biology. Temporary DNA modifications are suspected of having a significant impact on a wide range of processes, from carcinogenesis to the adaptation of an organism to the environment. The full list of effets of each of these modifications remain relatively unknown at the moment, although we can already categorize epigenetic marks into two major groups based on their effect on expression : Enhancers, that promote the transcription of the targeted gene; and silencers that limit or silence transcription the target gene. Other major players in epigenetics are the proteins that affix or remove these marks: histone deacetylases, methylases,...

This project aims to reconstruct the existing relationships between markers and marks, using a standardised dataset summarising the presence of several marks and epigenetic markers at regulatory sites of 26893 human genes. Apart from the fact that the role of markers is to place or remove these marks, their action can be favoured/disadvantaged by the presence of one or more brands.

- Learning the human chromatin network from all ENCODE ChIP-seq data

The dataset will be provided.

Main steps:

- Get familiar with the dataset, understand the distinction between marks and targets
- Choose and run an appropriate network inference method. Here you can run the approach several times : do you want a network of interactions between targets? marks? marks and targets?
- Analyse your results, paying close attention as to how the links were removed/retained
- Use online resources to confirm on infirm some of the predicted (non)interactions in your networks

2.5 Project 5: Cell-specific heterogeneity of gene networks in the immune system

Correlation of expression between genes can offer useful hints regarding their function or underlying regulatory mechanism. Today, large amounts of expression data are publicly available, allowing researchers to estimate expression correlation over thousands of samples. However, extracting information from correlation data is not straightforward, because underlying expression data are generated by different laboratories working on different cell types and under different conditions. In order for the observed correlations to be meaningful, the data have to be normalized and corrected for any source of bias that comes from the experimental design and not a biological phenomenon.

One such corrected dataset is Immuno-Navigator, a dataset comprising 38 cell types related to the immune system. You will use the methods presented during the class to study the heterogeneity of gene regulatory networks in different cell types.

- Immuno-Navigator, a batch-corrected coexpression database, reveals cell type-specific gene networks in the immune system, PNAS 2016, Vandenbon et al.
- Immuno Navigator online portal

Main steps:

- Download the dataset and understand the normalization used for creating the database. A particularity of the data is that you have the expression of probes, which have to be mapped to genes (see e.g. biomaRt package for mapping).
- Choose and run an appropriate network inference method. Once again, there is a small added difficulty in dealing with probe expression instead of genes.
- Start from one of the examples of heterogeneity shown in Vandenbon et al. and expand it with a network approach.

2.6 Project 6: Master Regulators for Metastatic behavior in Osteosarcoma

An osteosarcoma (OS) or osteogenic sarcoma (OGS) (or simply bone cancer) is a cancerous tumor in a bone. Specifically, it is an aggressive malignant neoplasm that arises from primitive transformed cells of mesenchymal origin (and thus a sarcoma) and that exhibits osteoblastic differentiation and produces malignant osteoid. Osteosarcoma is the most common histological form of primary bone cancer. It is most prevalent in teenagers and young adults. Overall survival of patients with metastatic disease is approximately twenty percent. Mechanisms behind the development of metastases in osteosarcoma are unknown. To identify gene signatures that play a

role in metastasis, a study performed genome-wide gene expression profiling on pre-chemotherapy biopsies of osteosarcoma patients who developed metastases within 5yrs and patients who did not develop metastases within 5yrs. In genetics, a master regulator is a gene at the top of a gene regulation hierarchy, particularly in regulatory pathways related to cell fate and differentiation. When analyzing the signature of a specific behavior in a disease, you can obtain the transcription factors that are master regulators for that phenomenon, that is, responsible for the behavior you see. In this case, we're talking about metastatic behavior. RTN is an R package specialized at inferring gene regulatory networks, based on ARACNe.

Tips

- You can always check the manual page for the functions and packages you use. There is valuable information in there! You can check documentation by:
  - Browsing the vignette. Do this by running the following command in R: browseVignettes('packagename')
  - ?NameOfFunction
  - ??wordRelatedToWhatYouWantToKnow

Project step by step:

- You should download the gene expression data generated by the study (GSE21257);
- You should install the RTN package and infer the regulatory network based on the downloaded gene expression data;
- You should install snow package for parallel processing in RTN;
- You should install RedeR package to better visualize the network you inferred;
- By performing a differential gene expression analysis between the metastatic and the non-metastatic biopsies, you will obtain a signature for the metastatic behavior of this cancer in these patients;
- You should run the Master Regulator Analysis to infer putative Master Regulators by using the inferred network and the obtained signature.
- Get biological insight into these putative master regulators at the Human Protein Atlas and at Gene Cards.

Papers

1. Buddingh EP, Kuijjer ML, Duim RA, Bürger H et al.Tumor-infiltrating macrophages are associated with metastasis suppression in high-grade osteosarcoma: a rationale for treatment with macrophage activating agents. Clin Cancer Res 2011 Apr 15;17(8):2110-9. PMID: 21372215
2. Castro M, de Santiago I, Campbell T, Vaughn C, Hickey T, Ross E, Tilley W, Markowetz F, Ponder B, Meyer K (2016). "Regulators of genetic risk of breast cancer identified by integrative network analysis." Nature Genetics, 48, 33. doi: 10.1038/ng.3458.
3. Fletcher M, Castro M, Wang X, de Santiago I, O'Reilly M, Chin S, Rueda O, Caldas C, Ponder B, Markowetz F, Meyer K (2013). "Master regulators of FGFR2 signalling and breast cancer risk." Nature Communications, 4, 2464. doi: 10.1038/ncomms3464.
4. Castro MA, Wang X, Fletcher MN, Meyer KB, Markowetz F (2012). https://doi.org/10.1186/gb-2012-13-4-r29 Genome Biology, 13(4), R29. doi: 10.1186/gb-2012-13-4-r29.