Systems biology RESYS projects - 2019/2020

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

October 14, 2019

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 shortly 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: Tumorigenesis

Many biomolecular networks ensure the proper functioning of vital processes of living organisms. Mutations in the genes that make up these networks can lead to deregulation which can cause complex genetic diseases such as cancer. Many studies are seeking to identify these causal genes, looking in particular at upstream "regulator" genes intervening upstream of metabolism cascades, in order to find new therapeutic targets. Finding these genes proves to be difficult as there is a significant diversity of mutations between tumours of differents patients with the same cancer, and the tissue of the same tumour may also have a large heterogeneity. Deregulations allowing tumor growth can find their source in separate signal paths, or at different levels of the same paths, rarely pointing to a clear-cut origin.

Despite this intra- and inter-patient heterogeneity, the cells at the origin of the tumour need to to successively cross certain common barriers that induce and maintain an uncontrolled proliferation, meaning there probably is a common pool of tumour-inducing mutations for all cancers. For example, it seems at first advantageous for a (future) cancerous cell to acquire mutations that reduce its probability of apoptosis and increase its division rate. In a second step, mutations favouring the non-verification of errors in the genome and/or the failure to repair them are also favourable to tumor development. Among other desirable properties that make cancerous cells more resilient and, we find increased angiogenesis, insensitivity to anti-factors of growth and the ability to invade other tissues. For breast cancer, we find that we can categorize into four subtypes of cancer: luminal A, luminal B, basal-like and HER2.

This project proposes to study datasets containing information on mutations identified in tissue samples as well as the levels of expression and the number of copies of genes, in order to reconstruct mutation pathways and/or interactions between genes involved in tumor growth in the different subtypes of breast cancer.

- Articles
    - Hallmarks of Cancer: The Next Generation (Cell 2011, Hanahan et al.)
    - The clonal and mutational evolution spectrum of primary triple-negative breast cancers (Nature 2012, Shah et al.)
    - Sequence analysis of mutations and translocations across breast cancer subtypes (Nature 2012, Banerji et al.)
    - The landscape of cancer genes and mutational processes in breast cancer (Nature 2012, Stephens et al.)
    - Comprehensive molecular portraits of human breast tumours (Nature 2012, Koboldt et al.)
- Possible datasets
    - cBioPortal
    - COSMIC: 1 dataset

Main steps:
This project is more "open" than the others : you will chose the dataset and set your own goals.

- Set your goal in light of the tumorigenesis literature and for which you could bring the start of an answer using available, public datasets
- Compose your dataset using online resources, being careful not to introduce any undesirable bias in your data collection (e.g. control cohort vs metastatic patients is not the same as controls vs all treated patients)
- Choose and run an appropriate network inference method
- Explain your results, confirm or infirm some predicted interactions (or absence of interaction) using the literature

2.2 Project 2: Selection constraints regarding genome evolution

On evolutionary timescales, the conservation of genes conferring a selective advantage is intuitive. This is in contrast to the conservation of genes that are particularly sensitive to harmful mutations, i.e. likely to induce complex genetic diseases in individuals. However, some recent studies have shown that there is a retention bias of copies of "dangerous" genes from two complete duplications of genomes that occurred about 500MA ago, which gave birth to the vertebrates. These studies show that the retention of these copies, called "ohnologs", is the consequence of their involvement in complex genetic diseases and the development of cancers.

This project proposes to study the interactions between different genetic properties of genes encoding for a protein in humans (such as involvement in cancers, participation in com- multiprotein plexes, the essential character...) in order to understand the mechanisms that have favoured the maintaining of copies of some of these genes. Particular attention will be paid to the role of the selection pressure, quantified by the Ka/Ks ratio, which compares the number of mutations nonsynonyms (modifying the protein sequence) to the number of synonymous mutations (not modifying amino acids from proteins).

- Learning causal networks with latent variables from multivariate information in genomic data. (PLoS computational biology 2017, Verny et al.)
- Human Dominant Disease Genes Are Enriched in Paralogs Originating from Whole Genome Duplication (PLoS Comp Biol 2014, Singh et al.)
- On the expansion of "dangerous" gene repertoires by whole-genome duplications in early vertebrates. (Cell Rep 2012, Singh et al.)
- Evolution and cancer : expansion of dangerous gene repertoire by whole genome duplications (Med Sci 2013, Affeldt et al.)
- Natural selection on protein-coding genes in the human genome (Nature 2005, Bustamante et al.)
- Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models (Mol Biol Evol 2000, Yang et al.)
- Ohnologs in the human genome are dosage balanced and frequently associated with disease (PNAS 2009, Makino et al.)

The dataset will be provided.

Main steps:
This project is relatively straightforward for conducting the analysis but requires careful interpretation and contextualizing with the literature.

- Understand the dataset provided (cf Verny et al. 2017)
- Preprocess the data so that it can be used for network inference
- Choose and run an appropriate network inference method
- Compare the results to the literature, draw your conclusions regarding the presence of ohnologs in modern genomes

2.3 Project 3: 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

2.4 Project 4: 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)

2.5 Project 5: 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

2.6 Project 6 : Causal graph simulations

One of the major difficulties in developping network inference methods is that there rarely is a ground truth to compare the results to. And because datasets are so different, the best method for a given problem may not be able to produce sensible results on another dataset which presents different relationships or a different underlying graph structure. A common solution to this problem is to simulate your own netwoks and datasets under different assumptions and see if the method is able to produce good results.

In this project you will have to create a script to generate causal graphs and datasets. Given a number of nodes and a number of edges (or mean degree), you will create a Directed Acyclic Graph at random and then draw node distributions that satisfy the (conditional) dependences you can derive from the graph. Finally, you can use your simulations to benchmarck the different network inference methods seen in class.

- Tetrad

- Causal Structure Learning

Main steps:

- Develop a script to simulate observations from a random causal network. It should have several options, such as the number of nodes, number of edges, type of graph (random, small-world, power-law) and the type of distributions (discrete, continuous)
- Write another script that measures the performance of a network inference approach given the ground truth
- Use your method to compare at least two of the methods seen in class. Explain which performed better and why.

2.7 Project 7: 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

2.8 Project 8: 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.

2.9 Project 9: 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.

Study 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