Table of contents

Matrix factorization applied to multi-omics datasets with transfer learning

David Hirst * [1]

[1] Aix Marseille Univ, INSERM, MMG, Marseille Medical Genetics, Marseille, France – Aix Marseille Univ, INSERM, MMG, Marseille Medical Genetics, Marseille, France – France

Matrix factorization is a popular method for disentangling the mixtures of biological signals that underlie multi-omics data. The resulting lower dimensional representations can be used to infer the extent to which latent processes differ across biological conditions. However when a multi-omics dataset is generated from only a limited number of samples, the effectiveness of matrix factorization is reduced. Therefore transfer learning approaches to matrix factorization have previously been proposed and applied to single omics data. A transfer learning approach to matrix factorization involves using information previously inferred from a large, heterogeneous learning datset to supplement the factorization of a small target dataset. In this study I simulated multi-omics datasets in order to assess the extent to which transfer learning approaches improve the application of matrix factorization to small target datasets. I focused on the Bayesian matrix factorization method MOFA, and evaluated approaches with respect to their ability to uncover groundtruth latent structure. Across varied simulation configurations, transfer learning approaches improved the quality of the factorization when compared to factorization of the small target dataset without transfer learning.

Keywords: matrix factorization, multi, omics, transfer learning

*Speaker

Optimal transport for automatic alignment of non-targeted metabolomic data

Marie Breeur * [1], George Stepaniants [2], Pekka Keski-Rahkonen [1], Philippe Rigollet [2], Vivian Viallon

[1] Centre International de Recherche contre le Cancer - International Agency for Research on Cancer – Organisation Mondiale de la Santé / World Health Organization Office – France
[2] MIT Mathematics Department – United States

Untargeted metabolomic profiling through liquid chromatography-mass spectrometry (LC-MS) allows the measurement of a wide range of metabolites in a biospecimen. Untargeted features whose intensity are measured in untargeted metabolomics studies are only defined through their mass-to-charge ratio (m/z) and retention time (RT) and are therefore not immediately identifiable. Furthermore, m/z and RT measured under different conditions are subject to variations and features common to two different studies cannot be directly identified. This limitation hampers the external validation of results and more generally the comparison of results across different studies. It also prevents the pooling or meta-analysis of untargeted metabolomics data, thus limiting the statistical power of untargeted metabolomics studies.
Here we develop an unsupervised method to automatically match features from two LC-MS untargeted datasets by combining information on their mass-to-charge ratios (m/z), retention times (RT) and signal intensities. Our approach primarily pairs features with compatible signal intensities by making use of the Gromov-Wasserstein distance (an extension of optimal transport designed to couple sets by taking advantage of their structure) between the features within each dataset. An additional constraint allows us to restrict this coupling to pairs of features sharing similar m/z. Finally, the deviation of the RTs between the two studies is estimated in order to retain only those pairs with compatible RTs in our final matching.

We performed an extensive simulation study to evaluate the empirical performance of our method and compare it to another recent approach that uses the same type of information (m/z, RT and signal intensities). Our method outperformed its competitor in terms of sensitivity and specificity under most scenarios we considered. When applied to real untargeted metabolomics data acquired in sub-studies nested within a large European cohort, for which a small subset of features had previously been matched manually by an expert biochemist, our approach again performed well. Unlike other existing methods, our approach requires the setting of only a few parameters, and is implemented using an open-source programming language to facilitate its use and possible future developments.
Such work could have multiple applications in metabolomics, from the comparison of acquisition protocols to the pooling or meta-analysis of data from different studies. This would allow a better use of the increasingly available non-targeted metabolomic data, for example in cancer epidemiology.

Keywords: Untargeted metabolomics, Data alignment, Features matching, Optimal transport, Gro-

*Speaker

mov, Wasserstein distance

Post-clustering difference testing: valid inference and practical considerations

Benjamin Hivert * [1,2,3], Denis Agniel [4], Rodolphe Thiébaut [1,2,3,5], Boris Hejblum [1,2,3]

[1] Univ. Bordeaux, Inserm Bordeaux Population Health Research Center, SISTM team, UMR 1219, Bordeaux F33076, France – Université de Bordeaux (Bordeaux, France) – France
[2] Inria SISTM team – Inria Bordeaux Sud Ouest – France
[3] Vaccine Research Institute, VRI – Hôpital Henri Mondor, Créteil F-94000, France – France
[4] Rand Corporation, Santa Monica, CA 90401, USA – United States
[5] CHU Bordeaux, Service d'information médicale – CHU de Bordeaux; University of Bordeaux; Inserm U1219 – France

Clustering is an unsupervised learning approach commonly used to uncover heterogeneity by grouping observations into homogeneous and separated subgroups, also called clusters. In many applications, clustering is followed by hypothesis testing to identify the features that separate the estimated clusters allowing their interpretation. For example, in RNA-seq gene expression analysis, clustering of samples (patients or cells) using all genes is often performed testing for differentially expressed genes between those estimated clusters. However, several subgroups may actually contain only units derived from the same homogeneous population: clustering then artificially creates differences between these spurious subgroups, directly contributing false positives during the inference step. Such clustering-induced differences do not represent real biological truths, but simply arise from the use of the same data twice, inflating the Type I error of traditional hypothesis testing.
We propose two novel differential analysis methods that account for the initial clustering step and its uncertainties, and obtain valid p-values and inference. In a more general context, post-clustering differential analysis could be translated as the search for the individual variables, among all those used for clustering, that separate clusters. We first extend the work of Gao et al. (2022), porting selective inference to the univariate setting and testing for a mean shift between two estimated clusters for a variable of interest (after clustering is performed using all variables). Alternatively, we propose a multimodality test which relies on a more restrictive definition of subgroups (using unimodality and multimodality to characterize homogeneity and separation, respectively). We benchmark the performance of both approaches in extensive numerical simulations as well as in applications to real, low-dimensional, biomedical datasets.

However, in high-dimension, correlated variables and low but repeated signals make it difficult to define clusters. In this setting, the two proposed tests are suffering from the curse of dimensionality and a careful examination of their behavior showed limitations of their applicability in high-dimension. Therefore, we leverage a novel technique, data fission, that splits the information contained in each observation into two parts. Thus, data fission allows the clustering and the inferences steps to be performed on two (conditionally) independent datasets, retaining all the needed properties of traditional hypothesis testing. Data fission requires a tradeoff between the amount of information consumed in clustering and the amount kept for testing. We demonstrate how this tradeoff can affect either the quality of the clustering or the statistical power. We also provide practical advice on how to tune this tradeoff, as well as numerical

*Speaker

considerations for high-dimensional Gaussian-distributed data. We apply this new approach to bulk RNA-seq data from French COVID patients to quantify heterogeneity within this disease and identify genes driving this heterogeneity.

Keywords: Circular analysis, clustering, Dip Test, double, dipping, hypothesis testing, multimodality test, selective inference

Gene expression and regulatory networks: bridging the gap between mechanistic modeling and machine learning

Ulysse Herbach * [1,2], Elias Ventre [3,4], Thibault Espinasse [4], Gérard Benoit [5], Olivier Gandrillon [3]

[1] Inria Nancy - Grand Est – Institut National de Recherche en Informatique et en Automatique – France
[2] Institut Élie Cartan de Lorraine – Université de Lorraine, Centre National de la Recherche Scientifique : UMR7502 – France
[3] Laboratoire de biologie et modélisation de la cellule – École Normale Supérieure - Lyon, Université Claude Bernard Lyon 1, Institut National de la Santé et de la Recherche Médicale : U1210, Centre National de la Recherche Scientifique : UMR5239 – France
[4] Institut Camille Jordan [Villeurbanne] – Ecole Centrale de Lyon, Université Claude Bernard Lyon 1, Institut National des Sciences Appliquées de Lyon, Université Jean Monnet [Saint-Etienne], Centre National de la Recherche Scientifique : UMR5208 – France
[5] Institut de Génétique et Développement de Rennes – Universite de Rennes 1, Centre National de la Recherche Scientifique : UMR6290, Structure Fédérative de Recherche en Biologie et Santé de Rennes – France

Inferring graphs of interactions between genes has become a textbook case for high-dimensional statistics, while models describing gene expression at the molecular level have come into their own with the advent of single-cell data. Linking these two approaches seems crucial today, but the dialogue is far from obvious: statistical models often suffer from a lack of biological interpretability, and mechanistic models are known to be difficult to calibrate from real data.
We recently introduced two strategies that exploit time-course data, where single-cell profiling is performed after a stimulus: HARISSA, a mechanistic network model driven by transcriptional bursting (1), and CARDAMOM, a scalable inference method seen as model calibration (2). Thanks to this correspondence, it is possible to combine the two approaches so that the same model can be used simultaneously as an inference tool, to reconstruct biologically relevant networks, and as a simulation tool, to generate realistic transcriptional profiles in a non-trivial way through gene interactions (3).

More specifically, I will show how this mechanistic model, describing an arbitrary number of interacting genes, can be calibrated to return both an interpretable graph and a quantitative simulation tool. Interestingly, the calibration procedure turns out to be very similar to the learning step of a set of single-layer perceptrons, suggesting a natural generalization with more layers. The edges of the network ultimately have an explicit causal definition in a probabilistic paradigm: it turns out that the variability observed in the data can be explained by biological stochasticity alone, which then plays a functional role.

This work is in collaboration with Elias Ventre, Thibault Espinasse, Gérard Benoit and Olivier Gandrillon.

(1) U. Herbach (2021). Gene regulatory network inference from single-cell data using a self-consistent proteomic field.

*Speaker

https://arxiv.org/abs/2109.14888

(2) E. Ventre (2021). Reverse engineering of a mechanistic model of gene expression using metastability and temporal dynamics. *In Silico Biology*, 14(3-4), 89-113.
https://content.iospress.com/articles/in-silico-biology/isb210226
(3) E. Ventre, U. Herbach, T. Espinasse, G. Benoit, and O. Gandrillon (2022). One model fits all: combining inference and simulation of gene regulatory networks.
https://www.biorxiv.org/content/10.1101/2022.06.19.496754v1

Keywords: gene regulatory networks, causal inference, data simulation, transcriptional bursting, single cell transcriptomics

Improvement of variables interpretability in kernel PCA

Mitja Briscik * [1], Marie-Agnès Dillies [2], Sébastien Dejean [1]

[1] Institut de Mathématiques de Toulouse, UMR5219, Université de Toulouse, CNRS, UPS, Cedex 9, 31062 Toulouse, France – Université de Toulouse Paul Sabatier – France
[2] Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, F-75015 Paris, France – Institut Pasteur de Paris – France

The recent advancement in high-throughput biotechnologies is making large multi-omics datasets easily
available. Bioinformatics has recently entered the Big Data era, thus requiring new methods to optimize the
analysis of post-genomic data, considering the high complexity and heterogeneity involved.
Kernel methods offer a theoretical framework for the omics data's high dimensionality and heterogeneous
nature (Schölkopf et al. 1997). For instance, the kernelized version of Principal Component Analysis provides
a non-linear solution to reduce the sample space dimensions. However, the kernel data transformation leads
to the so-called pre-image problem since the original features are lost during the data embedding process. In
the literature, few attempts offer an interpretation of the kernel principal components axes that are described
only through pairwise similarity of the sample points. Among others, Reverter et al. (2014) proposed a
method to visualize the variables into the 2D PCs plot as arrows without providing a variable importance
ranking, thus requiring previous knowledge about which variables to display. On the contrary, Mariette et
al. (2018) proposed a variable importance selection method based on random permutation to identify the
most influential variables for every principal component. The latter method does not come with a variable
representation and can be rather computationally expensive.
In the present work, we introduce a new approach based on Reverter et al. (2014) idea, with the main difference that it gives a data-driven features importance ranking and contrarily to the procedure in Mariette
et al. (2018), it does not have a random nature while being considerably faster.
Reverter et al. (2014) showed how every variable could be associated with a real-valued function in the input
space. We are interested in their projection onto the linear subspace of the feature space induced by the
kernel. The derivative of the projected curve for a specific variable is computed at each sample

*Speaker

point to
obtain the directions of maximal local growth of the given variable. Our method is based on the computation
of the lengths of the gradient vectors for every variable at each sample point as they represent how steep the
direction given by the partial derivative of the induced curve is. Thus, we compute the mean of all norm
vectors associated with every observation for every variable. Next, the variables are ranked from the highest
mean of the norms to the lowest, indicating which variables are the most influential for the sample points
representation into the PC axes.
We first tested our procedure by comparing it with the results obtained using Mariette et al. (2018) method-
ology available in the mixKernel R package. Then we compared the different KPCA data representations
obtained using our most important features, randomly selected or least important ones. Both approaches
have been carried out on publicly available transcriptomics datasets, showing that the method successfully
finds the most relevant genes linked with the chosen kernel principal components. A further step will be
validating the selected variables from a biological point of view or via an appropriate supervised learning
algorithm.

Keywords: Kernel Principal Component Analysis, Unsupervised learning, Features selection methods, Omics data

From single-cell differential expression analysis to differential transcriptome analysis with kernel based testing.

Anthony Ozier-Lafontaine * [1,2], Bertrand Michel [3], Franck Picard [4]

[1] Laboratoire de Mathématiques Jean Leray – Centre National de la Recherche Scientifique : UMR6629, Nantes université - UFR des Sciences et des Techniques – France
[2] Nantes Université - École Centrale de Nantes – Nantes Université – France
[3] Ecole Centrale De Nantes (ECN) – Ecole Centrale de Nantes, École Centrale de Nantes – France
[4] Laboratoire de biologie et modélisation de la cellule – École Normale Supérieure - Lyon, Université Claude Bernard Lyon 1, Institut National de la Santé et de la Recherche Médicale : U1210, Centre National de la Recherche Scientifique : UMR5239 – France

Single-cell RNA sequencing (scRNAseq) is a high-throughput technology quantifying gene expression at the single-cell level, for thousands of cells and tens of thousands of genes. A major statistical challenge in scRNAseq data analysis is to distinguish biological information from technical noise in order to compare conditions or tissues. Differential Expression Analysis (DEA) is usually performed with univariate two-sample tests and thus does not account for the multivariate aspect of scRNAseq data that carries information about gene dependencies and underlying regulatory networks and pathways. Applying multivariate two-sample tests would allow to perform Differential Transcriptome Analysis (DTA), to assess for the global similarity of the compared datasets.
We propose a kernel based two-sample test that can be used for DEA as well as for DTA. The Maximum Mean Discrepency (MMD) test is the most famous kernel two-sample test (1), it is based on the distance between the mean embeddings of the empirical distributions in an high-dimensional feature space, obtained through a non-linear embedding called the feature map. Our package implements a normalized version of the MMD test derived from the non-linear classification method KFDA (2), then regularized by a kernel PCA-like dimension reduction (3). Besides reaching state of the art performances in DEA with competitive computational cost, the non-linear discriminant transformation obtained from the KFDA approach offers visualization tools highlighting the main differences between the two conditions in terms of cells, allowing to identify condition-specific sub-populations.

(1) Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A Kernel Method for the Two-Sample-Problem. page 8, 2007.

(2) Zaid Harchaoui, Francis Bach, and Eric Moulines. Testing for Homogeneity with Kernel Fisher Discriminant Analysis. arXiv:0804.1026 (stat), April 2008. arXiv: 0804.1026.

(3) Zaid Harchaoui, Felicien Vallet, Alexandre Lung-Yut-Fong, and Olivier Cappe. A regularized kernel-based approach to unsupervised audio segmentation. In 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 1665–1668, Taipei, Taiwan, April 2009. IEEE.

*Speaker

Keywords: Differential Expression Analysis, Single, cell RNA sequencing, Kernel methods, Statistical testing

Bayesian inference of parental allele inheritance in fetus for noninvasive prenatal diagnosis

Ghislain Durif * [1,2], Marie-Claire Vincent [3,4], Cathy Liautard Haag *

3

[1] Institut Montpelliérain Alexander Grothendieck – Centre National de la Recherche Scientifique : UMR5149, Université de Montpellier, LBMC – France
[2] Laboratoire de biologie et modélisation de la cellule – École Normale Supérieure - Lyon, Université Claude Bernard Lyon 1, Institut National de la Santé et de la Recherche Médicale : U1210, Centre National de la Recherche Scientifique : UMR5239 – France
[3] Laboratoire de Génétique Moléculaire, Institut Universitaire de Recherche Clinique, Université de Montpellier, CHU Montpellier – CHRU Montpellier, Montpellier Université – France
[4] Physiologie médecine expérimentale du Cœur et des Muscles [U 1046] – Institut National de la Santé et de la Recherche Médicale : U1046, Centre National de la Recherche Scientifique : UMR9214, Université de Montpellier : UMR9214 – France

The field of noninvasive prenatal diagnosis (NIPD) has undergone significant progress over the last decade. Direct haplotyping has been successfully applied for NIPD of a few single-gene disorders. However, technical issues remain for triplet-repeat expansion diseases (a.k.a. trinucleotide repeat disorder). Developping an NIPD approach for couples at risk of transmitting dynamic mutations is thus challenging but crucial. For instance, fetal genotyping using circulating cell-free fetal DNA (cff-DNA) from maternal blood might not be able to detect complex genetic patterns in the fetal DNA that could have been inherited from a parent at risk. In such family, a workaround would be to directly detect which haplotypes among the pair of parental homologous chromosomes have been inherited by the fetus, for an entire targeted region of the genome. In combination with haplotype phasing of the parent(s) at risk, it would allow to determine if the haplotype region carrying the pathogenic variation was transmitted to the fetus or not.

We present a Bayesian approach that is able, not only to infer the fetal genotype, but more importantly to directly infer the fetal allele origin from the parental phased haplotypes at each locus in a target chromosome region. In particular, our model aims to identify the parental haplotype of origin for the genetic material inherited by the fetus. To do so, only haplotype data from both parents and genotype data from circulating cell-free DNA (cf-DNA) in maternal plasma (i.e. a mix of maternal and fetal DNA) are used. On contrary to existing fetal genotyping models which consider all loci independently, we infer the allele origin jointly on all loci in the targeted region. Because of combinatorial issues, we cannot directly derive the joint posterior but we rather use a Markov chain Monte Carlo (MCMC) procedure (specifically a Gibbs sampler) to estimate the full posterior over the entire region, and a maximum a posteriori (MAP) estimation to determine the allele inheritance patterns over the entire region.

We performed analyses using blood samples from families with Huntington's disease or my-

*Speaker

otonic dystrophy type 1. We were able to perform the Bayesian inference of parental haplotype transmissions for five fetuses. The predicted variant status of four of these fetuses was in agreement with the invasive prenatal diagnosis findings. Conversely, no conclusive result was obtained for the NIPD of fragile X syndrome. Although improvements should be made to achieve clinically acceptable accuracy, our study shows that linked-read sequencing and parental haplotype phasing can be successfully used for NIPD of triplet-repeat expansion diseases.

Our approach is implemented as a Python package with a command line interface (CLI). The source code can be found in a dedicated repository (https://github.com/gdurif/nipd), and the corresponding work has been published (https://hal.archives-ouvertes.fr/hal-03716132).

Keywords: noninvasive prenatal diagnosis, bayesian inference, MCMC, Gibbs sampling, haplotyping, genotyping

COTAN: scRNA-seq data analysis based on gene co-expression

Silvia Galfre' * [1], Francesco Morandin [2], Marco Fantozzi , Daniel Puttini [2], Marco Pietrosanto [3], Federico Cremisi [4], Manuela Helmer-Citterich [5]

[1] University of Pisa – Italy
[2] University of Parma – Italy
[3] Univesity of Rome Tor Vergata – Italy
[4] Scuola Normale Superiore di Pisa – Italy
[5] University of Rome "Tor Vergata" – Italy

Estimating the co-expression of genes in single-cell is crucial. Due to the low efficiency of scRNA-seq methodologies, sensitive computational approaches are critical to accurately infer transcription profiles in a cell population. We introduce COTAN, a statistical and computational method, to analyze the co-expression of gene pairs at the single-cell level. Due to the extreme sparsity of the count matrix, the basic idea is to study the zero UMI counts' distribution instead of focusing on positive counts. This is done with a specific and new mathematical model that leads to a generalized contingency tables framework. With this, different scores and information can be evaluated or extracted and they can be used for gene correlation studies and gene or cell clustering. COTAN is an R package that complements the traditional single-cell RNA-seq analysis and it is available at http://bioconductor.org/packages/release/bioc/html/COTAN.html or https://github.com/seriph78/COTAN.

Keywords: scRNA sequencing, gene correlation, contingency tables

*Speaker

Detection and characterization of the DNA double strand break landscape in colorectal cancer cell lines

Alexandra Mancheno-Ferris * [1], Fabio Iannelli [1], Marta Falcinelli [1], Giulia Dell'omo [1], Matteo Cabrini [1,2], Fabrizio D'adda Di Fagagna [1,2]

[1] IFOM, Istituto FIRC di Oncologia Molecolare – Italy
[2] IGM, Institute of Molecular Genetics - CNR (National Research Council) – Italy

Colorectal cancer (CRC) is the 3rd most prevalent cancer with a 5-year survival rate of 85-90%. However, when it becomes metastatic (mCRC), this rate decreases dramatically to 12%. Although the molecular mechanisms involved are well described, treatments are still ineffective, and this malignancy has been ranked the 2nd most deadly type of cancer1,2. The genetic alterations involved in the tumorigenesis include the accumulation of mutations, mismatch repair (MMR) alterations and a dysfunctional DNA damage response (DDR), specifically the homologous recombination which is involved in the repair of DNA double strand break (DSB) 3–7. A molecular characterization of DDR in this type of tumor could be of fundamental importance to open the field to new biomarkers and therapies.
To characterize the landscape of endogenous DSBs in mCRC we have opted for an approach using the power of OMICS integration. To detect DSBs we took advantage of the Break Labeling In Situ (BLISS) technique, which allows their genome-wide identification with high resolution. We also developed a novel data analysis pipeline tailored for endogenous DNA damage detection.

To set up our pipeline on reproducible samples and conditions, we choose to work with mCRC cell lines. These cell lines are known to recapitulate the behavior of CRC primary tumors 8. We applied this approach to a total of 11 cell lines, and then focused on the best three cell lines in terms of replicate reproducibility and based on several quality checks.

Here we show that DSBs tend to accumulate in coding sequences and enhancers. Moreover, taking advantage of published ChIP-seq histone marks for mCRC patient-derived organoid9, we observe that DSBs correlate with enhancer markers (H3K4me1) and anticorrelate with promoter marker (H3K4me3) confirming the accumulation of DNA damage in the surrounding enhancer regions.

By an *in-silico* motif detection we show that DSBs are enriched for TEA domain binding TEAD transcription factor members. Finally, by a co-localization computational research, we scan for secondary transcription factor motifs located in proximity of TEA motifs discovering that other specific transcription factors are associated with the primary TEA motif. These preliminary results suggest that, in enhancer regions, TEAD transcription factors are likely to form complex and this could be affected by endogenous DSB occurring in mCRC cells. These results will be integrated with chromatin conformation and structural genomics available data to reconstruct the network of enhancer-target genes in CRC. We will then use the acquired information to study which molecular pathways are altered and to determine new potential therapeutic targets,

*Speaker

1. Siegel, R. L., *et al. CA Cancer J Clin,* 2019.

2. Mauri, G., *et al.* A. *Ann Oncol*,2020.

3. Lorans, M., *et al. Clin Colorectal Cancer,* 2018.

4. Sinicrope, F. A. & Sargent, D. J. *Clin Cancer Res*, 2012.

5. Jongen, J. M. J. *et al. Oncotarget*, 2017.

6. Vogelstein, B. *et al. New England Journal of Medicine,* 1988.

7. Vitelli, V. *et al. Annu Rev Genomics Hum Genet*, 2017.

8. Mouradov, D. *et al. Cancer Res*, 2014.
10. della Chiara, G. *et al. Nat Commun*, 2021.

Keywords: OMICS, colorectal cancer, BLISS, DNA reparation

Variable selection in sparse multivariate GLARMA models: Application to germination control by environment

Marina Gomtsyan * , Céline Lévy-Leduc [1], Sarah Ouadah [2], Laure Sansonnet [3], Christophe Bailly [4], Loïc Rajjou [5]

[1] AgroParisTech (AgroParisTech) – Institut national de la recherche agronomique (INRA) : UMR518, AgroParisTech – AgroParisTech 75231 Cedex 05, Paris - France, France
[2] Mathématiques et Informatique Appliquées – Institut National de la Recherche Agronomique : UMR0518, AgroParisTech – France
[3] UMR MIA-Paris – AgroParisTech, INRA - Université Paris-Saclay – France
[4] Sorbonne Université (SU) – Sorbonne Université – 4 place Jussieu 75005 Paris, France
[5] Institut Jean-Pierre Bourgin – Institut National de la Recherche Agronomique : UMR1318, AgroParisTech – France

We propose a novel and efficient iterative two-stage variable selection approach for multivariate sparse GLARMA models, which can be used for modelling multivariate discrete-valued time series. Our approach consists in iteratively combining two steps: the estimation of the autoregressive moving average (ARMA) coefficients of multivariate GLARMA models and the variable selection in the coefficients of the Generalized Linear Model (GLM) part of the model performed by regularized methods. We explain how to implement our approach efficiently. Then we assess the performance of our methodology using synthetic data and compare it with alternative methods. Finally, we illustrate it on RNA-Seq data resulting from polyribosome profiling to determine translational status for all mRNAs in germinating seeds. Our approach, which is implemented in the MultiGlarmaVarSel R package and available on the CRAN, is very attractive since it benefits from a low computational load and is able to outperform the other methods for recovering the null and non-null coefficients.

Keywords: multivariate GLARMA, sparsity, variable selection, seed quality, gene expression

*Speaker

Assessing the potential of imputed Low coverage sequencing for association studies

Raphaël Blanchet * [1], Fabien Laporte [1], Valentin Crusson [1], Audrey Donnart [1], Romain Bourcier [1], Christian Dina [1], Richard Redon [1]

[1] Nantes Université, CHU Nantes, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France – Université de Nantes, CHU Nantes, CNRS, INSERM, l'institut du thorax – France

Next generation sequencing has revolutionized the field of biomedical genetics over the past 2 decades. The plummeting of costs and the reduction of sequencing time have allowed the construction of massive databases regrouping up to hundreds of thousands of genomes. However, the cost of whole-genome sequencing (WGS) remains a limitation when working on large cohorts. Single nucleotide polymorphism (SNP) array genotyping followed by variant imputation based on reference panels has remained an alternative, in particular for common variants. Yet, by fixing the interrogated markers, this technique introduces important biases, especially when working on non European populations. Low pass sequencing paired with SNP imputation therefore stands as a promising alternative. This technique, which consists in sequencing whole genomes at low coverage (down to less than 1X), has been shown to be more accurate than SNP arrays (1), but its potential compared to costly 30X WGS for association study is still to be investigated.
To assess the relevance of this method, we have used available WGS data from 2,600 subjects including 350 cases with Brugada syndrome (BRS). All genomes were downsampled in order to simulate low-pass runs with sequencing depths of 1X, 2X and 4X. Imputation of the simulated dataset was then performed using the GLIMPSE software (2) and the 1000G reference panel. Imputed genotypes were then compared to those obtained from the 30X sequenced genomes. When considering variants with allele frequencies from 0.5 to 2% in European populations, we observed a correlation of 83% between the SNP genotypes imputed from low-pass sequencing and those directly called from WGS at 30X, when considering a mean coverage of 1X. This correlation rises up to 93% when downsampling to 4X. We also found that WGS data with mean coverage of 1X were sufficient to replicate a know genetic association between common variation at the SCN5A locus and the Brugada syndrome (3). On the other hand, rare variants association could not be replicated, even with 4X downsampled data.

In conclusion, low-pass sequencing stands as a cost-effective solution for association studies on common (low-frequency) variants. The imputation of rare variants (with a minor allele frequency below 1%) remains dependent on the size of the reference panel used for imputation. To overcome this difficulty, we propose to perform simultaneously low-pass sequencing and whole-exome sequencing at higher depth, based on the same library preparation.

(1) Vylyny Chat, Robert Ferguson, Leah Morales, Tomas Kirchhoff Ultra Low-Coverage Whole-Genome Sequencing as an Alternative to Genotyping Arrays in Genome-Wide Association Studies. Frontiers in Genetics 12 (2022) : 1664-8021

(2) Simone Rubinacci, Diogo Ribeiro, Robin Hofmeister, Olivier Delaneau. Efficient phasing

*Speaker

and imputation of low-coverage sequencing data using large reference panels. Nature Genetics 53.1 (2021): 120-126.
(3) Connie R Bezzina , Julien Barc, Yuka Mizusawa, Carol Ann Remme, Jean-Baptiste Gourraud, Floriane Simonet et al. Common variants at SCN5A-SCN10A and HEY2 are associated with Brugada syndrome, a rare disease with high risk of sudden cardiac death. Nat Genet. 2013 Sep;45(9):1044-9

Keywords: genomics, low coverage sequencing, imputation

c-RegMap portal: a co-regulatory influence network view of cancer heterogeneity and plasticity

Geoffrey Pawlak * [1], Mohamed Elati [2]

[1] Cancer Heterogeneity, Plasticity and Resistance to Therapies - UMR 9020 - U 1277 – Institut Pasteur de Lille, Institut National de la Santé et de la Recherche Médicale : U1277, Université de Lille : UMR9020, Centre Hospitalier Régional Universitaire [Lille] : UMR9020, Centre National de la Recherche Scientifique : UMR9020 – France

[2] CANTHER – CNRS : UMR9020, Institut National de la Santé et de la Recherche Médicale - INSERM, Université de Lille, Droit et Santé – France

background

Cancer studies performed in a variety of laboratories and by a number of large-scale projects have given an unparalleled amount of information on tumors and *in vitro* models. Consolidating these data into an easily accessible and intuitive system-level format is crucial to accelerate systems oncology model development.

methods

We combined network biology with machine learning and visualization tools to execute a cycle of systems oncology model development: inference of the co-regulatory networks (from transformed cells in vitro), interrogation of the tumors in vivo using the inferred networks, and intervention with the network (feeding back to the in vitro tumor models).

results

Here we introduce c-RegMap a powerful web-based tool to help researchers to rapidly access a co-regulatory influence network view of cancer heterogeneity and plasticity. c-RegMap allow user to: explore the similarities and differences between cancer subtypes and identify their possible core regulators; identify rare subtypes; align tumor and cell line transcriptional profiles; and define new targets related to the different states and plasticity of the tumours (undifferentiated vs differentiated, therapy sensitive vs resistant cells, etc.).

c-RegMap has a very intuitive interface, and no bioinformatics skills are required. For all the networks and plots that are generated, the user can run different annotation (classification, genomic alteration and clinical), add new transcriptome data, and the raw data to reproduce the plots can be downloaded for future analysis or publications.

conclusion

The identification of regulatory networks and the study of their plasticity should allow to identify efficient therapeutic strategies and will pave the way for precision oncology.

*Speaker

Keywords: Machine learning, system biology, coregulatory networks

A Phylogenetic Framework to Simulate Synthetic Inter-species RNA-Seq Data

Mélina Gallopin * [1], Olivier Lespinet *

, Paul Bastide *

[1] Institut de Biologie Intégrative de la Cellule – CNRS, Université Paris Sud, Université Paris Saclay – France

Inter-species RNA-Seq datasets are increasingly common, and have the potential to answer new questions about the evolution of gene expression. Single species differential expression analysis is now a well studied problem that benefits from sound statistical methods. Extensive reviews on biological or synthetic datasets have provided the community with a clear picture on the relative performances of the available methods in various settings. However, synthetic dataset simulation tools are still missing in the inter-species gene expression context. In this work, we develop and implement a new simulation framework. This tool builds on both the RNA-Seq and the Phylogenetic Comparative Methods literatures to generate realistic count datasets, while taking into account the phylogenetic relationships between the samples. We illustrate the usefulness of this new framework through a targeted simulation study, that reproduces the features of a recently published dataset, containing gene expression data in adult eye tissue across blind and sighted freshwater crayfish species. Using our simulated datasets, we perform a fair comparison of several approaches used for differential expression analysis.

*Speaker

Preprocessing Strategies for Bayesian Phylogeographic Analysis Using Large-Scale Genomic Sequence Data

Yimin Li * [1,2], Augustin Clessin [3], Nena Bollen [1,2], Samuel Hong [4], Simon Dellicour [1,2], Guy Baele [5]

[1] Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Belgium – Belgium
[2] Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles, Belgium – Belgium
[3] Master Biosciences, École Normale Supérieure de Lyon, Université Claude Bernard Lyon 1, Université de Lyon – Université de Lyon, Université Lyon 1 – France
[4] Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Belgium – Belgium
[5] Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Belgium – Belgium

The ongoing SARS-CoV-2 pandemic has been posing a huge threat to public health, economic development and social interactions since the end of 2019. Different SARS-CoV-2 variants keep emerging throughout this pandemic and are important to study in terms of their evolution, local and/or global dispersal, impact on transmissibility, severity, and immunity. First detected in December 2020, SARS-CoV-2 lineage B.1.525 contains several mutations of biological significance. The E484K mutation and ΔY144 deletion tend to drive immune escape, while the D614G mutation and ΔH69/V70 deletion can increase transmissibility and infectivity. With more than 10,000 genomes from this lineage being now available (Nov 2022), conducting a fully-integrated Bayesian phylogeographic analysis based on the complete data set would not be feasible in practice. We explore different strategies for reducing this data set to a representative set of genomic sequences that enables us to infer the origin and reconstruct the dispersal history of this lineage. Initial data exploration strategies such as provided in TempEst have yielded different results depending on the selected data, with the complete data set seemingly evolving at half the evolutionary rate as a data set that consists only of high-quality genomes. We first explore various maximum-likelihood and Bayesian inference methodologies - paying special attention to different molecular clock and tree prior specifications - on the core high-quality data set to establish a consensus for both TMRCA and mean evolutionary rate, which we compare to estimates from other SARS-CoV-2 lineages. We subsequently evaluate the temporal signal in the remaining genomes, grouping by time, sequencing lab and country, to determine which genomes bias the temporal signal in the core data set and warrant further investigation. To this end, we use several popular inference packages, such as BEAST, TreeTime and Chronumental. When the final data set has been constructed, we aim to employ several subsampling procedures to avoid sampling bias, as this might impact estimation of important phylogenetic and phylogeographic parameters. Targeted at analysing thousands of viral sequences, our work aims to provide a reproducible genomic data (pre-)processing pipeline for (SARS-CoV-2) phylogeographic inference analyses.

*Speaker

Keywords: Big Genomic Data, Data Quality Check, Bayesian Phylogenetics, Bayesian Phylogeography

Representation and quantification of Module Activity from omics data with rROMA

Matthieu Najm * [1,2,3], Matthieu Cornet [1,2,3,4], Luca Albergante [1,2,3], Andrei Zinovyev [1,2,3], Isabelle Sermet-Gaudelus [4], Véronique Stoven [1,2,3], Laurence Calzone [1,2,3], Loredana Martignetti [1,2,3]

[1] INSERM U900 – Institut National de la Santé et de la Recherche Médicale - INSERM – France
[2] Center for computational biology – MINES ParisTech, PSL Research University – France
[3] Institut curie – Institut Curie, PSL Research University – France
[4] Institut Necker Enfants Malades - U1151 – Institut National de la Santé et de la Recherche Médicale - INSERM, Assistance publique - Hôpitaux de Paris (AP-HP) – France

In many analyses of high-throughput data in systems biology, calculating the activity of a set of genes (or module) rather than focusing on the differential expression of individual genes has proven to be efficient and informative. Here, we present rROMA, a user-friendly and interactive R package for fast and accurate computation of the activity of gene sets with coordinated expression.
Quantification of activity by rROMA is based on the simplest uni-factor linear model of gene regulation that approximates the expression data of the module by its first principal component. The algorithm also calculates the statistical significance of the estimated module activity and gives the list of genes contributing the most to the module activity.

When interested in a disease or subgroups of a disease (e.g. in cancer), rROMA can be applied to identify, among a set of biological pathways derived from external databases, those that vary the most in disease transcriptomics. These pathways and specially the genes contributing the most to their activities are ideal candidates for building a boolean model of the disease.
We applied rROMA to cystic fibrosis, highlighting biological mechanisms potentially involved in the establishment and progression of the disease and the associated genes. Source code and documentation are available at https://github.com/sysbiocurie/rROMA.

*Speaker

Overcoming spillover for subcellular spatial omics

Benjamin Rombaut * [1], Ruth Seurinck [1], Yvan Saeys [1]

[1] University College Ghent – Belgium

Methods for profiling RNA and protein expression in a spatially resolved subcellular manner are rapidly evolving, enabling comprehensive characterization of cells and tissue architecture. While existing cell segmentation methods enables locating individual cells, cell annotation is still an open problem, especially for complex and dense tissues. Expression signals from one cell can overlap with neighboring cells, resulting in a spatial spillover that can affect downstream analysis. Several works have proposed methods to deal with spillover using compensation, pixel analysis, density estimation, neighborhood analysis or deep learning. These methods can suffer from signal loss, difficulty in scaling to large datasets, or requirements for additional data sources such as scRNAseq on the same tissue. In this work, we present a preliminary investigation into various spillover methods, their characteristics and impact on cell annotation for various subcellular platforms and tissue types.

Keywords: spatial transcriptomics, spatial proteomics, spatial spillover, compensation

*Speaker

Robust differential expression analysis at the sub-gene level

Jeroen Gilis * [1,2,3], Lieven Clement [1,2]

[1] Applied Mathematics, Computer science and Statistics, Ghent University, Ghent, 9000, Belgium – Belgium
[2] Bioinformatics Institute, Ghent University, Ghent, 9000, Belgium – Belgium
[3] Data Mining and Modeling for Biomedicine, VIB Flemish Institute for Biotechnology, Ghent, 9000, Belgium – Belgium

Differential expression (DE) analyses typically aim to identify genes for which the average expression differs between groups of observations. While DE analyses are performed routinely for both bulk and single-cell RNA-seq (scRNA-seq) data, modelling gene expression in single cells still poses some unresolved challenges. scRNA-seq data are much noisier than their bulk RNA-seq counterpart and are characterized by a large fraction of zero counts. In addition, many of the existing DE tools do not scale to the vast number cells that are profiled in droplet-based scRNA- seq protocols.

Most DE analyses are being performed at the level of genes. However, most multi-exon genes are subject to alternative splicing and can thus produce a variety of functionally different transcripts from a single genomic locus. As such, the biological question of interest often lies in assessing DE at the level of isoforms. With the advent of pseudo-alignment tools like Salmon or kallisto, fast and accurate quantification at the level of isoforms is now possible. This has resulted in a growing number of studies that are aiming to identify differential expression in isoforms (differential transcript expression, DTE), or to identify isoforms that display a change in their relative usage within their corresponding gene (differential transcript usage, DTU).

In this talk, I will discuss our recent advances in the development of DTU analysis tools. In particular, I will focus on how parameter estimates can be made more robust against the noise, sparsity and outliers that are present in scRNA-seq data, without sacrificing scalability. In addition, I will discuss how we can leverage equivalence class counts to unlock droplet scRNA- seq data for sub-gene level differential expression analysis.

Keywords: differential expression, robust, transcript usage, single, cell

*Speaker

SPArrOW: a workflow for subcellular resolution spatial transcriptomics assays

Lotte Pollaris * [1], Benjamin Rombaut [1], Wouter Saelens [2], Charlotte Scott [1], Martin Guilliams [1], Ruth Seurinck [1], Yvan Saeys [1]

[1] University College Ghent – Belgium
[2] Laboratory of Systems Biology and Genetics, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland – Switzerland

Spatial transcriptomics links each gene profiling measurement to a spatial location and can range from high-throughput pseudobulk techniques such as Visium to targeted (sub)cellular techniques. Methods that work on a subcellular level, require specific processing to generate gene-by-cell count matrices, that can be used to perform further downstream analysis.

We designed SPArrOW, a SPAtial Omics Workflow. This workflow provides a modular, scalable, versatile and interactive pipeline to process a variety of Spatial Transcriptomics methods, including Molecular Cartography (RESOLVE), MERSCOPE(Vizgen), Xenium (10x) and StereoSeq (BGI). SPArroW creates count matrices starting from stained images and the coordinates of the measured transcripts. The pipeline consists of five modular steps, which allow us to plug in new exciting algorithms at any time. Firstly, the image is processed to remove artefacts and enhance quality. Secondly, this image is segmented, identifying the cells. Thirdly, transcripts are allocated to the different cells. Afterwards, the counts are normalized. Lastly, we provide celltype or cell state annotation, which can be reference-based (scRNA-seq), or based on marker genes. Depending on the dataset and user preferences, different options are available at every step. By creating a napari plug-in, the pipeline is interactive and easy to use, making it possible to perform analyses without coding. SPArroW performs well on various datasets, both plant and animal tissue, different organs and different platforms. By enabling the creation of count matrices, SPArroW provides the ideal starting point for answering biological hypotheses from the data, including spatial clustering and colocalization analyses. It is an important first step to spatially analysing intercellular communication.

Keywords: pipeline, processing, spatial transcriptomics

*Speaker

From spatial transcriptomics to tissue morphogenesis

Lorette Noiret * [1], Adrien Leroy [2], Eric Van Leen [3], Elif Lale Alpar [2], Maria Balakireva [2], Cyril Kana Tepakbong [2], Stéphane Pelletier [2], Isabelle Gaugué [4], Boris Guirao [2], Stéphane Rigaud [5], Floris Bosveld [6], Yohanns Bellaiche [1]

[1] Polarity, Division and Morphogenesis – Institut Curie, PSL Research University, CNRS UMR3215, INSERM U934, UPMC Paris-Sorbonne, 26 Rue d'Ulm, 75005, Paris, France – France
[2] Polarity, Division and Morphogenesis – Institut Curie, PSL Research University, CNRS UMR3215, INSERM U934, UPMC Paris-Sorbonne, 26 Rue d'Ulm, 75005, Paris, France – France
[3] Bayer Pharmaceuticals – Germany
[4] MERIT – Institut de recherche pour le développement [IRD] : UR2R26100 – France
[5] Image Analysis Hub – Institut Pasteur de Paris – France
[6] Polarity, Division and Morphogenesis – Institut Curie, PSL Research University, CNRS UMR3215, INSERM U934, UPMC Paris-Sorbonne, 26 Rue d'Ulm, 75005, Paris, France – France

During development, the formation of functional tissues and organs entails tissue shaping or morphogenesis. Morphogenesis is modulated by genetic expression profiles and control of signaling pathways activity. However, tools that allow a systematic exploration of genetic factors controlling morphogenesis are lacking. To identify the role of gene expression in morphogenesis, we present a spatial transcriptome reconstruction method which relies on single cell RNA sequencing and atlases of known gene expression patterns to reallocate single dissociated cells onto the tissue. We apply our framework to the development of Drosophila dorsal thorax (notum) at the onset of its morphogenesis. We validate that the method allows for the identification of novel gene expression patterns (validation on 23 novels genes, with median correlation 0.75). We performed non negative matrix factorization to decompose the spatial transcriptome tissue in meta-regions and correlate them with signaling pathways. We combined this analysis with an existing description of the cell properties at the scale of the tissue (e.g. cell area, rate of cellular proliferation, tissue deformation). Last, we performed multivariate statistical analyses to explore the links between gene expression patterns and specific cell and those tissues properties. Subsequently, we apply our spatial transcriptomics method to mutant condition (Toll8 RNAi) and explore the possibility to predict the spatial differential expression patterns.

Keywords: spatial transcriptomics, morphogenesis, RNAi condition

*Speaker

Author Index