Monte Carlo methods for data mining in complex systems.

ECT* 5/12/2025

Michele Caselle – University of Torino and INFN
caselle@to.infn.it

Outline

- **Introduction:**
  complex systems as “component systems”

- **Bayesian inference and “Topic Modeling”**

- **Case studies and results**

- **Conclusions**

Component Systems

Many complex systems are composed by elementary units (“components”) which interact among them. They are characterized by “emergent properties” which are the result of these interactions.

Examples:

System	Components
Books	Words
Cells	Proteins/Genes
Lego toys	Lego bricks

These systems exist in different realizations, and each realization is characterized by a distinct pattern in the abundance of the components.

“Statistics of Shared Components in Complex Component Systems.” A. Mazzolini, et al. Physical Review X, 8:021023, (2018)

“Emergent statistical laws in single-cell transcriptomic data.” S. Lazzardi, et al. Phys Rev E. (2023) 044403.

Component Systems

Once components are identified, a specific realization of a system (e.g., a specific book, a LEGO set, a genome) can be represented by its parts list, which is the subset of the possible elementary components (e.g. words, bricks, genes), with their abundances, present in the realization.

.

Realizations

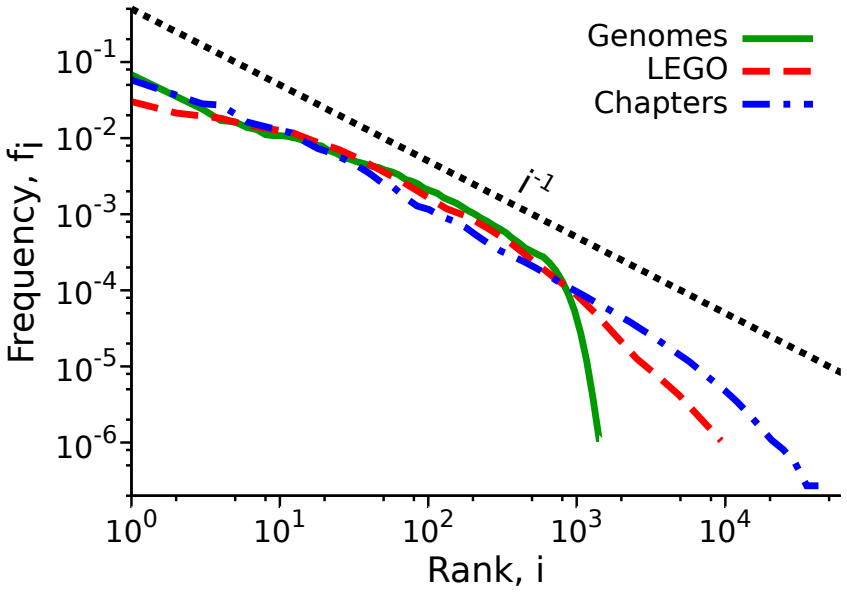
Components

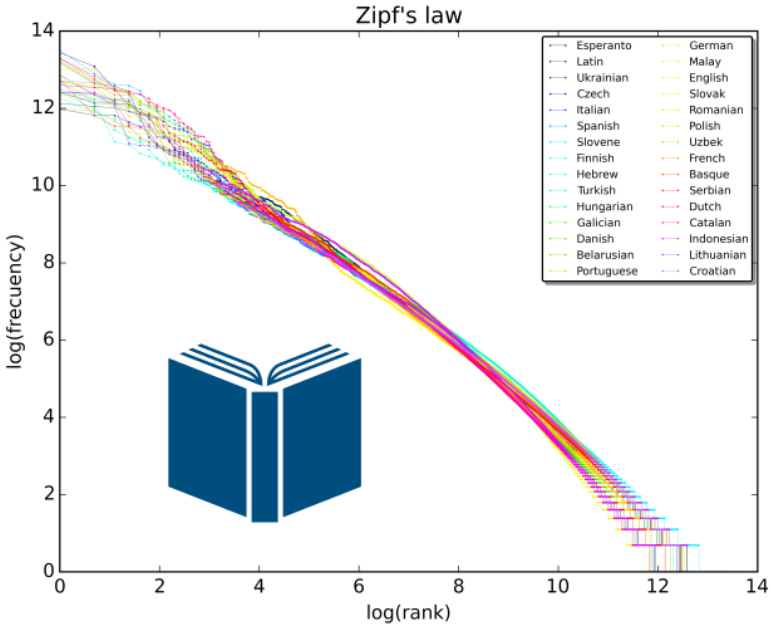
	P1	P2	P3	P4	P5
	100	10	0	36	21
	45	0	6	124	0
	163	6	0	72	0
	43	32	12	0	2
	0	0	0	0	80

Component Systems

Despite these differences many of these systems share the same universal laws which are a signature of complexity. An important issue in the study of complex systems is to understand the origin of these universal laws. Example: **Zipf law**

.





Component Systems

Occurrence patterns of components across realizations are expected to reveal relevant architectural constraints. For example, the bricks present in each LEGO set clearly reflect a target architecture that can be built with them following the instruction booklet.

Inferring such constraints from the statistics of components may answer important question about the nature of a system:

- Undertstand the general principles characterizing different classes of component systems
- Understand the basic features of the underlying generative processes.
- Identify the modular structures within the system and their functional role

The cell as a “Component System”

samples

genes

	P1	P2	P3	P4	P5
	100	10	0	36	21
	45	0	6	124	0
	163	6	0	72	0
	43	32	12	0	2
	0	0	0	0	80

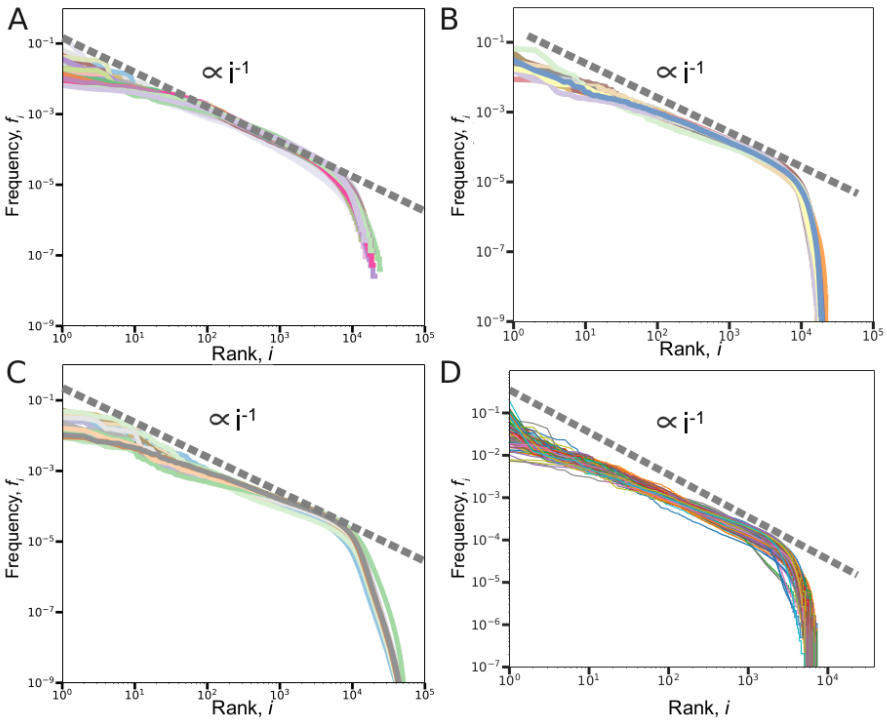
Cells are complex systems whose regulation involves a complex interplay of several genes.

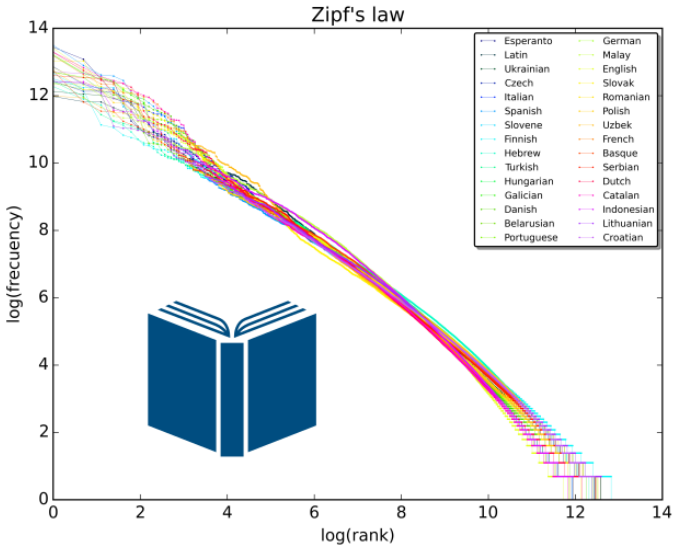
Samples may be single cells, tissues or patients

The entries of the matrix are the result of gene expression experiments and count the copies of mRNAs for each gene in the sample

“Emergent statistical laws in single-cell transcriptomic data.” S. Lazzardi, et al. Phys Rev E. (2023) 044403.

Zipf’s law for single cell data





A: Mouse Cell Atlas: each curve is an average over single cells of an organ
B: Tabula Muris: each curve is an average over single cell of an organ
C: Bulk data taken from GTEX, each line is an organ
D: 100 cells taken from heart in Tabula Muris: each line is a single cell, with no average

“Emergent statistical laws in single-cell transcriptomic data.” S. Lazzardi, et al. Phys Rev E. (2023) 044403.

Inferring the hidden architecture of cells: a physicist point of view

- Cells are complex systems whose regulation involves a complex interplay of several genes.

- Different cell populations can be characterized by "marker genes"

- Gene expression is intrinsically modular: different moduli are the metabolic pathways of the cell.

- Instead of lists of mutually exclusive elements of these pathways it is better to look for probability distributions to belong to a given pathway

- Instead of looking for few selected markers it is better to look for a distributed "marker ability" of genes which cooperate to characterize the given cell type

Topic modeling

This is a difficult task which requires innovative tools. Our proposal is to use a **“Topic Modeling”** approach.

Topic modeling is a tool which is widely used in Natural Language processing. Its use in molecular biology is (motivated by the similarity between the two complex problems: automatic classification of texts and automatic detection of cell lines

Genes	---------->	Words
Samples	---------->	Texts
Expression level of a given gene	---------->	Frequency of the word in the text

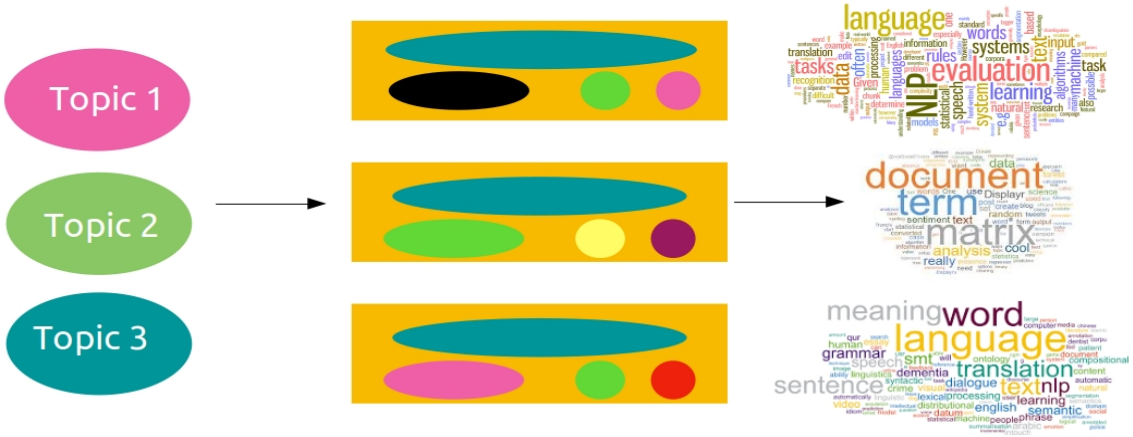
Why Topic Modeling?

Topic modeling is a clever and succesfull way of extracting information.
Main advantages:

- allows for a probabilistic description of clustering
- simultaneous clustering of both sides of the bipartite network:

$$P(\theta \mid G) \propto P(G \mid \theta) P(\theta)$$

.

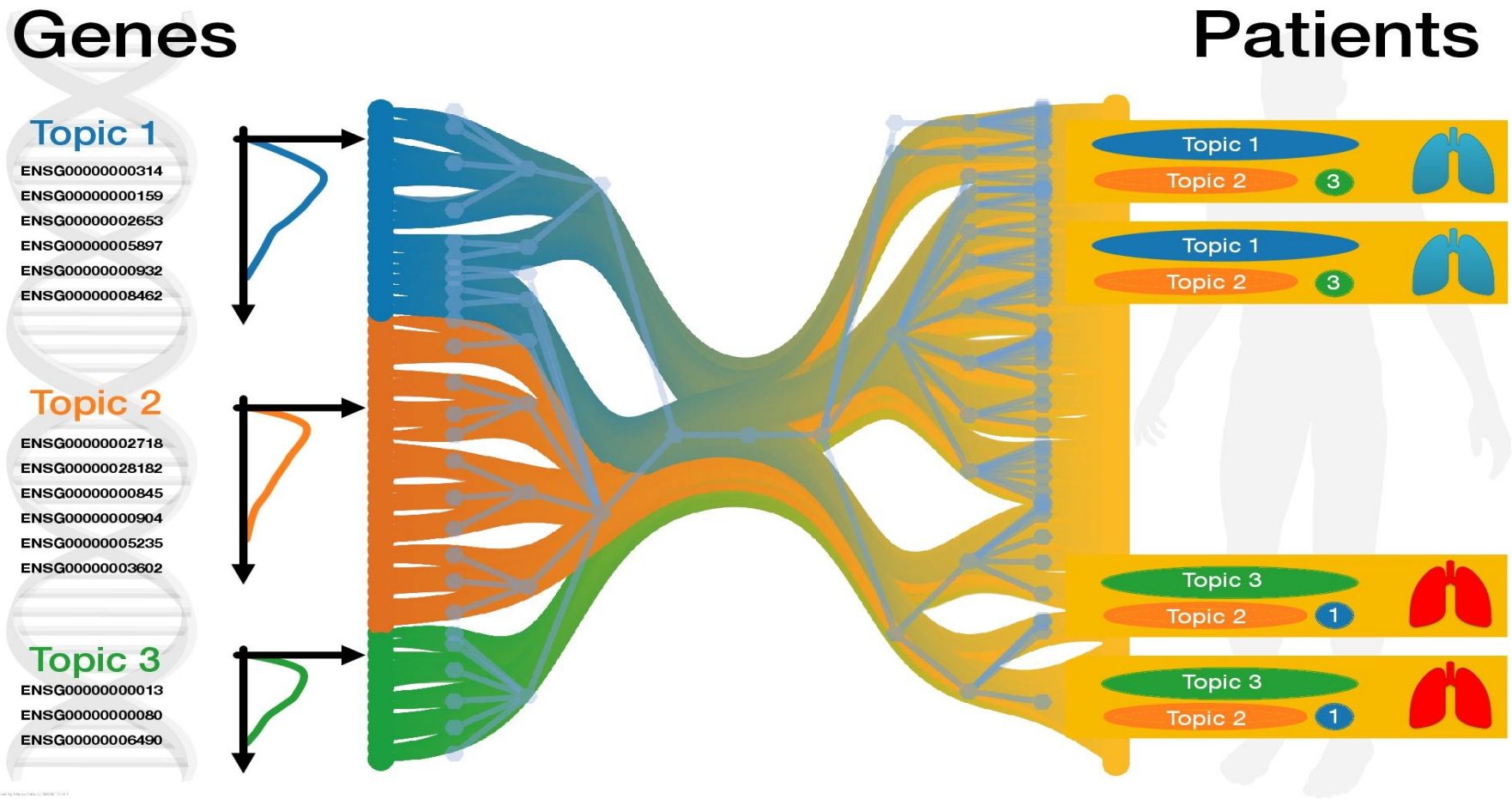


"Latent Dirichlet Allocation." D. Blei, A. Ng, M.Jordan.
Journal of Machine Learning Research 3 (2003) 993-1022

"A network approach to topic models" M. Gerlach, T. P. Peixoto, and E. G. Altmann
Science Advances 4:7 (2018)

Topic Modeling

$$P(\theta|G) \propto P(G|\theta)P(\theta)$$



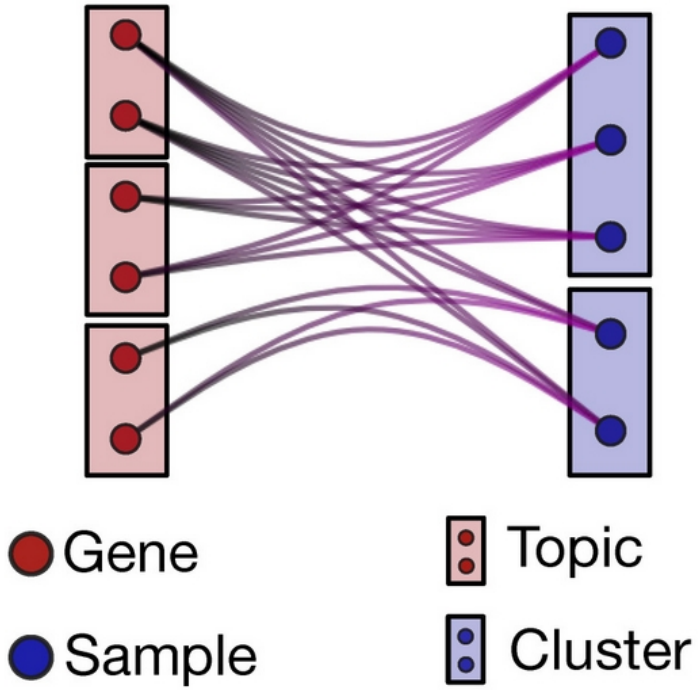
Gene expression data as bipartite networks

Our main goal is the identification of the "*latent variables*" in gene expression data. These data can be represented as bipartite networks. Finding the latent variables is the same as performing *community detection* in a bipartite graph .

samples

genes

	P1	P2	P3	P4	P5
	100	10	0	36	21
	45	0	6	124	0
	163	6	0	72	0
	43	32	12	0	2
	0	0	0	0	80



A topic modeling analysis of TCGA breast and lung cancer transcriptomic data
F Valle, M Osella, M. C. Cancers 12 (2020), 3799

Minimum Description Length

- Hierarchical Stochastic Block Model (hSBM) is a kind of generative model that tries to maximize the probability that the model describes the data

$$P(\theta|\mathcal{A}) \propto P(\mathcal{A}|\theta)P(\theta).$$

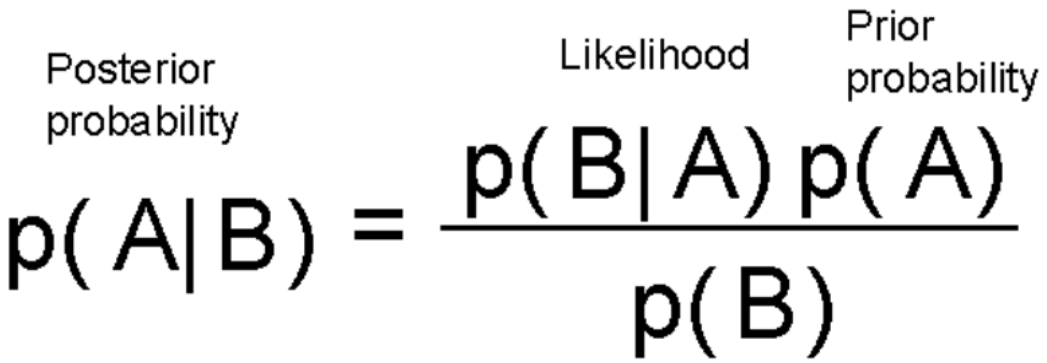
Its approach is completely non-parametric and aims to maximize the posterior probability or, equivalently to minimize the description length

$$\Sigma = -lnP(\mathcal{A}|\theta) - lnP(\theta)$$

Where A is the gene expression matrix, and theta encodes the partition in clusters and topics

Bayesian Inference and Minimum Description Length

Bayesian inference is based on the Bayes Theorem



Which states that our belief (the prior probability) is updated to the posterior probability after we observe the data (likelihood)

Bayesian Inference

In Bayesian inference the same theorem can be translated as follows

$$P(h|D) = \frac{P(D|h).P(h)}{P(D)}$$

Which states that we try to find which hypothesis
h describes the data D, given the data.
In general the space of possible hypothesis is infinite
and we want to maximize the probability that
one particular h is most likely to originate the given data

Bayesian Inference

$$P(h|D) = \frac{P(D|h).P(h)}{P(D)}$$

This amount to find the "Maximum a Posteriori" value of h

h / argmax {P(h|D)}

Since P(D) does not depend on h we have

h_MAP =argmax{P(D|h)P(h)}

Minimum Description Length

From which we have:

$$h_{MAP} = \arg\max P(D|h).P(h)$$
$$= \arg\max \log_2(P(D|h).P(h))$$
$$= \arg\max [\log_2 P(D|h) + \log_2 P(h)]$$
$$= \arg\min [-\log_2 P(D|h) - \log_2 P(h)]$$

Where log2 P(h) is the length of a code which describes the hypothesis (i.e fixes all the free parameters which define our model)
While log2 P(D|h) is the length of a code which describes the ability of the hypothesis to describe the data (i.e. the number of errors that we make when we use h to describe D).

There is a trade off between log2 P(h) and log2 P(D|h)

The algorithm

The algorithm is a sort of Simulated Annealing in which the goal is to minimize the Minimum Description Length (MDL)

The main issue is to find an easy way to define the MDL.
Three steps:

- Discretize the entries of the matrix as positive integers (which is a natural choice in gene expression data and in many complex systems applications)

- Construct a bipartite network in which genes and samples are connected by a number of links equal to the weight

- Write the MDL as the probability distribution of links given the distribution of blocks within the two layers of the bipartite network.

Nonparametric Bayesian inference of the microcanonical stochastic block model. Tiago Peixoto, Phys. Rev. E 95, 012317 (2017) arXiv:1610.02703

The algorithm: MDL

The core of the program is the definition of the MDL: P(k,e,b|A) = P(A| k,e,b) P(k|e,b) P(e|b) P(b)

.

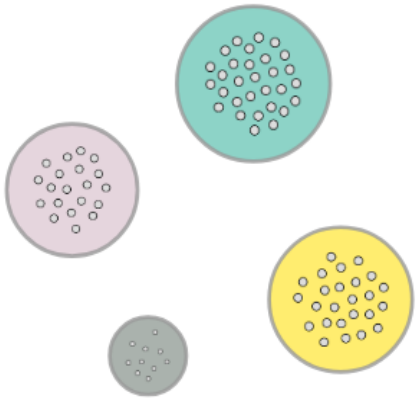
A is the dataset of interest

P(b) denotes the Partition in blocks

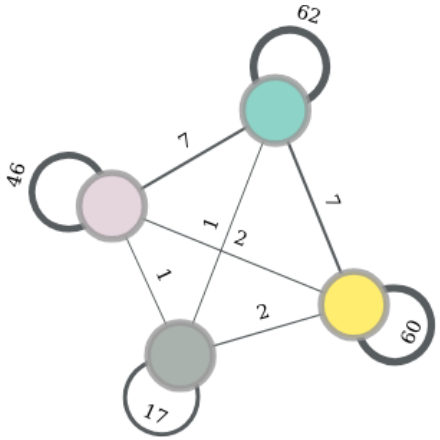
P(e|b) the edge counts given the chosen partition

P(k|e,b) the degree of the nodes given the distribution of links

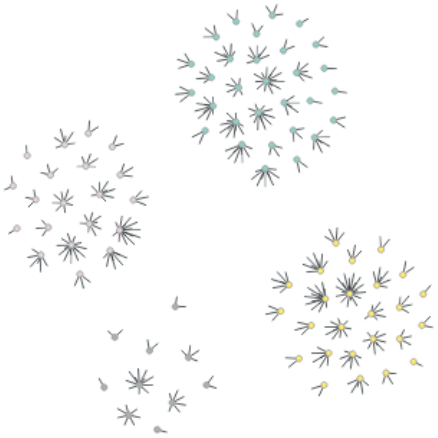
P(A| k,e,b) the probability that given this choice of partition, link distribution and degree of the nodes we find the matrix A



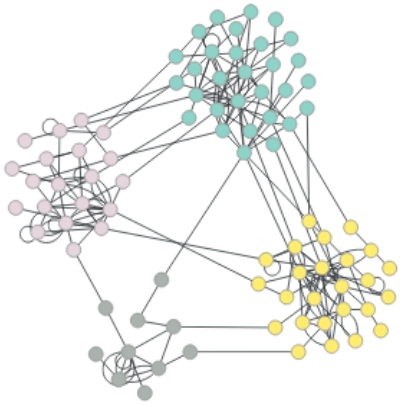
(a) Node partition, $P(\boldsymbol{b})$.



(b) Edge counts, $P(\boldsymbol{e}|\boldsymbol{b})$.



(c) Degrees, $P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b})$.



(d) Network, $P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$.

The algorithm: MDL

The core of the program is the definition of the MDL:

.

$$P(\boldsymbol{A},\boldsymbol{k},\{\boldsymbol{e}_l\},\{\boldsymbol{b}_l\}) = P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}_1)\times P(\boldsymbol{k}|\boldsymbol{e}_1,\boldsymbol{b}_1)\times P(\{\boldsymbol{e}_l\})\times P(\{\boldsymbol{b}_l\})$$

$$= \frac{\prod_i k_i!\prod_{r<s}e_{rs}!\prod_r e_{rr}!!}{\prod_r e_r!\prod_{i<j}A_{ij}!\prod_i A_{ii}!!}\times\prod_r\frac{\prod_k \eta_k^r!}{n_r!}q(e_r,n_r)^{-1}\times$$

$$\prod_{l=1}^{L}\prod_{r<s}\left(\!\!\binom{n_r^l n_s^l}{e_{rs}^{l+1}}\!\!\right)^{-1}\prod_r\left(\!\!\binom{n_r^l(n_r^l+1)/2}{e_{rr}^{l+1}/2}\!\!\right)^{-1}\times\frac{\prod_r n_r^l!}{B_{l-1}!}\binom{B_{l-1}-1}{B_l-1}^{-1}\frac{1}{B_{l-1}}.$$

Where A is the dataset of interest
P(b) denotes the Partition in blocks
P(e|b) the edge counts given the chosen partition
P(k|e,b) the degree of the nodes given the distribution of links
P(A| k,e,b) the probability that given this choice of partition, link distribution and degree of the nodes we find the matrix A

Nonparametric Bayesian inference of the microcanonical stochastic block model.
T. Peixoto, Phys. Rev. E 95, 012317 (2017) arXiv:1610.02703

The algorithm

- The system is initialized with a random partition of samples and genes.

At each step a sample is chosen at random and moved from one cluster to another one (chosen again at random) and the move is accepted if there is a decrease in the MDL . The same is done on the gene layer.

The process stops when the MDL does not decrease any more.

The whole process is repeated O(10) times and the realization with the minimum MDL is chosen.

Nonparametric Bayesian inference of the microcanonical stochastic block model. Tiago Peixoto, Phys. Rev. E 95, 012317 (2017) arXiv:1610.02703

The algorithm

However this procedure has a very slow convergence toward the minimum. One can improve its behaviour with two tricks:

- Chose the move according to the probability:

$$p(r \to s|t) = \frac{e_{ts} + \epsilon}{e_t + \epsilon B},$$

Which is the probability of accepting the move of a node with a neighbour t from group r to group s, being e_ts the number of edges between groups t and s, e_t the total number of edges connected to group t, B the total number of blocks and epsilon a parameter greater or equal to zero.

- Use a "Multilevel Agglomerative Algorithm": which consists in equilibrating the chain for a larger number of groups, and then merging the groups using the same algorithm used for the block membership moves

The algorithm

The above choice does not fulfill detailed balance, so it must be corrected leading to the following acceptance

$$a = \min\left\{ e^{-\beta \Delta\Sigma} \frac{\sum_t p_t^i p(s \to r|t)}{\sum_t p_t^i p(r \to s|t)}, 1 \right\},$$

Where:

1) Σ is the MDL, i labels the node that we are moving, p^i_t is the fraction of links starting from i and reaching a node in block t

2) p(s → r|t) must be computed after the proposed r → s move (i.e. with the new values of e_tr), whereas p(r → s|t) must be computed before.

3) The parameter β plays the role of an inverse temperature, and can be used to sample partitions according to their description length (β = 1) or to find the ground state (β → ∞)

The algorithm

The two non-trivial features of the algorithm are:

1) the number of clusters and topics is directly encoded in the MDL and is thus an output of the program.

2) The MDL is chosen so that to allow a hierarchical organization of partitions: from time to time a move is allowed to create a new level of partitions

The problems of interest are typically non-convex with several competing minima, thus different runs of the program may lead to different partitions. If the dataset contains reliable information then these different realizations differ only for small changes in the cluster and topic organization.

Output of the program

$$P(\text{topic}|\text{sample}) = \frac{\text{number of half-edges on sample coming from topic}}{\text{number of half-edges on sample}}$$

$$P(\text{gene}|\text{topic}) = \frac{\text{number of half-edges to topic going to gene}}{\text{number of half-edges to topic}}.$$

We use the centered P(topic|sample)

$$\overline{P}(\text{topic}|\text{sample}) = P(\text{topic}|\text{sample}) - \frac{1}{R}\Sigma_{s\in\text{samples}}P(\text{topic}|\text{s}),$$

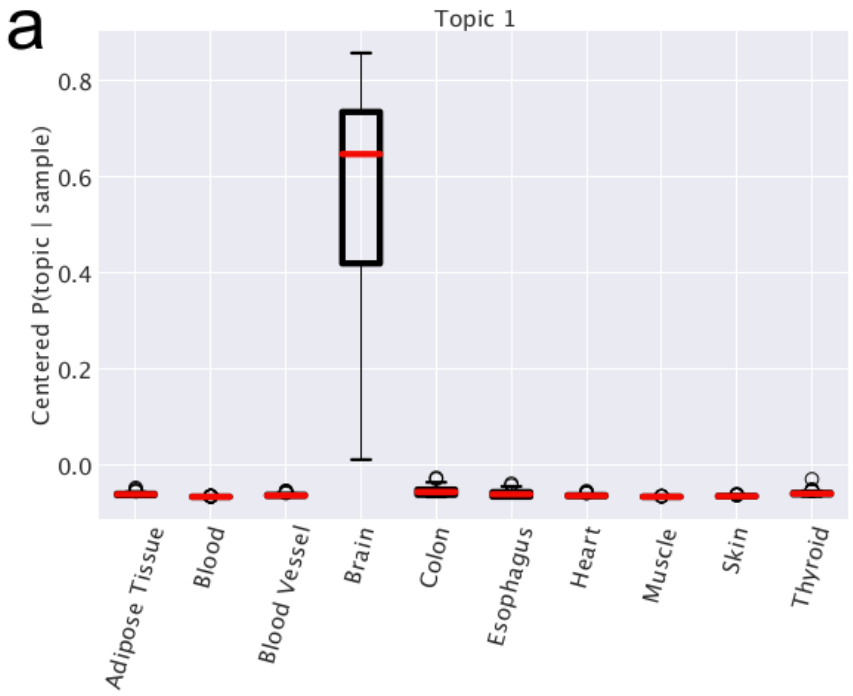
The algorithm

- hSBM (and its improved version nSBM) is freely available:

https://github.com/BioPhys-Turin/nsbm

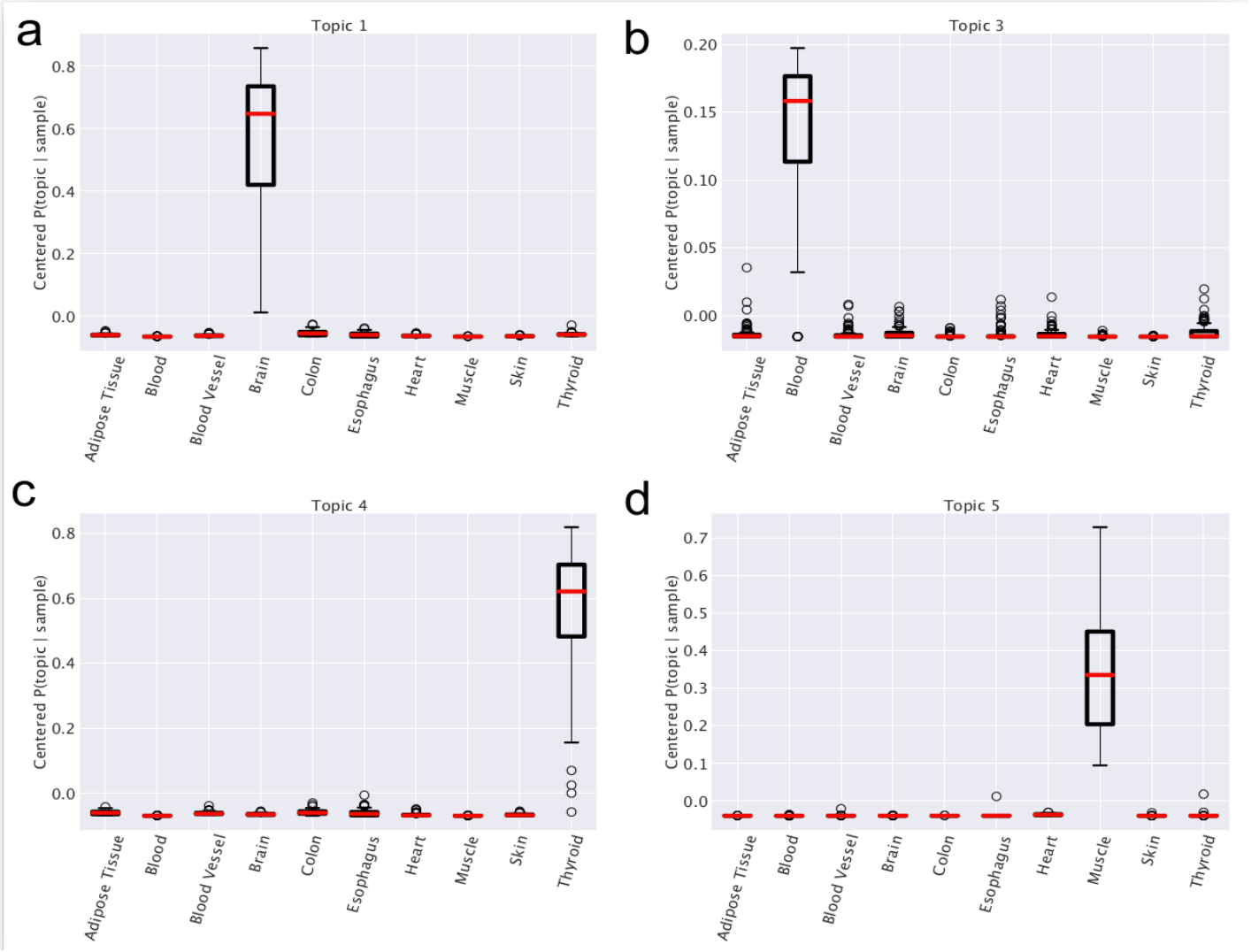
From Topic Modeling to Marker genes

Topics are nothing else than sets of genes. Each gene is associated to a probability to belong to the topic: ***P (gene | topic)*** If a topic is strongly associated to a particular cell type (i.e. a large ***P (topic | sample)*** for samples belonging to a given cell type) then the genes belonging to the topic are likely to be markers of that cell type.

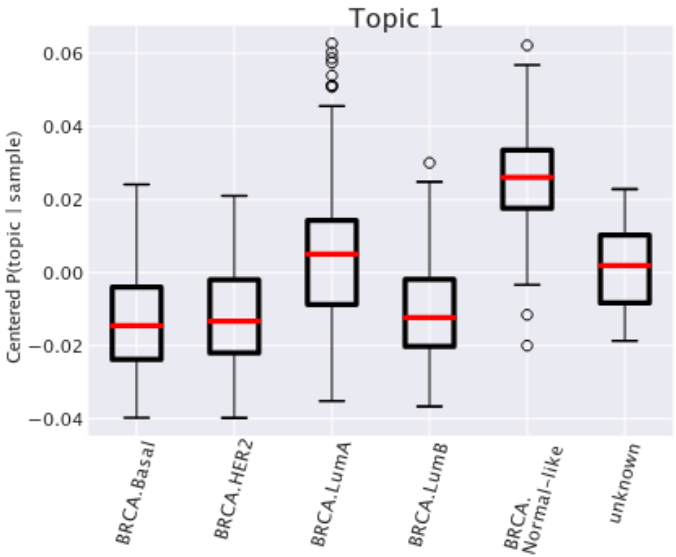


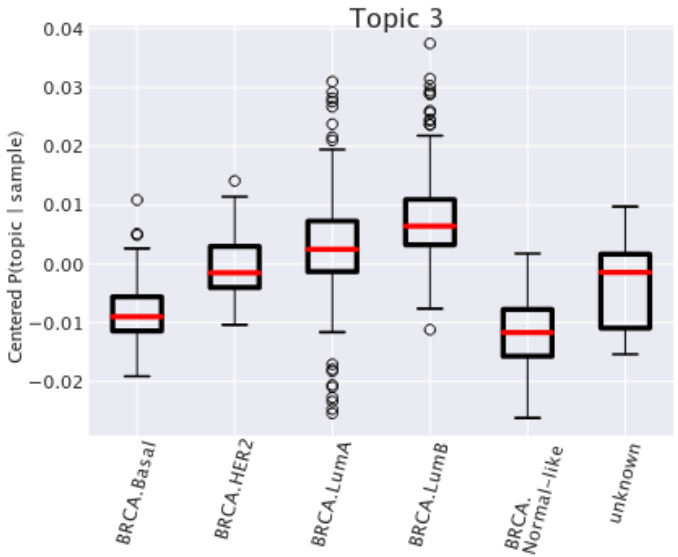
Example: topic modeling analysis of gene expression data of ten human tissues (data taken from GTEX). Topic 1 is strongly enriched in Brain. The genes contained in it are candidate markers

From Topic Modeling to Marker genes



Markers of Cancer subtypes





Term	FDR *q*-Value
Topic 1 (72)	
SMID_BREAST_CANCER_LUMINAL_A_UP	3.03×10^{-40}
SMID_BREAST_CANCER_NORMAL_LIKE_UP	3.23×10^{-15}
Topic 3 (50)	
SMID_BREAST_CANCER_BASAL_DN	4.89×10^{-9}
VANTVEER_BREAST_CANCER_METASTASIS_UP	3.46×10^{-8}

Data on breast cancer samples from **TCGA**.

A **Gene Ontology** analysis confirms the quality of the markers

Topic Modeling: benchmarks

Tests of Topic modeling results:

1) On the clusters side:
   - ability to reconstruct know classifications of samples.
     Scoring function: Normalized Mutual Information (NMI)
   - clusters collect together samples with similar biological features:
     Example: survival probability

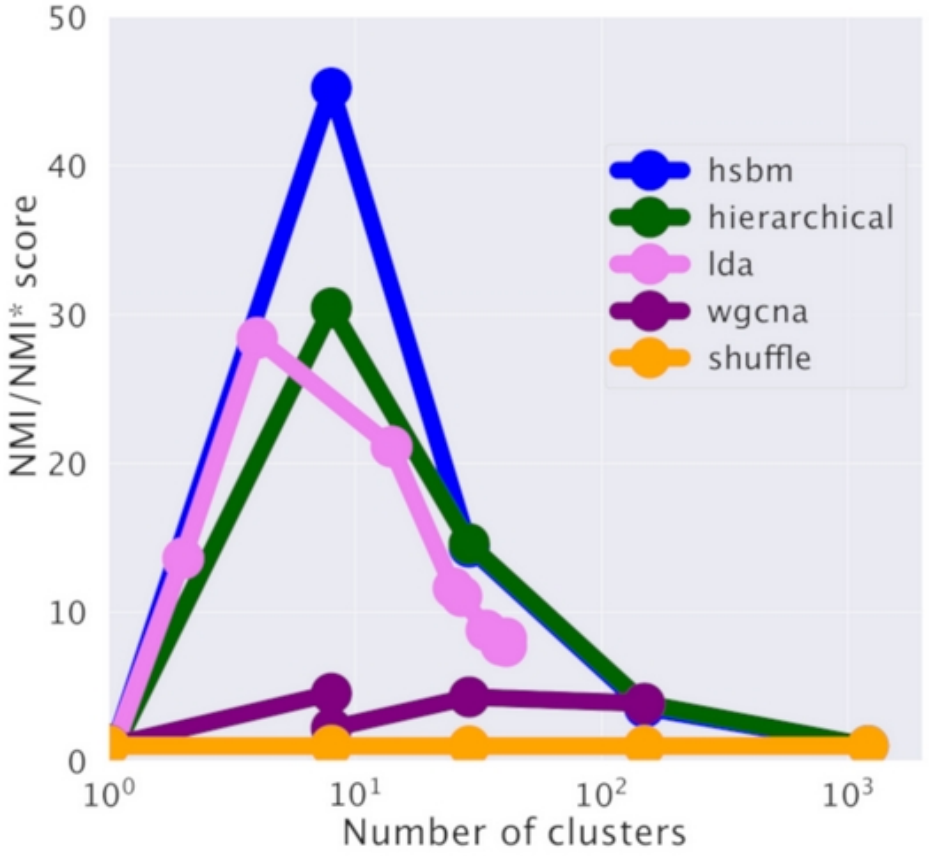
2) On the topic side:
   - ability to associate relevant biological information to the topics.
     Tools: Gene Onotology, Genomic annotations

   - Topics as a reduced representation of the complexity of the dataset
     Tools: performance of classifier constructed in the topic's space

A topic modeling analysis of TCGA breast and lung cancer transcriptomic data
F Valle, M Osella, M. C. Cancers 12 (2020), 3799

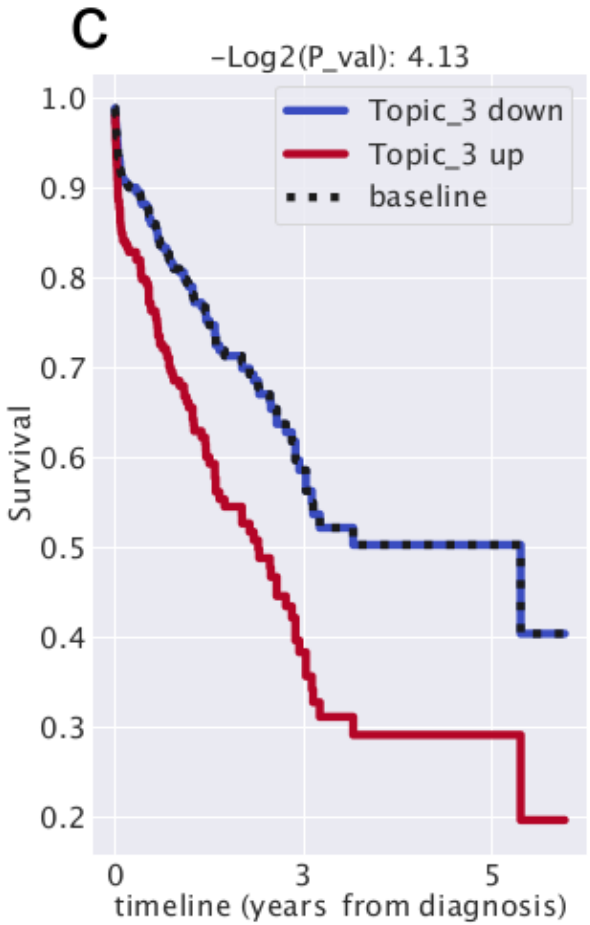
Topic Modeling: breast cancer

hSBM outperforms other clustering algorithms



Topics are markers of survival

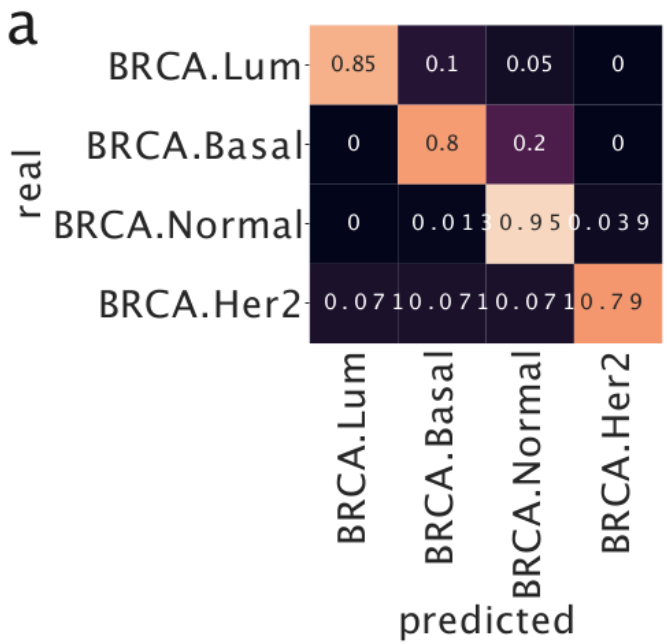
Cohorts of patients selected among those with low (or high) presence of a given topic show relevant differences in survival probability.



A topic modeling analysis of TCGA breast and lung cancer transcriptomic data
F Valle, M Osella, M. C. Cancers 12 (2020), 3799

Topics can be used as classifiers

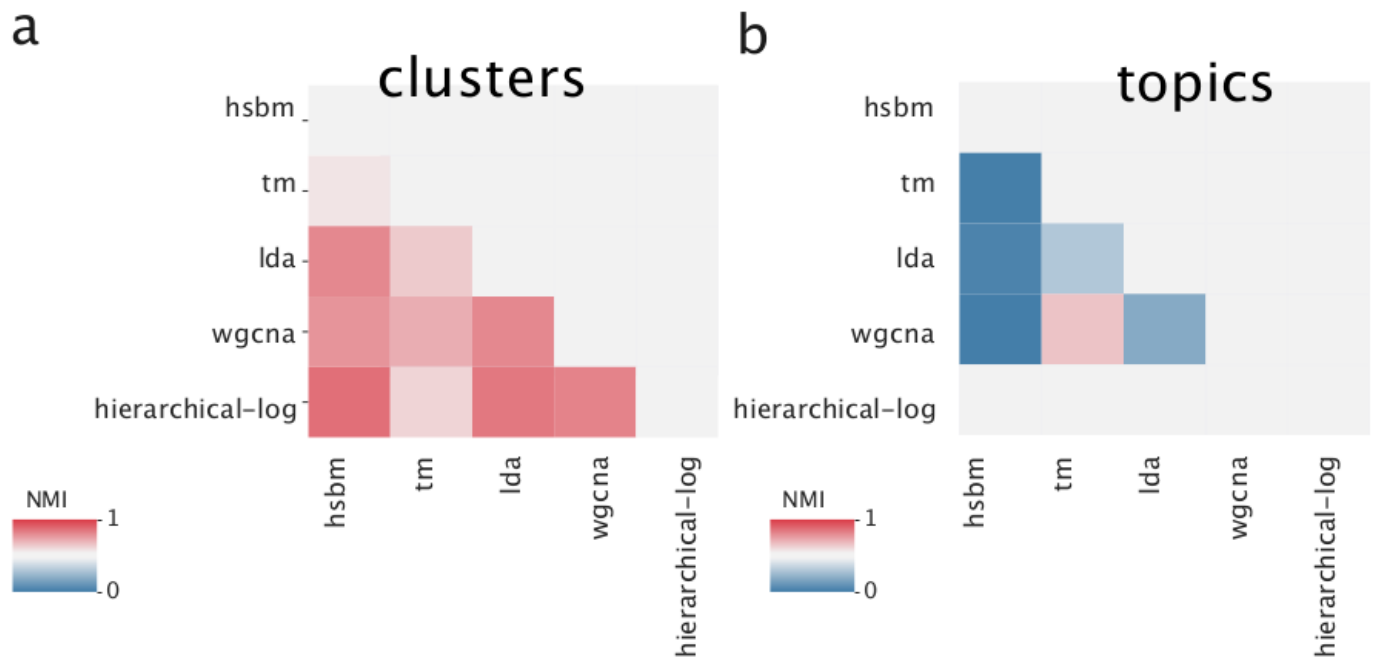
Topics can be used as embedding space to train a classifier which shows very good performances in classifying a new sample into breast cancer subtypes.



A topic modeling analysis of TCGA breast and lung cancer transcriptomic data
F Valle, M Osella, M. C. Cancers 12 (2020), 3799

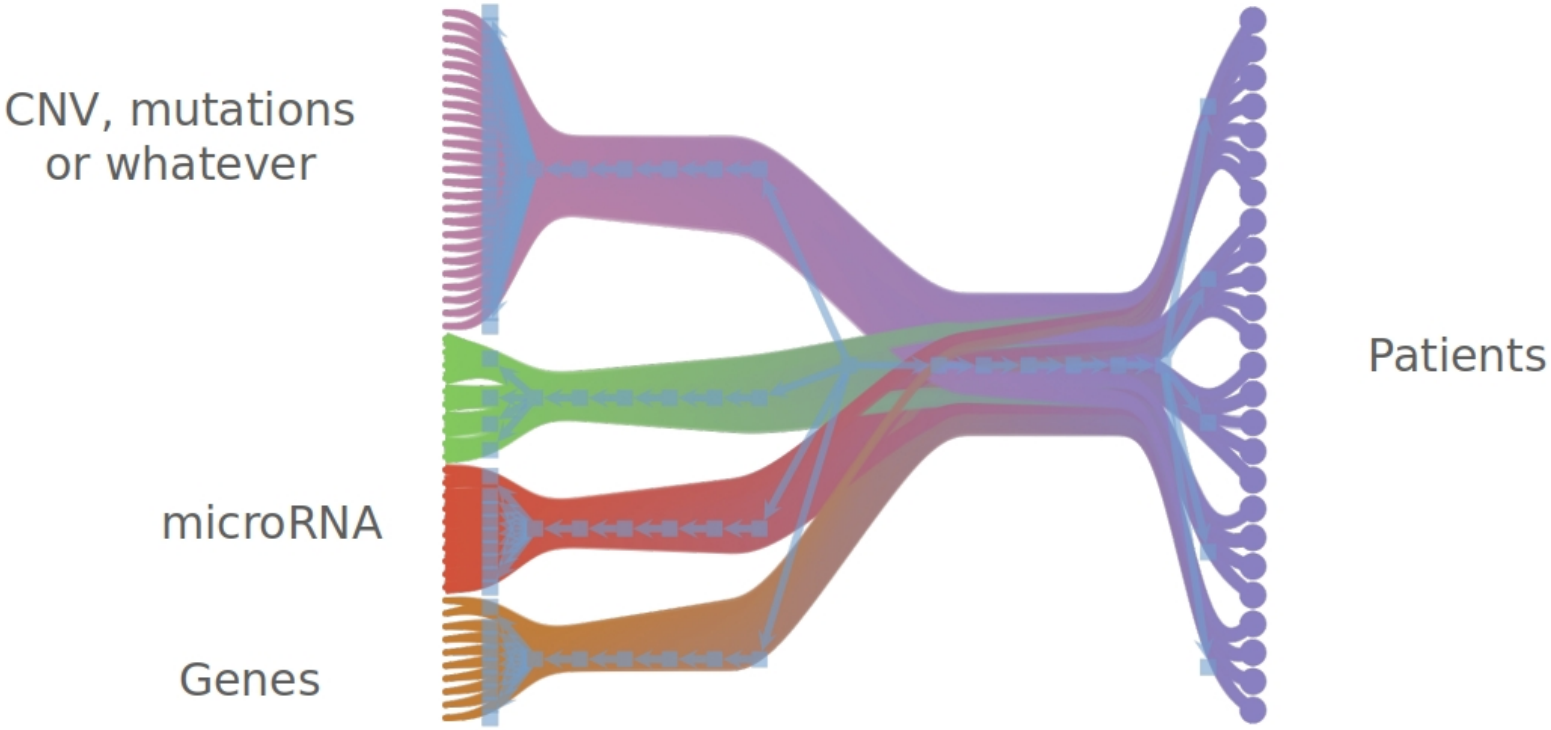
Topics strongly depend on the chosen prior

The organization in cluster does not depend too much on the prior but the gene content of the topics is strongly affected by the choice of the prior



Exploring the latent space of trascriptomic data with topic modeling.
F. Valle, M. C. M.Osella NAR Genom. Bioinform. 7(2) (2025)

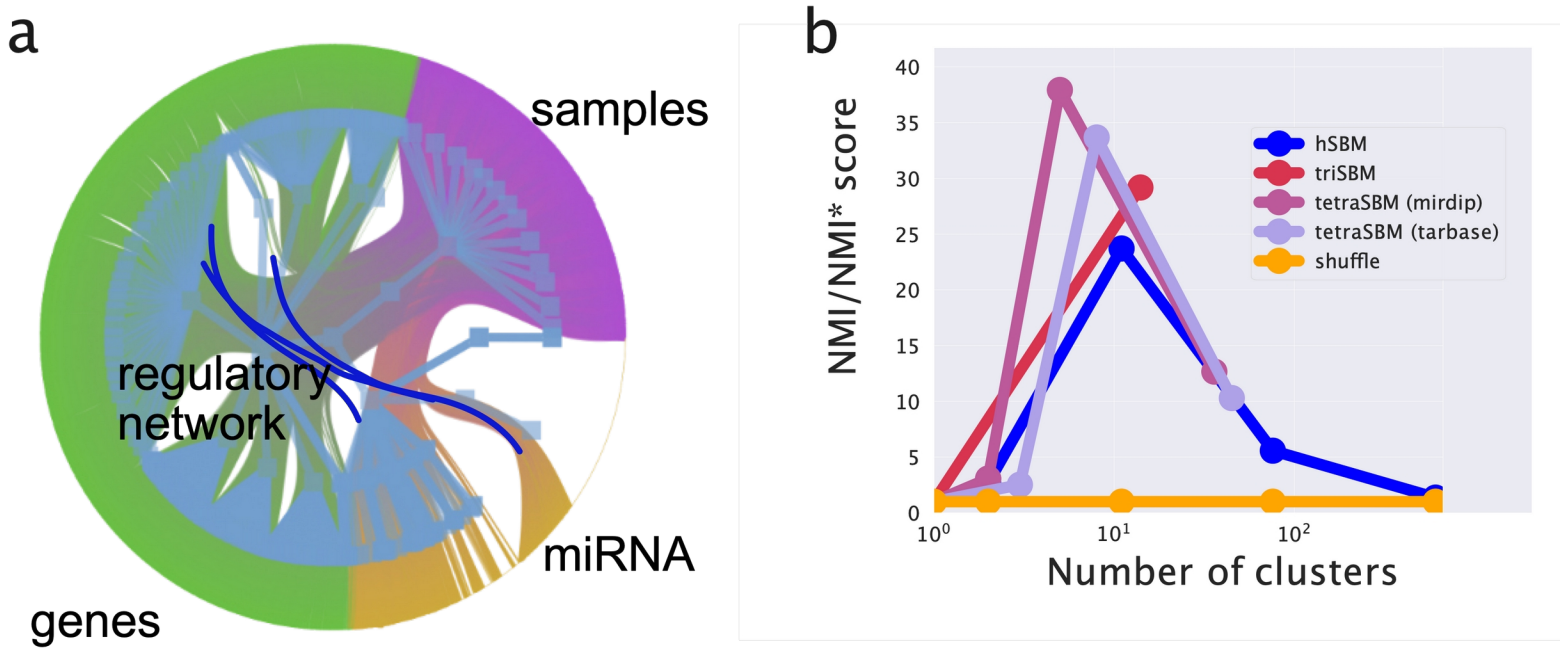
Extension of hSBM to multi-omics: nSBM



Multiomics Topic Modeling for Breast Cancer Classification
F Valle, M Osella, M. C. Cancers 14 (2022), 1150

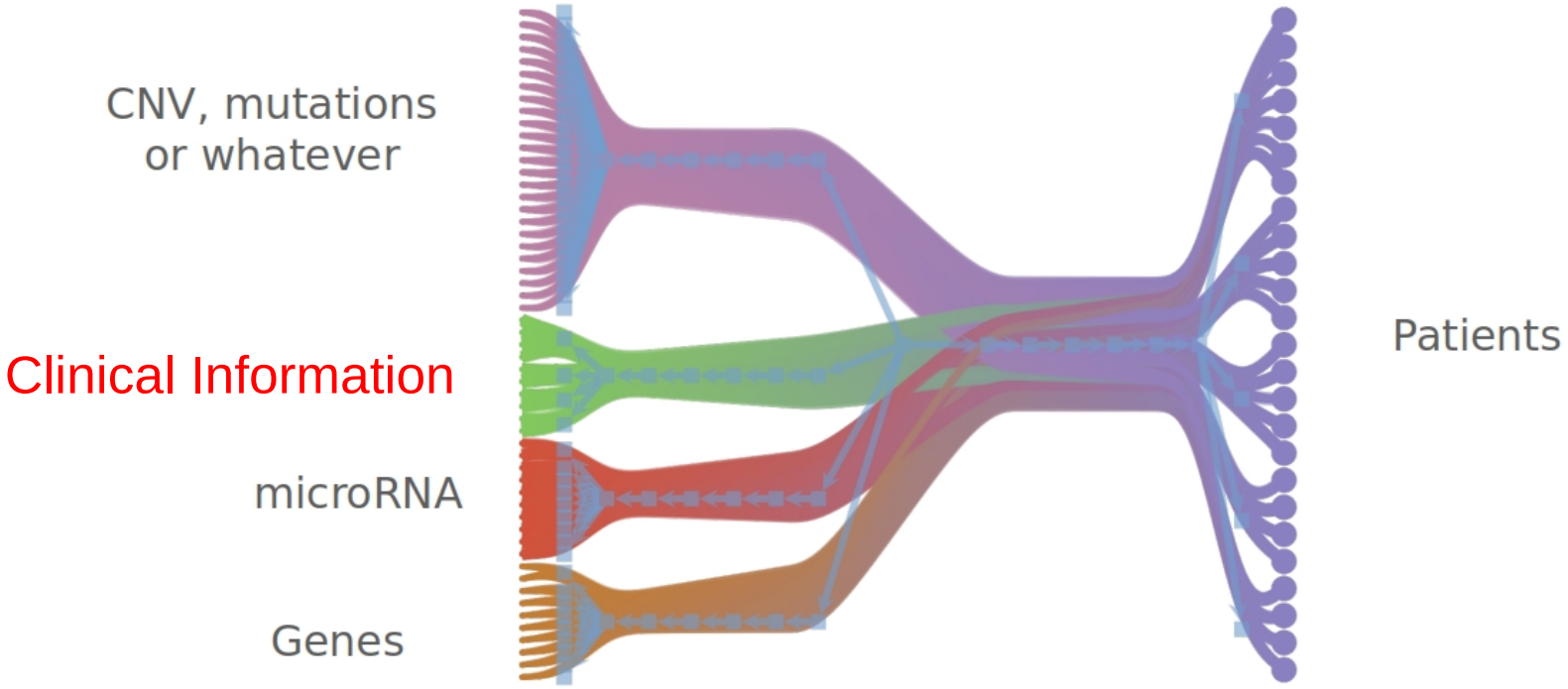
Adding regulatory info in nSBM

In the nSBM framework one can also include regulatory information, adding new links among different sectors of the n-partite network. This additional information improves the quality of the classification



Adding human annotations in nSBM

In the nSBM framework one can also include a layer with human annotations or clinical information on the samples. In this case the goal is not any more to classify samples but to improve the quality of topic and thus of markers.



Take home message

- Complex systems typically have a modular structure which contains relevant information.

- Bayesian inference based on the minimum description length is a powerful tool to identify this modular organization.

- Topic modeling based on a hierarchical version of Stochastic Block Modeling (i.e. Bayesian inference with a flat prior) outperform other types of modular analysis

- Topics modeling can be used to study gene expression data with remarkable biological results:
Identification of cancer subtypes,
Identification of marker genes
Definition of an embeddig space to construct efficient subtype classifiers

References

A topic modeling analysis of TCGA breast and lung cancer transcriptomic data.
F. Valle, M. Osella, M. C. Cancers 12 (2020), 3799

Multiomics Topic Modeling for Breast Cancer Classification
F .Valle, M. Osella, M. C. Cancers 14 (2022), 1150

Identification of Interpretable Clusters and Associated Signatures in Breast Cancer Single-Cell Data: A Topic Modeling Approach.
G. Malagoli et al. Cancers 12 (2024), 1350

Exploring the latent space of trascriptomic data with topic modeling.
F. Valle, M. C. M.Osella NAR Genom. Bioinform. 7(2) (2025)

Topic modeling analysis of Allen Human Brain Atlas
L. Pizzini, F. Valle, M. Osella, M. C. Sci Rep. (2025) 6928.

Download

- hSBM (and its improved version nSBM) are freely available:

https://github.com/BioPhys-Turin/nsbm

Topic Modeling Collaboration

UNIVERSITÀ DI TORINO

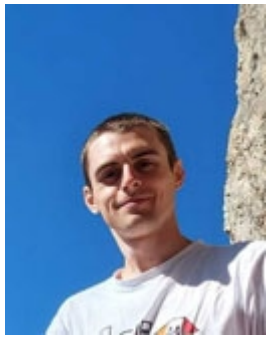
INFN Istituto Nazionale di Fisica Nucleare

Matteo Osella

Filippo Valle

Letizia Pizzini

Loredana Martignetti



Gabriele Malagoli