Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

FISICA

Ciclo XXXIII

Settore Concorsuale: 02/D1

Settore Scientifico Disciplinare: FIS/07

STATISTICAL AND NETWORK DYNAMICS APPROACHES TO CANCER GENOMICS DATA ANALYTICS

Presentata da: Tommaso Matteuzzi

Coordinatore Dottorato	**Supervisore**
Prof. Michele Cicoli	**Prof. Gastone Castellani**

Esame finale anno 2021

Contents

Introduction

Cancer development is a complex process involving almost all the layers of biological hierarchies, from genome to gene regulatory network and signalling between cancer cell and its micro-environment. However, at its foundations cancer is an evolutionary process.

During the lifetime of an organism, its cells naturally undergo a random mutational process. As a consequence, some cells can acquire a selective advantage, i.e. an increased reproduction rate, on the others. When a cell acquires a set of mutations that allows it to proliferate without control, its progeny has the potential of invading tissues and metastasize [79].

In the last decades, cancer genome wide studies provided strong evidence supporting the contribution of somatic mutations to tumour development. Cancer genomes display on average from tens to hundreds somatic mutations, comprising single-base substitutions, deletions and insertions of one or a more bases as well as chromosomal abnormalities. At the same time, however, these studies highlighted a wide heterogeneity in mutational patterns, across and within cancer types, to the extent that two different tumours almost never show an identical somatic mutation profile or, sometimes, a single mutation in common. Moreover, not all somatic mutations found in cancer genomes are involved in the development of the disease. While the so called *driver* mutations confer a selective advantage to cancer cell and are positively selected during cancer evolution,*passengers* mutations do not confers growth advantage and do not contribute to cancer development [83].

A challenge of past and current research has been the identification of tumour-causing mutations. However, it is increasingly acknowledged that cancer is a perturbation of cell state, resulting from the interplay of dysfunctional molecular constituents and efforts are needed to understand the mechanisms by which a genetic alteration pattern results in a specific phenotype. Moreover, without a systemic view of cancers it remains difficult to develop therapeutic strategies for treating them [21].

Uncovering interactions among genetic alterations and understanding their consequences at multiple scales is thus central in modelling genotype- phenotype associations.

In this thesis we focused on some statistical and physical methods which attempt to tackle the problem of cancer genetic heterogeneity and its relationship to higher level biological properties.

Recently, systematic experimental screens have made available several reconstructions of the network of interactions among macromolecular constituents of the cell. Such network, often referred to as the interactome, allows to gain a system level view of mutational patterns, providing a framework to understand how mutations act together to give rise to the cancer phenotype.

A mutation often result in the impairment of a protein function and, from a network point of view, it can be seen as a node removal. It is thus of interest to study how cancer related alterations impact the global network topology.

Since different reconstructions of the interactome exist, we collected from the web twenty interactome reconstructions, selected among the most widely used in genomic integration studies and, in the first chapter of this thesis, we compare them from a topological perspective by analysing their global and local properties. We then study their overall resilience under nodes perturbation and we compare the impact of random node deletions to removal of cancer related genes derived from mutational data of several cancer types.

Notwithstanding their daunting heterogeneity, cancer genetic alterations are thought to impact some specific gene groups. In other words, one expects that, instead of being caused by specific gene mutations, cancer stems from the impairment of one or more biological functions due to mutations of genes taking part in them. This hypothesis, along with the observation that different patterns of mutations lead to different responses to treatments and in turn in different survival outcomes, highlights the importance of stratifying patients based on their genetics and cytogenetic alterations.

To this end, in the second chapter, we focus on hierarchical non parametric bayesian methods. Latent topic models, such as Latent Dirichlet Allocation and Hierarchical Dirichlet Process allow to model hidden structures in the data and fit well with the hypothesis that cancer mutations impact specific gene groups in different proportions.

In the second part of the chapter, we study a cohort of 2043 patients affected by Myelodysplastic Syndromes, characterized by a panel of more the fifty genetic mutations and chromosomal abnormalities. By applying Hierarchical Dirichlet Process and Bayesian Networks, we draw a picture of the genetic landscape of the disease.

From a more general perspective, the view of cancer as an evolutionary process frequently implies the assumption of a direct and univocal genotype-phenotype relationship. However, as for cell differentiation, such genetic deterministic view is not sufficient to account for several experimental observations as, for example, the similarity of gene expression patterns in cancer with different mutational profiles [46].

In the third chapter, we focus on the hypothesis of cancer as an abnormal attractor in the epigenetic landscape of the cell. In the first part, we set the mathematical framework of gene regulatory networks and we introduce the concept of epigenetic landscape. Since, recently, the introduction of single cell sequencing made available gene expression profiles of thousands of cells, we study the connection between the empirical distribution of cell in the gene expression state space with network laplacian based manifold reconstruction techniques and their application for inferring the epigenetic landscape from data.

Chapter 1

Interactome Reconstructions Topology Comparison and Resilience

1.1 Introduction

In the cell, proteins interact with one another and with other molecular components, such as metabolites and nucleic acids, to perform biological process. Genetic alterations can lead to disruptions of some of these processes which, in turn, result in higher phenotypic effects, for example increased cell reproduction rate and cancer [4, 46].

Understanding how a given genetic alteration pattern links to a specific phenotype, i.e. inferring genotype-phenotype maps, is a central task to explain the genetic architecture of complex diseases. To this end, the characterization of the complex web of macromolecular interactions occurring within human cells, usually referred as the **interactome**, is essential since it allows to capture systems-level patterns (e.g. active network regions, disease modules) and go beyond the knowledge attainable analysing each genetic perturbation as if it affected the phenotype by acting independently [21, 52].

As a map to guide our understanding of how alterations perturb the system as a whole, the interactome is currently being used in several approaches and many network-based methods have been developed to solve problems in integrative analyses, namely, to understand molecular behaviours, to find disease subtypes and to predict an outcome or phenotype [8, 55]. Indeed, the interactome represents a powerful framework to integrate omics datasets [14, 13, 82, 27, 87].

Loosely, the interactome can be defined as a graph where each node represents a gene product (generally a protein) and an edge represents some kind of, directed or undirected, interaction (e.g molecular docking) between two of them.

In contrast to human genome and transcriptome, the interactome is not uniquely defined and several recontructions of it can be defined depending on the nature of interactions considered. Moreover, even when interactions are of a specific kind, for example biophysical PPI, their mapping is still far from completeness [60]. For this reason, a unique reference model is not available for the interactome and, currently, different reconstructions exist.

1.1.1 Gene-Centered Interactomes

All interactome reconstructions are *gene-centered*. Nodes are gene identifiers and edges represent different types of interactions involving genes and gene products, see figure 1.1.

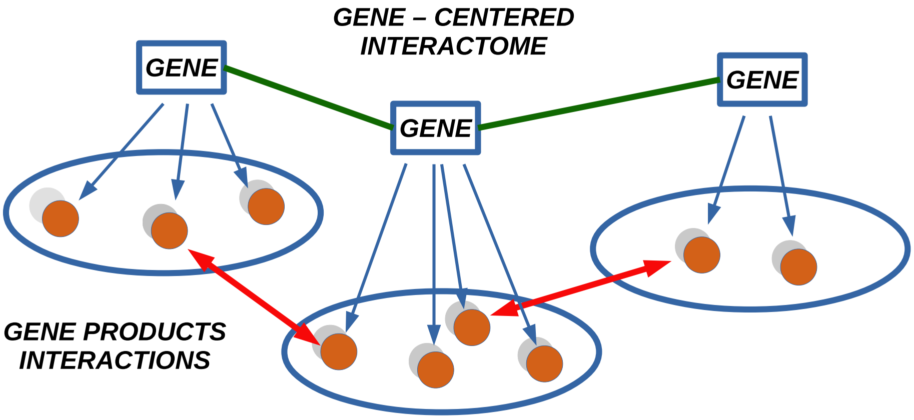


FIGURE 1.1: Gene-Centered Interactome Representation. Since to each gene can be associated different gene products (orange dots), e.g different protein isoforms, the gene-centered view, at the cost of losing some information, allows to simplify network analysis and integration of omics datasets which usually carry information at gene level.

This representation simplifies the many types of players (e.g. DNA sequence, protein isoforms) and interactions (PPI, protein-DNA) actually involved, providing a useful model to integrate many other data types that are attributable to genes, like the scores (e.g. p-values, fold-changes, etc.) emerging from omics assays.

In such gene-centerd view, a node represents the gene itself or any of its products, while edges accommodate both biophysical (direct) and functional (indirect) interactions.

Biophysical interactions mainly include PPI and protein-DNA interactions (PDI). Therefore, a PPI between genes A and B represents any PPI between any pair of products of the two genes; while a PDI between A and B indicates the binding between any protein encoded by A to gene B.

Functional interactions represent any type of biological relation between two genes which does not involve a direct contact, for example: co-expression relations, genetic interactions and links between enzymes that catalyse adjacent reactions in metabolic pathways.

Gene-centered interactomes differ in terms of types of interactions included, data sources and assembling procedure. Among those available in current literature, we distinguished three classes:

- *High-throughput biophysical* interactomes (**HTBP**), state-of-the art in terms of reconstructing the interactome in a biological model, where PPIs are detected by means of a high-throughput assays [60] (for example, yeast two hybrid screening or affinity purification followed by mass spectrometry and co-fractionation).

- *Integrative* interactomes (**ITC**),which integrate data from both primary databases and meta-databases. Primary databases collect experimental data from small and large scale studies, while meta-databases integrate and unify interactions from multiple primary databases.

- *Integrative-predictive* interactomes (**IP**), which contain interactions collected from multiple sources as well as predicted interactions, hypothesized on the basis of a series of evidences, principally co-expression, co-participation in molecular pathways or even co-occurrence in scientific publications [80].

In such heterogeneous and incomplete scenario, which lacks a reference model, it is not trivial to decide which interactome or interactomes is most appropriate given a research task (for example, cancer mutated gene prioritization). To guarantee a good coverage of the totality of the genes, it is common to perform network-based analysis using interactomes defined combining multiple sources [14]. In some works, the results obtained using different interactomes on the same data are compared assessing the variation of the studied outcome (e.g. tumor stratification [43]) or joined in a consensus [56]. However, quite often, a single interactome is used [82]. Recently, a benchmark on the performance of several interactomes on a particular task, namely disease prioritization, found that the choice of interactome matters greatly [44].

In the first part of this chapter, we study the topological properties of 20 interactomes reconstructions to shed light on their heterogeneity, redundancy and specificity from topological and applicative perspectives. In section 1.2 we introduce the database considered. We then compare them in terms of degree distribution, centrality, clustering and sharing of hubs in section 1.3. In the second part of the chapter, we try to characterize interactome resilience, that is, how topology is affected by multiple gene failures and in particular by cancer related gene alterations.

The knowledge emerging from our analyses summarizes the current situation and can be useful to guide the choice of interactomes in future applications.

1.2 Interactomes Databases Collection and Harmonization

We selected a panel of 19 interactomes comprehensive of the most widely used interactomes in the literature and representative of the three classes introduced above. In addition, since HTBP interactomes suffer from a high rate of false negative interactions detection and are generally more sparse the the others, we studied the interactome resulting from their union (acronym BUN in table 1.1). In Table 1.1, are reported interacrome version, class and size. For sake of brevity, in the following we will refer to interactomes using the assigned acronyms (ID).

The original genes or protein identifiers chosen by the authors of each interactome (Entrez Gene id, gene symbols, Uniprot, Ensembl transcript, Ensemble gene, Ensemble protein, iRefIndex icrogid) were mapped to Entrez gene identifiers. Mappings between Entrez Gene identifiers and other identifiers were collected from Entrez Gene FTP site ftp://ftp.ncbi.nih.gov/gene (26/02/2019), Uniprot FTP site https://www.uniprot.org/downloads, R package biomaRt (26/02/2019), and, where available, by the authors of the interactomes (STRING: https://string-dborg/mapping_files/entrez, iRefIndex: https://irefindex.vib.be/wiki/index.php). Some interactomes included a minor number of interactions involving identifiers from non human species that we discarded.

Apart from the four HTBP interactome, that were independently derived, the remaining interactomes share interaction sources databases.

All interactomes included a largest connected component (LCC), which involved more than the 99% (median value) of the total genes of the interactome, and a few minor components: only the LCCs were considered for our study.

The two interactomes derived from STRING and designated as S04T and S07T were obtained selecting only the links with confidence score 0.4 and 0.7, respectively. The other two interactomes derived from STRING, S04 and S07, were obtained recalculating the confidence score without the contribution of text mining, by means of the script provided at http://string-db.org/download/combine_subscores.py.

When multiple pairs of Ensembl protein identifiers, characterized by different STRING confidence scores, mapped to the same pair of Entrez gene identifiers, the highest score was considered as representative of the interaction between the two genes. iRefIndex complexes were transformed into a list of binary interactions following the so-called spoke model (interactions occur only between the bait protein and each of the others) if the bait protein was indicated, and, otherwise, to the matrix model (all-pairs interactions) (see https://irefindex.vib.be/wiki/index.php/README_MITAB2.6_for_iRefIndex_15.0).

ID	Name (version) [REF]	Class	# Interactions	# Genes
BX	Bioplex (4a) [48]	HTBP	56 401	10 880
CF	Cofrac15 [41, 84]	HTBP	15 513	3 191
HURI	HURI [59, 75, 88]	HTBP*	27 084	8 029
QU	QUBIC [42]	HTBP	14 696	4 379
BN	Biana [33]	Integrative	339 698	13 246
HINT	HINT (April 2019) [29]	Integrative	164 255	14 372
HP	HIPPIE (2.2) [5]	Integrative	404 020	18 038
INCT	Intact (2019_07_03) [66]	Integrative	174 388	15 539
IR	irefindex (15.0) [73]	Integrative	476 437	17 522
DMND	Diamond [36]	Integrative	138 045	13 244
NCBI	NCBI (15/09/2017) [19]	Integrative	326 859	17 655
CP	ConsensusPathDB (guildify 2.0) [3]	Integrative	273 005	16 066
MN	MULTINET [51]	Integrative	105 573	13 387
FP60	FPCLASS [54]	Integrative-predictive	258 107	10 403
IBMP	InBio_web (core 2019_02_26) [72]	Integrative-predictive	652 636	17 458
S04	String, CS >0.4 (v11) [80]	Integrative-predictive	490 587	15 800
S04T	String including TM, CS >0.4 (v11) [80]	Integrative-predictive	986 054	18 863
S07	String, CS >0.7 (v11) [80]	Integrative-predictive	357 054	12 747
S07T	String including TM, CS >0.7 (v11) [80]	Integrative-predictive	417 012	16 721
BUN	Biolplex, Cofrac15, Huri, Qubic	HTBP	109261	13925

TABLE 1.1: Interactomes Databases. Version and size of each database are reported. For sake of clarity of the text, we assigned to each interactome a short ID. (*) the interactome contains a minor number of biophysical interactions manually curated from small studies.

1.3 Overall Properties

Comparing and classifing networks from a global perspective is a central problem in network science. *Global properties* can be defined as statistical properties of the network as a whole, as, for example, the degree distribution, the mean network connectivity or the mean distance between nodes. Networks with similar global properties, and in turn dynamical systems defined on them, often share similar behaviours. For example, networks with the same degree distributions show analogous diffusion [64] and synchronization patterns [74] or resilience under node or link removal [26].

However, when two networks are two partial reconstructions of the same underlying physical system, as is the case for the interactomes, also a *local structure* comparison is of interest. In fact, when a significant number of interactions lacks, global properties could differ but some local properties, as for example the centrality or the neighbourhood of a specific node, could be preserved. In what follow we will first compare the 20 interactome from a general topological perspective.

As a first step, we compared interactomes in term of sizes and overlaps of nodes and links. As shown in figure 1.2, they show relevant variations in terms of genes and interactions, not only between classes, as expected by the different designing principles, but also within the same class. For example, the HTBP class includes interactomes containing a number of genes ranging approximately from 3000 to 11000; a number, this latter, comparable with that of the smallest interactome of integrative-predictive class (FP60). Integrative and integrative-predictive interactomes are comparable in terms of gene number (from 11000 to 19000), but on average integrative-predictive interactomes have a higher link density (from 20 to 50 links per node on average).

Mutual overlaps of nodes and links are shown in figure 1.3. Dot size and color are proportional to interactions and nodes overlap respectively. As expected, integrative interactomes that share interaction sources, i.e. links derived from common databases, have many links and nodes in common, for example CP and IR. On the other side, HTBP interactomes, due to their independent derivation and different experimental techniques, prone to a high rate of false negative, have a small mutual overlap [60].

The portion of the exome covered by all of the interactomes is relatively small and only 1021 genes out of 20630 ($\sim 5\%$) are shared among all of them. Moreover, the similarity of interactome reconstructions on this common core

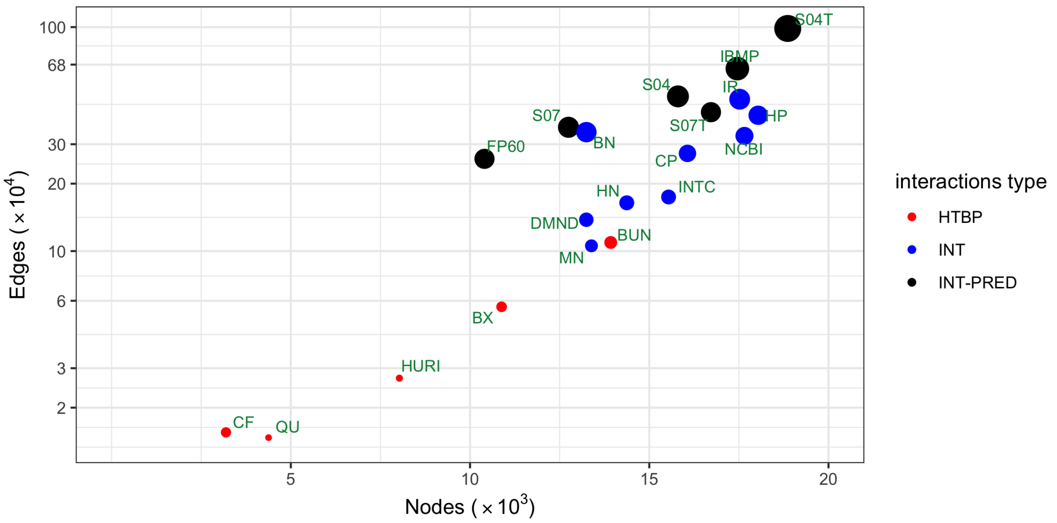


FIGURE 1.2: Number of interactions versus number of genes. Dot size is proportional to network density, i.e. $\frac{n.Link}{n.Nodes}$

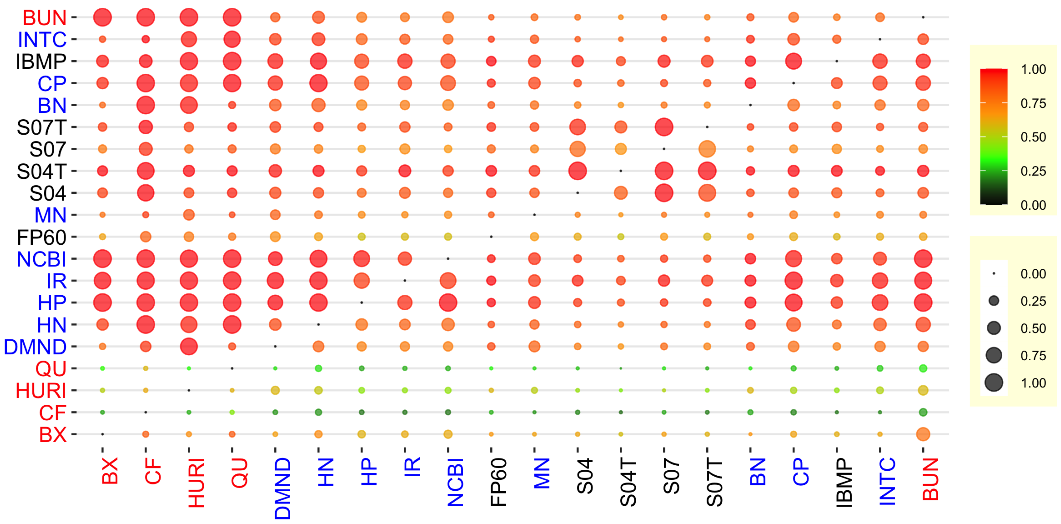


FIGURE 1.3: Mutual link and node overlap for each interactome pair. Dot size (colour) is proportional to the ratio between the number of links (node) shared by two interactomes and the total number of links (nodes) in one of them. Note that the matrix is not symmetric: a column gives how much that interactome is represented by the other while a row gives how much it represents the others.

varies greatly as shown in figure 1.4.

Higher density is associated with lower mean distance and higher clustering, with low density HTBP class on one side and ST04 on the other. Clustering is lower than 0.1 for most of the interactomes with less than 24 links per node on avarage (with the exception of CF and DMND), while for the others covers a wide range (0.2-0.7), see figure 1.5. Since the interactome is

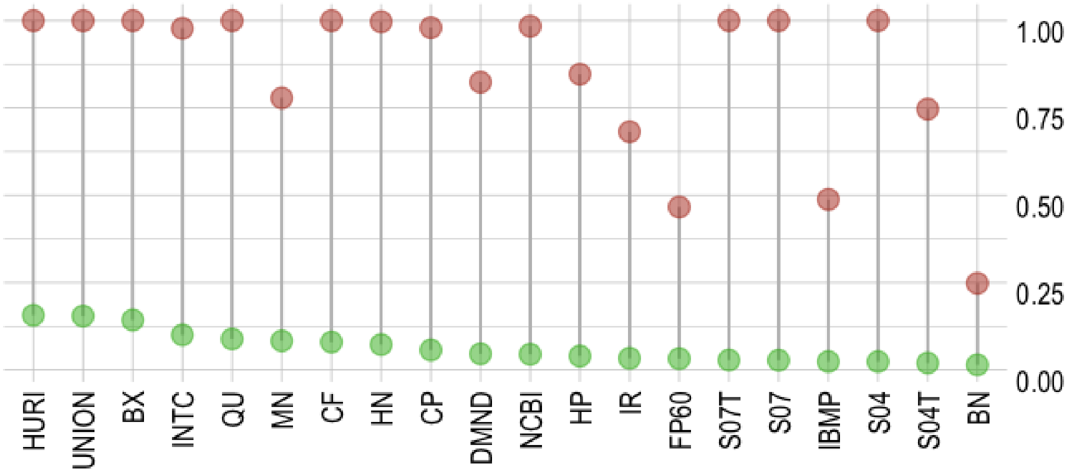


FIGURE 1.4: For each interactome, maximum and minimum fraction of sheared link (with respect to all others) on the common core of 1021 genes.

often claimed to be scale-free, we compared these quantities with those of Barabasi-Albert network model (BA) of the same density finding relevant differences. The mean distances and clustering coefficients of interactomes are always higher than those of BA nets. Moreover, the clustering coefficient of interactomes is highly variable while in BA nets it is almost constant (see figure 1.19 in Appendix).

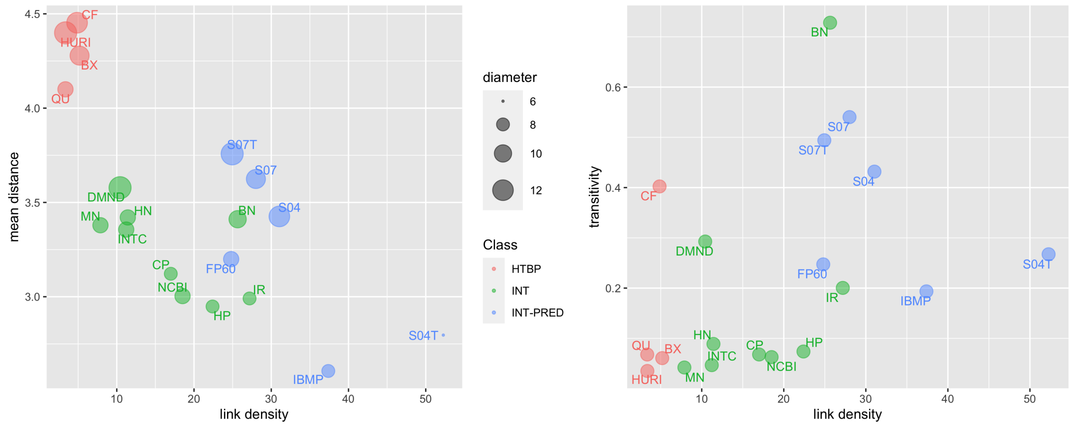


FIGURE 1.5: Mean Distance and mean clustering coefficient as function of link density. In the left panel dot size is proportional to the diameter of the interactome.

1.4 Degree Distribution

A central global property of a network is its degree distribution since it encodes many information about network physical properties.

On one hand, systems taking place on networks with similar degree distribution share common properties, such as diffusion or resilience. On the other hand, in evolving networks, the shape of the degree distribution is often related to the mechanism underlying their evolution. For example, the Barabasi-Albert model predicts a scale-free stationary degree distribution for evolving networks with preferential attachments [9] and many feature of the degree distribution of biophysical PPI interaction networks are reproduced by the duplication-divergence model and its variants [70, 49].

Interactomes and protein-protein interaction networks are often claimed to be scale-free [8]. A network is scale free if its degree distribution $P(k)$ follows a power law:

$$P(k) \propto k^{-\alpha} \tag{1.1}$$

where k is the degree and $\alpha > -1$. The power law distribution has a heavy tail and is invariant under scale transformations so that's not possible to define a 'typical' scale of network degree. In a log-log plot power law is characterized by a linear trend.

Actually, real interactome reconstructions, as the majority of real world networks [18], do not show a power law trend in the whole degree range.

The typical shape of a real interactome is shown in figure 1.6 along with a scale free network with the same number of nodes and links and same a exponent. The linear trend (in log-log scale) is typically observed only for an intermediate range of degree values, while a saturation (i.e. a higher number than expected) is observed for low degrees and a cut-off (i.e. a lower number than expected) is observe for high degree. The degree distributions of the interactome databases considered are reported in figure 1.18 in the appendix of this chapter.

1.4.1 Testing the Scale-Free Hypothesis

Even though interactomes are not genuine scale-free nets, testing if the scale-free hypothesis holds, at least for a sufficiently wide degree range, and if it is consistent across different reconstructions, can shed light on the nature and organization of the web of macromolecular interactions in the cell.

To this end, we fitted the overall degree distribution of each interactome recontruction with a power-law following the method proposed by Clauset et Al. [25] and implemented in the R package *poweRlaw* [37].

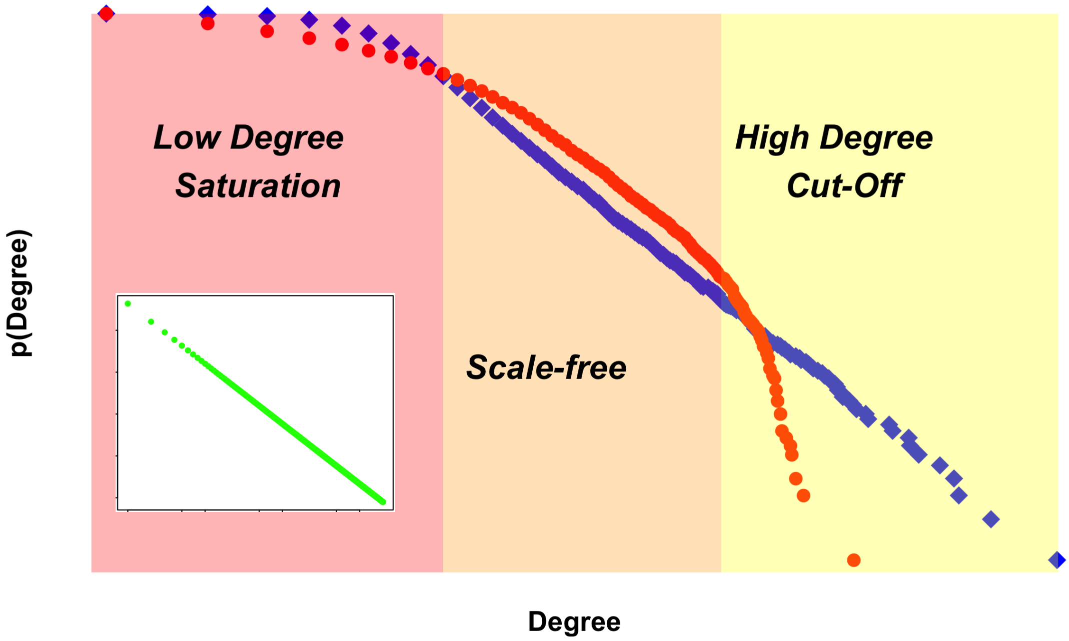


FIGURE 1.6: Typical Degree Distribution of PPI network (red dots) and Degree distribution (blue squares) of a network with the same number of nodes and links and degree sequence extracted from a power law with the same exponent of the PPI (log-log scale). The three regions highlight low degree saturation, scale free behaviour and high degree cut-off range. The inset show a pure power law with a linear trend on the whole degree range

This method is based on the hypothesis that the degree distribution follows a power law for degree greater then some threshold value K_{min} and jointly estimates K_{min} and the power law exponent α. For a discrete power law, the maximum likelihood estimator (MLE) for α, fixed K_{min}, is given by:

$$\hat{\alpha} = 1 + n \left[\sum_{i}^{n} ln \frac{x_i}{K - 0.5} \right]^{-1} \tag{1.2}$$

where x_i are the empirical data points. To estimate K_{min} the MLE for α is computed varying K in the range $[0, max(Degree)]$. K_{min} is the K which minimizes the Kolmogorov-Smirnoff (KS) distance between the data and fitted model cumulative density function. Since increasing K decreases the number of data points available for the fit of α, a minimum number of points in the tail is required for not to discard the power law hypothesis.

Goodness-of-fit was assessed by a semi-parametric bootstrap procedure. A fixed number, N_{bs} of synthetic distributions is generated in the following way:

- for degree lower than the estimated K_{min}, points are bootstrapped from the empirical data,
- for degree higher than the K_{min}, points are sampled from the best fit power law distribution.

Than, for each synthetic distribution fit, the KS statistics is computed. A p-value is defined as the fraction of times the KS of the fit of the synthetic distributions is greater than that for the empirical data fit. Therefore, a high p-value indicates that the power law fits real data as good as synthetic data and cannot be rejected.

Finally, uncertainties on K_{min} and α are estimated sampling (with replacement) from the original data set and re-estimating the parameters.

The scale free hypothesis was tested based on several criteria proposed in a recent comprehensive survey on scale-free networks [18]. Since our interactomes are all simple and undirected networks, we introduced a simplified version of the taxonomy proposed therein. We stratified interactomes in three different levels of plausibility of the scale free distribution hypothesis:

- **None**: interactomes for which the semi-parametric bootstrap has a p-value lower than 0.1, showing that the power law must be rejected.
- **Weak**: interactomes such that power law distribution cannot be rejected, i.e. semi-parametric bootstrap has a p-value grater than 0.1, and such that the fitted tail (data points $x_i > K_{min}$) contains at least 200 nodes.
- **Strong**: Interactomes satisfying weak constraints, such that no other distribution are favoured (i.e., better fits data) on the power law in the same degree range and with a power law exponent α in the range $2 < \alpha < 3$.

The picture resulting from analysis is quite heterogeneous. Figure 1.7 reports the estimated α with against the bootstrap p-value.

For six interactomes, in the INT and IP classes, the power law hypothesis must be rejected, having a zero p-value. All other interactomes show at least weak evidence for scale-freeness, for some degree range, with a p-value grater than 0.1 and a tail which contains more that 200 nodes. To test for strong evidence, power law hypothesis was compared with two alternative distribution: lognormal and exponential.

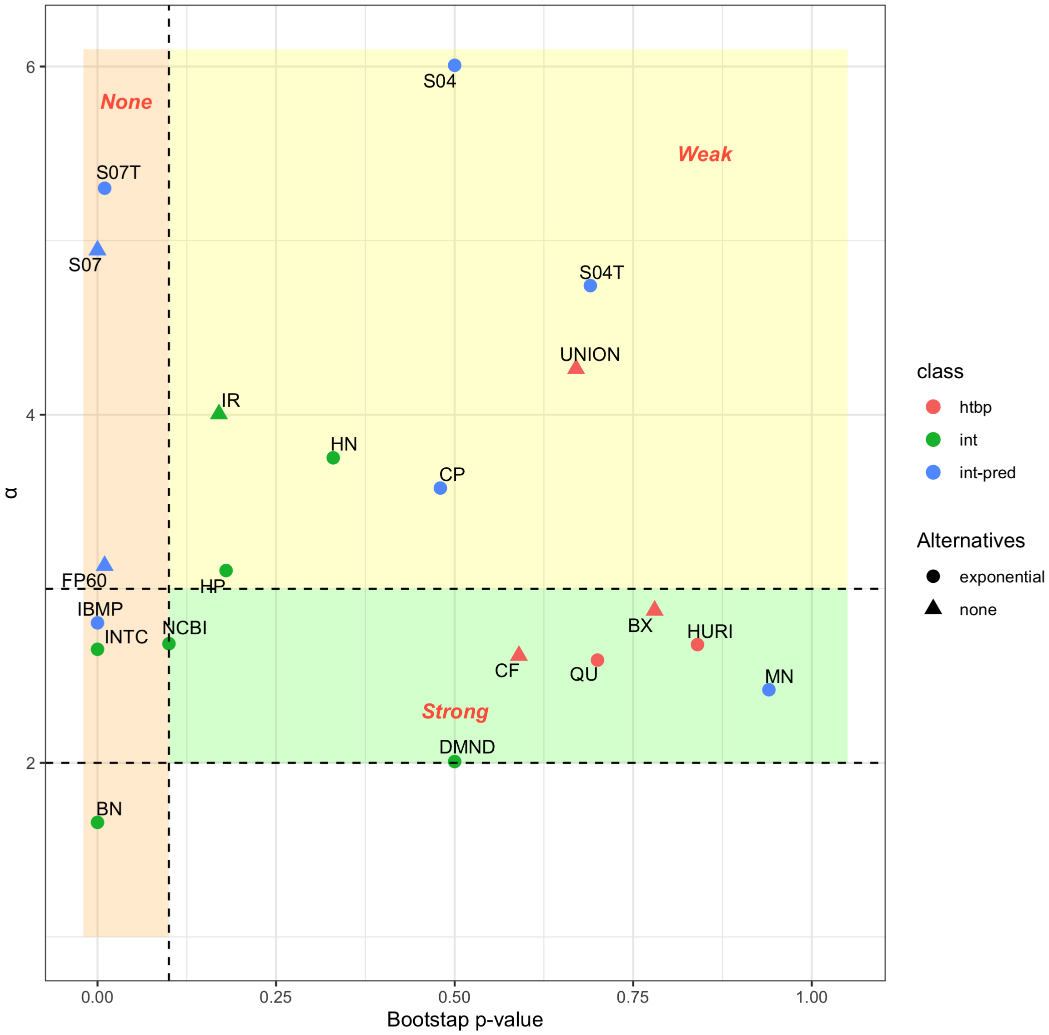


FIGURE 1.7: Evidence for scale-free hypothesis. Power law fitted exponent α vs bootstrap p-value. Colour indicates interactome class. Shape indicates if there exists a distribution which fits data better then power law.

Criteria for strong evidence are satisfied only by CF and BX in the HTBP class. NCBI, DMND (ITC class) and MN (IP) have an exponent in the range $[2, 3]$, consistent (within the error bars) with that of CF and BX (see figure 1.8). However, for this interactomes, power law fit is not significantly favoured on exponential fit. For the union of HTBP and IR the power law is favoured on alternatives and thus the evidence for these interactomes can be considered stronger than for others even though their exponent is not in the range $[2, 3]$.

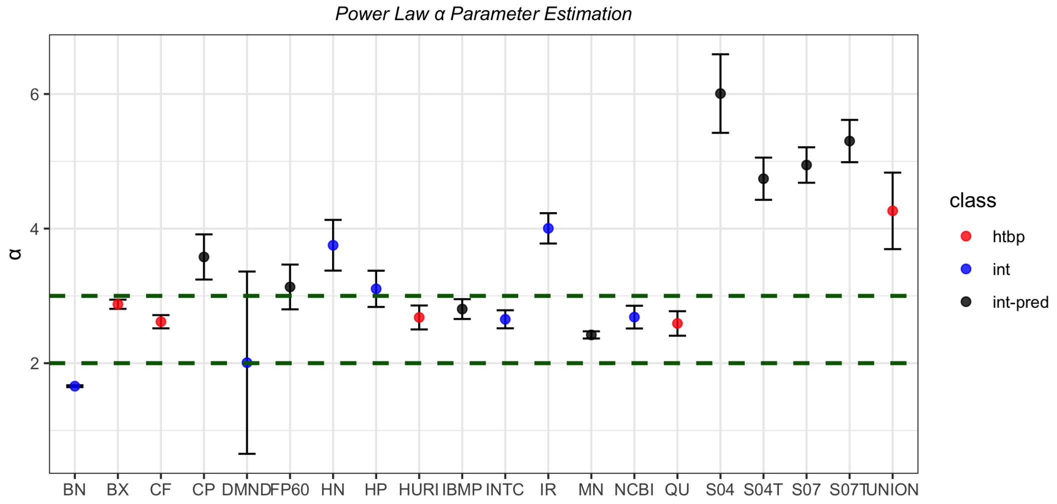


FIGURE 1.8: Estimated value of α exponent for each interactome. Error bars are computed with a bootstrap re-sampling of the empirical distribution, as described in the main text ($N_{rep} = 1000$, $n =$ n.of nodes of the interactome).

1.5 Shared Hubs

Network hubs are nodes with a high centrality. The reason why we are interested in hubs is two folded. First, even though not all interactomes are power law distributed and almost all of them show a high degree cut-off, their distributions has a heavy-tailed nature, so that nodes with high degree are rare but still have a high impact on their global topology [8]. Moreover, the high degree cut-off can be a consequence of limitations in the experimental techniques, as for the case of HTPB class.

Second, hubs play a central role in network-based gene prioritization, in particular they are crucial in network propagation algorithms which have been applied successfully to disease genes and genetic modules identification [27, 14, 56].

In this section, we investigate to which extent hubs are shared among the 20 interactomes reconstructions. In other words, we study how many hubs of a given reconstruction are hubs in the others. Moreover, we study if the number of hubs in the intersection is significant compared to chance.

To this aim, for each reconstruction, we defined hubs as the top 2 percentiles (right tail) of the degree distribution. For each interactome pair, we computed the intersections of hubs and we tested its significance with respect to a hypergeometric distribution null model.

Quantitatively, when considering the the top 2 percentile about 900 genes occur in at least 2 interactome when, 32 when considering at least 12 interactomes and none of the hubs is shared by more than 16 interactomes. The most recurrent hub is the histone deacetylase 1 (HDAC1, $< d >= 360$), which is included in the first 2 percentiles of 16 interactomes and available in all of them; followed by E1A binding protein p200 (EP300, $< d >= 542.5$), BRCA1 DNA repair associated (BRCA1, $< d >= 376$), heat shock protein 90 alpha family class A member 1 (HSP90AA1, $< d >= 433$), tumor protein p53 (TP53, $< d >= 553$) and heat shock protein family A (Hsp70) member 8 (HSPA8, $< d >= 433$), which appear in at least 17. In parenthesis is reported the average hub degree. The fraction of shared hubs is reported in figure 1.9 for each pair.

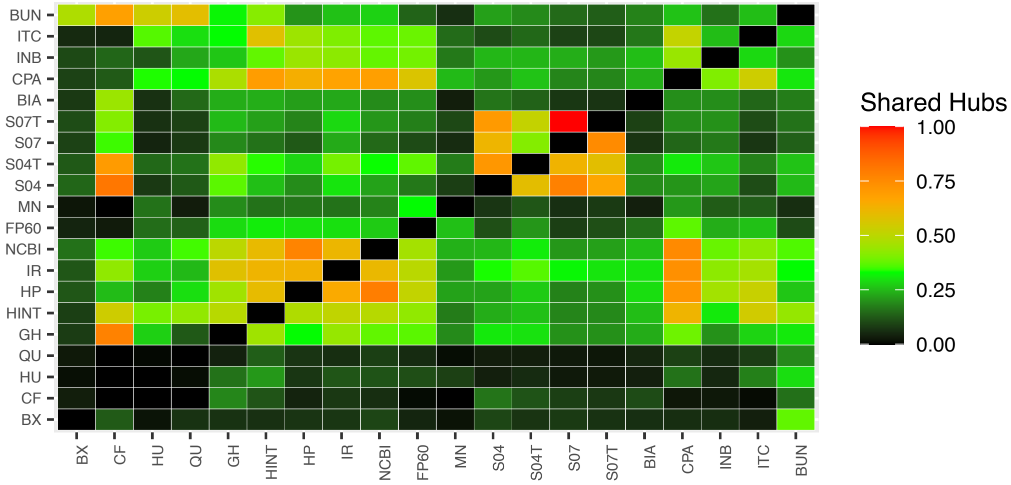


FIGURE 1.9: Fraction of shared hubs (top 2 percentile) for each interactome pair.

1.5.1 Null Model Comparison

Given the number of common hubs between two interactomes, a question is if this number is significative with respect to chance. To test this significance, we developed a simple null model based on the hypergeometric distribution. In particular, given two set of elements (in our case the sets of nodes of two interactomes) with a non empty intersection (the common nodes), the model gives the probability of choosing the same k elements by randomly draw from the two sets.

Hypergeometric Distribution

Given a population of size N made of two classes, where K objects belong to class A and $N-K$ to class B, the *hypergeometric distribution* describes the probability of k successes, *i.e.* draws belonging to class A, in n draws without replacement. The distribution function is given by:

$$p(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{1.3}$$

and has the following properties:

$$< k >= \frac{nK}{N} \tag{1.4}$$

$$Var =< k^2 > - < k >^2 = \frac{n(N-n)K(N-K)}{N^2(N-1)} \tag{1.5}$$

In the limit $N \gg n$ it is well approximated by a binomial distribution.

Networks with Different Sets of Nodes

Given a pair of network, A and B, let V_A and V_B be their sets of nodes with, respectively, N_A and N_B elements. Let N_C be the number of nodes in the common core, $V_C = V_A \cap V_B$.

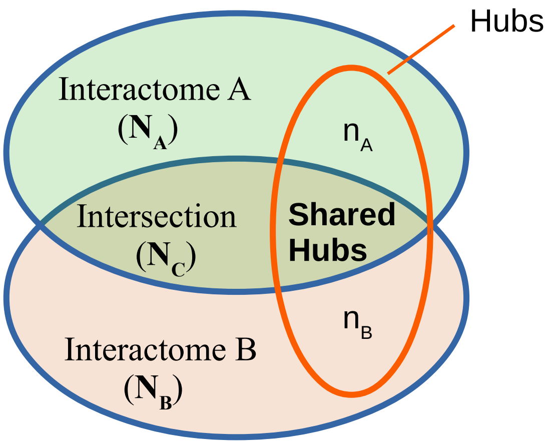


FIGURE 1.10: Illustrative representation of the intersection null model to characterize hubs sharing of interactomes reconstructions with respect to chance.

We are interested in the following question. Choosing randomly n_A nodes from V_A and n_B nodes from V_B, which is the probability, $p(k)$, for the intersection between the two sample to have dimension equal to k?

In other words, if we call the two samples S_A and S_B, we are interesting in finding the probability distribution:

$$p(k) = p(k; N_A, N_B, N_C, n_A, n_B) = p(|S_A \cap S_B| = k) \tag{1.6}$$

where $|\cdot|$ counts the number of elements in a set. We can think to the problem as two independent extractions in the following way:

- in the first extraction we are interested in the probability, $p_A(k_A)$ of finding k_A nodes of V_C in n_A draws from V_A, without replacement. In other terms, we want to find:

  $$p_A(k_A) = p(|S_A \cap V_C| = k_A)$$

  It turns out that p_A is an hyper-geometric with the following parameters :

  $$p_A(k_A) = \frac{\binom{N_C}{k_A}\binom{N_A - N_C}{n_A - k_A}}{\binom{N_A}{n_A}} \tag{1.7}$$

- If we call $I = S_A \cap V_C$, I is a subset of V_B. Then, in the second extraction we are interested in the probability, $p(k|k_A)$, that, in n_B draws from V_B without replacement, k vertex belong to I ($|I| = k_A$). As before, $p(k|k_A)$ is an hyper-geometric distribution:

  $$p(k|k_A) = \frac{\binom{k_A}{k}\binom{N_B - k_A}{n_B - k}}{\binom{N_B}{n_B}} \tag{1.8}$$

$p(k)$ can be written in in following form:

$$p(k) = \sum_{k_A=0}^{n_A} p(k|k_A)p(k_A) \tag{1.9}$$

It is possible to show that (see below):

$$< k > = \frac{N_C n_A n_B}{N_A N_B} \tag{1.10}$$

Networks with the Same Set of Nodes

If network A and network B as the same set of nodes we have:

$$V_A = V_B = V_C \tag{1.11}$$

$$N_A = N_B = N_C = N \tag{1.12}$$

Equation 1.7 do not make sense since V_C and V_A are the same set.

We have to compute the probability of finding k elements of a class with n_A elements (corresponding to the draws from the network A), performing n_B draws without replacement from the set of nodes of network B, which is a set of dimension N. This is given by an hyper-geometric distribution:

$$p(k) = \frac{\binom{n_A}{k}\binom{N-n_A}{n_B-k}}{\binom{N}{n_B}} \tag{1.13}$$

which for the same number of draws from network A and network B, i.e. $n_A = n_B = n$, becomes:

$$p(k) = \frac{\binom{n}{k}\binom{N-n}{n-k}}{\binom{N}{n}} \tag{1.14}$$

with mean value:

$$< k >= \frac{n^2}{N} \tag{1.15}$$

We computed, for each interactome pair, the p-value of the measured intersection with respect to the above model. In figure 1.11 are reported two examples of the null model distribution and the observed intersection. As expected, the significance of the overlap is high for most interactome pairs (p-value $<$ 0.01). However, this is not the case for the HTBP interactome class. In particular, the intersections between CF and HU with other HTBP reconstructions show a high p-value (fig. 1.11, top panel).

1.6 Local Properties Comparison

While global properties concern the network as a whole, local properties encode information at the level of the single node or link. Studying interactome reconstructions, the interest in local properties is two sided. Interactome reconstructions should encode the same underlying network, i.e. the real network of gene product interactions, and the same gene (node) should have similar local properties in different reconstructions. Moreover, network-based omics data analysis exploit local relationship between genes, for example to find disease modules or prioritizing genes. Since results are strongly dependent on the choice of the interactome, a better knowledge of

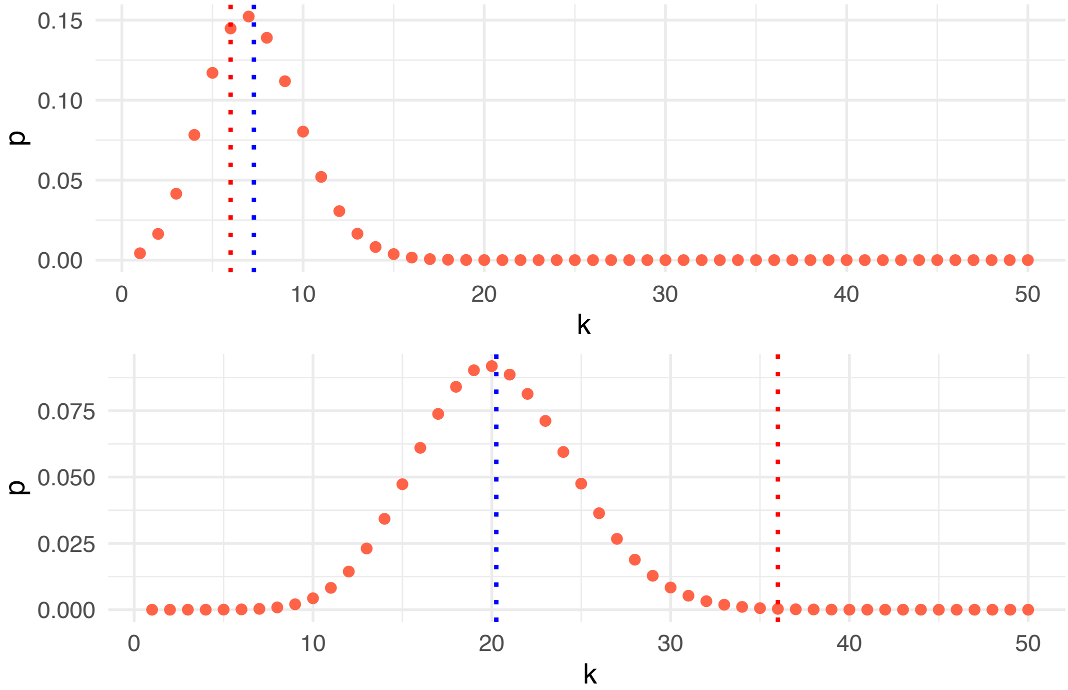


FIGURE 1.11: Examples of the null distributions for two interactome pairs. Blue line is the null model mean, red line is the observed intersection. Top panel shows CF-HU pair (high p-value). Bottom panel shows MN-BX pair (low p-value).

the relationships among local properties of different reconstructions can be a guide to asses network-based studies general validity and reproducibility.

Centrality can be generally defined as the *importance* of a node or a link in the network. Depending on the feature of interest several centrality measures have been defined. We based our analysis on the following four node-centered measures:

- **Degree**: for general undirected graph is the sum of the weights w_{ij} of the k_i links connected to a node i:

$$D(i) = \sum_{j=0}^{k} w_{ij} \tag{1.16}$$

  For unweighed graphs, for which $w_{ij} = 1 \; \forall i, j$, $D(i) = k_i$.

- **Betweenness** [32]: given a connected graph, a path between two nodes is a sequence of links which joins them. A shortest path is a path for which the number of links is minimum. If, sp_{hk} is the number of shortest paths between every node pair (h, k) in the network and $sp_{hk}(i)$

is the number of shortest paths passing through i, the betweenness of node i is defined as:

$$B(i) = \sum_{h \neq k} \frac{N_{hk}(i)}{N_{hk}} \tag{1.17}$$

- **Closeness** [76] for a node i let's call $l_{sp}(i, j)$ the length of the shortest path from node i to j. The closeness of node i is:

$$C(i) = \frac{1}{\sum_j l_{sp}(i, j)} \tag{1.18}$$

- ***k*-Spectral** [71] the network Laplacian is defined as $L = D - A$, where A is the adjacency matrix and D is the diagonal degree matrix ($D_{ii} = \sum_j A_{ij} = k_i$). L encodes key topological (for example, the Fiedler value) and physical information (e.g., it enters in the description of heat transport and diffusive processes on networks).

  The family of *k*-Spectral node centrality aims to measure the impact of the deformation associated to a node, on the laplacian matrix. If A is the adjacency matrix of an undirected network, a deformation of the network with respect to node i is defined as:

$$B^{(i)}_{kl} = \begin{cases} A_{kl} & \text{for}(k, l) \in \{(i, \cdot), (\cdot, i)\} \\ 0 & \text{otherwise} \end{cases} \tag{1.19}$$

  if $\tilde{L}$ is the laplacian of the deformation, the laplacian of the deformed original network is $L + \epsilon \tilde{L}$. The k- spectral centrality is defined as:

$$s_i^k = |\lambda_k'(0)| \tag{1.20}$$

  where $\lambda_k(\epsilon)$ is the k_{th} non null eigenvalue associated to the laplacian of the deformed graph. It turns out that for a node deformation:

$$s_i^k = \sum_{j=1}^{k} A_{ij}(\nu_i(0) - \nu_j(0))^2 \tag{1.21}$$

  where $\boldsymbol{\nu}(0)$ is the k-th eigenvector of L. In the following we will use the 1-spectral centrality.

1.6.1 Results

Local properties comparison was performed on each interactome and on the corresponding subnetwork, defined by the 1021 genes in common to all interactomes and the links among such genes in the considered interactome. In order to test the local similarities, we computed the Spearman's correlation between the aforementioned measures between nodes shared by each couple of interactomes.

We observed a high correlation and a similar distribution of correlation values (medians close to 0.43) for all the centrality measure apart from spectral centrality (figure 1.12-B).

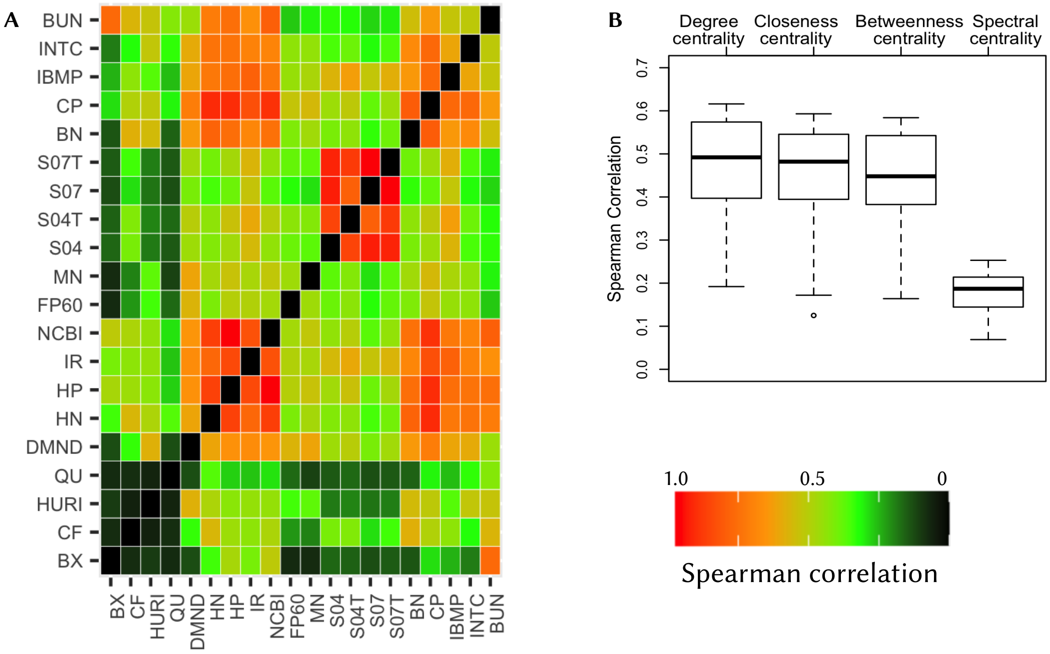


FIGURE 1.12: Left: Spearman Correlation of the Degree for each interactome pair. Other centrality measures show the same correlation pattern. Right: Boxplot of centralities correlation values.

The correlation analysis, figure 1.12(A), revealed that the four variants of STRING form a group on their own. On one hand, they show a high similarity among themselves, meaning that including text-mining derived interactions and varying confidence score did not affect significantly the local structure of the network: gene ranking by centrality is similar even if the links in S04 are twice as many than S07, an interesting observation since they have different global properties and degree distributions. On the other hand, they are much less correlated with other interactomes meaning that their local topology is different, even from interactomes of comparable size and density which have a high overlap with them (e.g. FP60). Another group comprises INT interactomes (with the exception of MN) and IBMP. This result reflects the fact that many of them share the same interaction sources.

A third group comprises the HTBP interactomes. Despite their global similarities, they show very different intra-class centrality profiles probably reflecting the different experimental techniques used and the high number of false negative of this techniques, as suggested by [60]. They are also poorly correlated with interactomes of other classes, even with those interactomes that use them as interaction sources. However, the correlation of their union, show a higher level of correlation with IBPs. FP60 and MN (integrative-predictive and integrative) are less correlated (~ 0.5), similarly to HTBP, even when they show a significant overlap with other interactomes (for examples, MN with IR).

A similar picture is obtained when we compute centrality considering only the 1021 genes shared by all 20 interactomes..

1.7 Network Resilience To Cancer Mutations

As mentioned in the introdution, cancer phenotypes are the outcome of a process of accumulation of genetic alterations by which some cells acquire a selective advantage on normal cells and spread across the healthy tissue. In the last decade, large scale genome wide studies (e.g, the TCGA project) have identified an increasingly large number of genes associated to cancer, shading light on its genetic basis.

Cancer is not the outcome of alterations of one or few specific genes. It is a the perturbation of cell state resulting from the interplay of dysfunctional molecular constituents [46]. The same cancer can originates from completely different sets of mutations and a wild heterogeneity of mutational landscapes its observed across patients and cancer types [20, 83].

Such variability hampers the efforts to link cancer genotype to phenotype. It is thus crucial to uncover interdependencies among gene alterations, i.e. how they 'collaborate' to give rise to cancer. A first step in this direction is to study regularities of disease genes location on biological networks as, for example, their proximity to hubs or their tendency to form disease modules [8, 43, 21].

Recent advances in experimental techniques [60] to uncover interactions among cell constituents, made available a number of different maps of biological interactions, among them the most studied in the interactome.

As mantioned above, the human interactome is a network where each node represents a gene and an edge represents some kind of interaction between two genes or their products. In this context, a mutation can be seen as the removal of a node (e.g. a misfolded protein) or in the disruption of an interaction. In other words, a mutation is a network failure which affects the flow of biological information in the cell [50, 92].

Resilience, or fault tolerance, quantify to which extent a network changes when one or more nodes (or links) are removed. The lower the change the higher the resilience.

In the context of biological networks, several resilience measures have recently been proposed. While some of them focus on overall resilience to sebsequent random nodes and links removals [53, 92] others try to asses the impact of specific perturbations on the network information flow [12].

In this section, we integrate information from several interactomes reconstructions with cancer mutational datasets to study the impact on the interactome of cancer mutations. Moreover, since interactomes reconstructions differs in terms of size and types of the interactions included, we compare the 20 databases introduced above characterizing their resilience to random nodes removal with respect to other network models.

1.7.1 Results

Networks are an efficient way to characterize the overall architecture of interactions among cell molecular constituents giving a picture of their global interdependencies. Studying to which extent the topology of a biological network is affected by failures, i.e. the removal of nodes or links, allows to advance hypothesis on the loss of biological functions and the increase of the risk of diseases.

We studied the resilience of the interactome networks under successive node deletions by means of a resilience measure which sizes the level of network fragmentation when an increasing fraction f of its nodes is removed. Fixed f, network fragmentation was quantified by a modified Shannon Diversity:

$$S'(f) = -\frac{1}{log(N)} \sum_{i=1}^{c} [p_i log\,(p_i) - f] \tag{1.22}$$

where c is the number of disconnected components and p_i their relative size. The overall network resilience is defined as:

$$R = 1 - 2\int_0^1 S'(f) df \tag{1.23}$$

and takes values in $[0, 1]$. The higher R the more stable the network.

In figure 1.13 is reported an illustrative example of the link between network fragmentation, S' and R (for details on the definition of S' and R see Appendix 1.8).

We studied the overall resilience of the selected 20 reconstructions of the human interactome under two different failure schemes:

- *Random Failures*: nodes to remove where chosen randomly with uniform distribution.
- *Targeted Failures*: nodes with a specific property where chosen with higher probability.

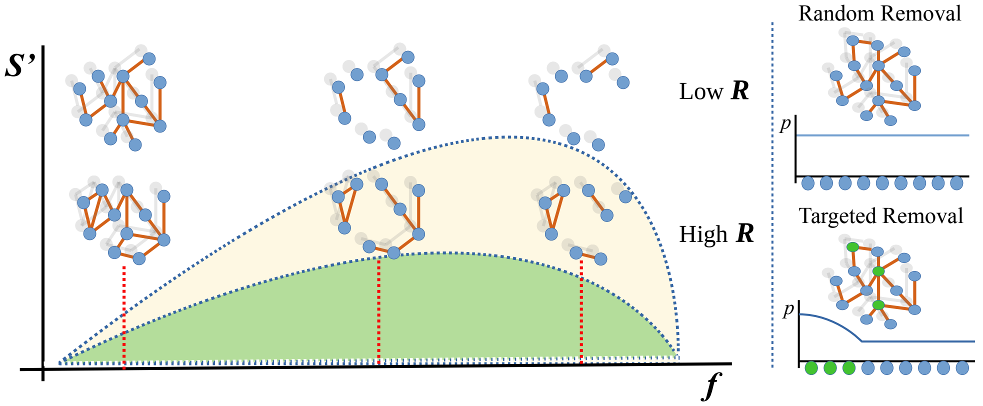


FIGURE 1.13: Left: Illustrative example of modified Shannon diversity S' as a function of the fraction of removed nodes f. The higher the network fragmentation the higher S'. R is lower when the integral of S' in $[0,1]$ in greater. Right: Node removal schemes. In random removal nodes are chosen with a uniform probability while in targeted removal, 'target' nodes are removed with higher probability.

1.7.2 Interactome Resilience to Random Nodes Failure

We compared interactome reconstructions resilience under random nodes removal. In figure 1.14 is reported resilience as a function of link density. As expected more dense networks are, in general, more resilient.

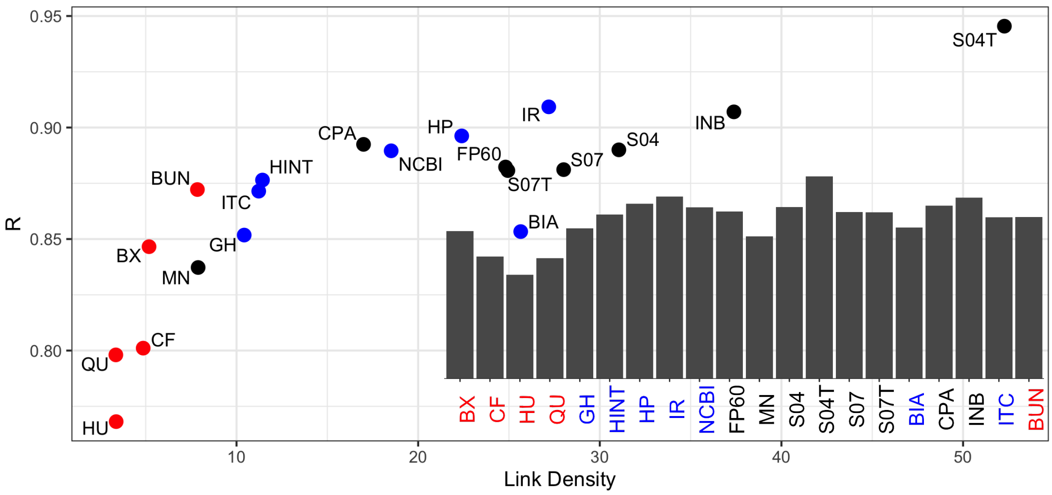


FIGURE 1.14: Interactome resilience as a function of link density. Colour indicates interactome class. Red = HTBP, blue = INT, black= ITP

We compared interactome resilience with respect to regular lattices (RL), Erdos-Ranyi (ER) and Barabasi-Albert (BA) network. Since $S'(f)$ depends on link density we compared each reconstruction to the above network models with the same density. Figure 1.15 show the typical trend of $S'(f)$ for interactome reconstructions and the related network models. Fixed f, $S'(f)$ is minimum when only one connected component in present. Continuous line in figure 1.15 is the theoretical $S'(f)$ for a network of complete graph.

We observe that, while regular and fully random graph start to disaggregate when more than half of their nodes is removed, interactomes start to break up much earlier, after the removal of few random nodes indicating a higher brittleness.

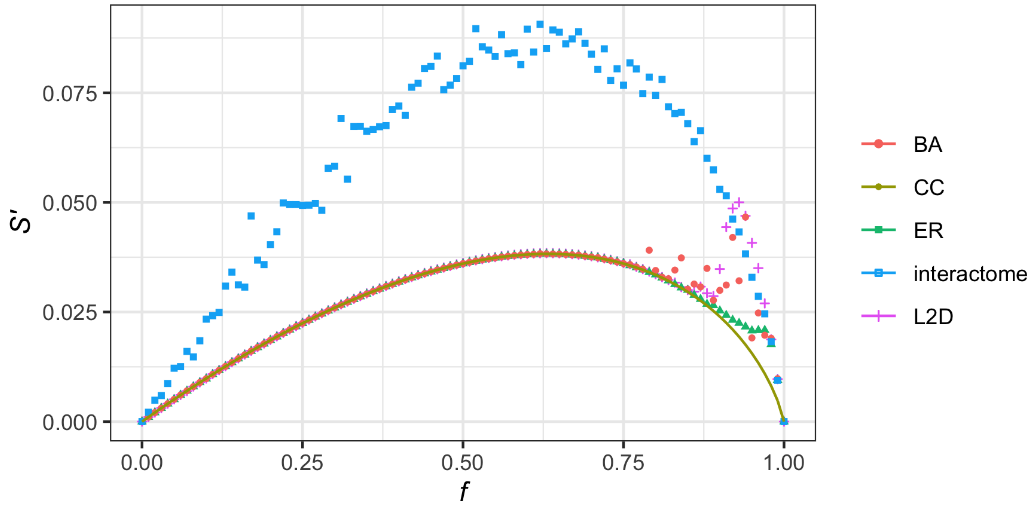


FIGURE 1.15: Comparison of Shannon diversity as a function of the fraction of removed nodes f of 3 different network models with the same link density of Bioplex interactome. Dotted curves: 2D lattice, Erdos-Renyi Graph, Barabasi-Albert Preferential Attachment, Interactome. Continuous line: theoretical Shannon diversity for one component network.

1.7.3 Impact of Cancer Mutations on Interactome Resilience

Having characterized the resilience of interactome recontructions to random nodes failures, we were interested in the impact of cancer related genes removal on the interactome topology, compared to random genes. We set up a removal scheme where most mutated gene associated to a given cancer were chosen with higher probability. If cancer mutations are not randomly distributed on the interactome, one would expect a change in the global resilience.

From the TCGA database (`https://tcga-data.nci.nih.gov`) we collected mutational datasets of three cancer types: Acute Myeloid Lekuemia (LAML), Prostate Adenocarcinoma (PRAD) and Lung Adenocarcinoma (LUAD). Since cancer mutational landscape is heterogeneous and the number of mutations per patients vary wildly among tumours, we chose these tumours types to cover the whole spectrum of mutational frequencies (see Table 1.2). For each tumour, we removed highly mutated samples (outliers) and we retained only genes present in at least one interactome.

TABLE 1.2: TCGA Cancer Mutational Datasets

	LAML	**PRAD**	**LUAD**
n. of patients	136	484	567
n. of mutated genes	1760	9146	18675
median n. of mutations	14	33	232

For each tumor we computed the frequency of mutation for each gene and assigned a probability of being removed proportional to its mutation frequency.

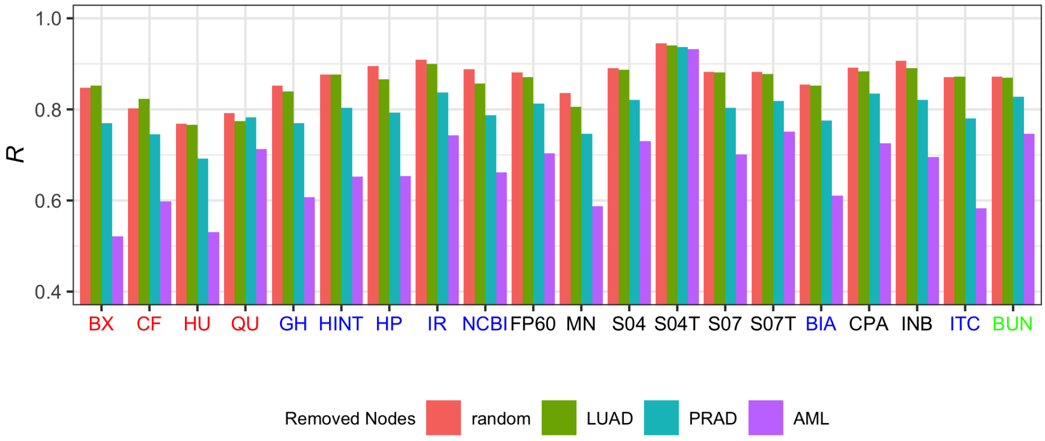


FIGURE 1.16: Resilience of each interactome recostraction under random and cancer related genes removal.

For all interactome reconstructions we observed a lower resilience to cancer mutations even thought with relevant variations among them and depending on the cancer type (figure 1.16). This result is in line with the 'local hypothesis' which states that mutated genes involved in the same disease tend to be neighbours on the interactome. Moreover, highly mutated cancer genes tends to be node with high degree with grater impact on the global network topology. Interestingly, the only exception to the overall picture is STRING with a 0.4 confidence score and with text mining derived interactions.

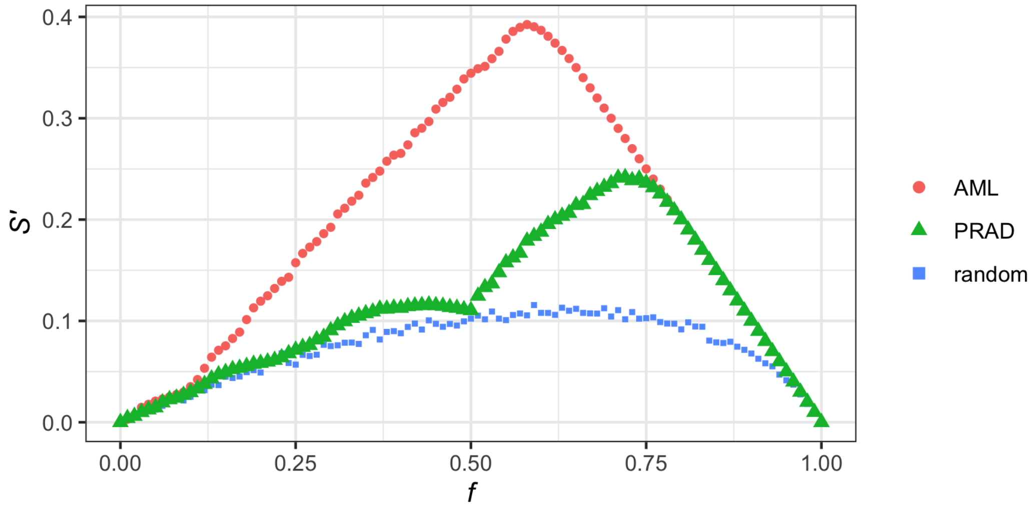


FIGURE 1.17: $S'(f)$ under random and cancer related genes removal for DMND interctome.

1.8 Conclusions

Currently available models of the human interactomes are incomplete. Given the increasing importance of network-based analyses of omics datasets, we compared 20 interactomes, including the three main types: high-throughput biophysical, integrative and integrative-predictive. We gave a picture of topological properties which revealed a relevant structural heterogeneity among the interactomes under study. Such heterogeneity goes beyond interactome size (number of genes and interactions) or density, and involves degree distribution shape and clustering coefficient.

Some interactomes showed a strong evidence in favour of being scale-free networks, while for others such definition is questionable. This is not surprising, since the debate about the “scale-freeness” of networks that model the interactome is still open.We found a significant overlap of hubs among interactomes when compared to random a random null model. However, considering that the studied interactomes can be seen as models of the same underlying reality, the observed overlap might be considered not satisfactory in some cases and indicates some relevant discrepancies on genes that play the role of hubs. The observation that the most shared hubs tend to have higher degrees might reflect a correlation between the amount of available evidences supporting the interactions and the relevance of a gene in one or more diseases.

Centrality measures revealed three groups of interactomes. The four versions of STRING keep a high similarity among themselves, despite the differences in interaction type and confidence. This similarity reflects a specificity in local properties in comparison to other interactomes of similar type and size (i.e., FP60 and IBW). The majority of integrative interactomes (MN excluded) and IBW forms a second group, which very likely reflects the use of common sources of interactions. Third, the specificity of the four HTBP interactomes reflects the different experimental approaches used to detect the interactions.

Despite their wide differences, interactome reconstructions show common patterns of resilience under both random node and mutated node removal. With respect to the other network model considered, all reconstructions, show higher brittleness indicating their higher modularity. On the other hand, the much lower resilience to cancer mutations highlights their not random distribution on the network supporting the hypotheses that cancer mutations impacts neighbouring nodes with higher probability.

Appendix

Degree Distribution of Interactomes

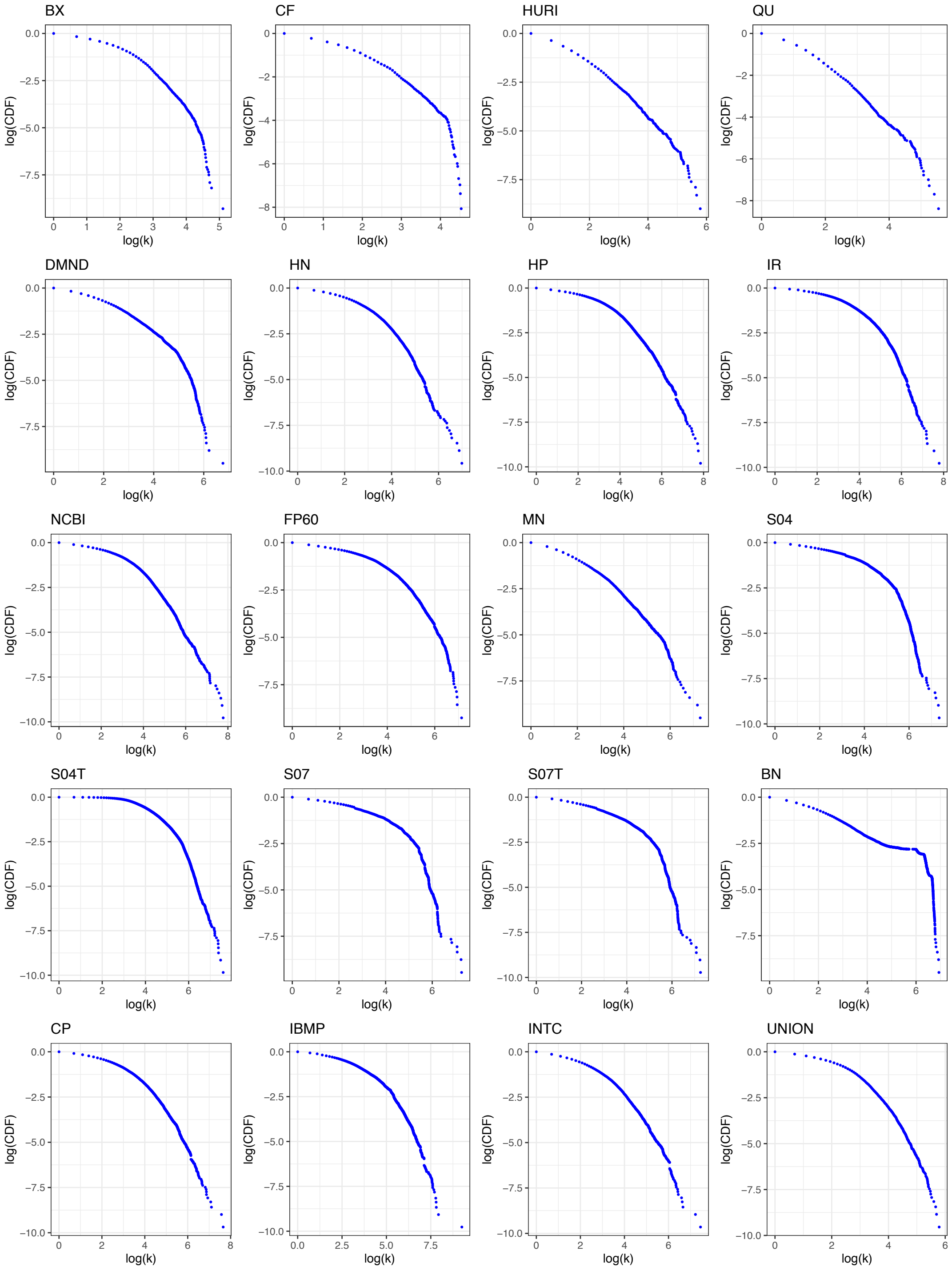


FIGURE 1.18: Interactome reconstructions degree distribution

BA properties comparison

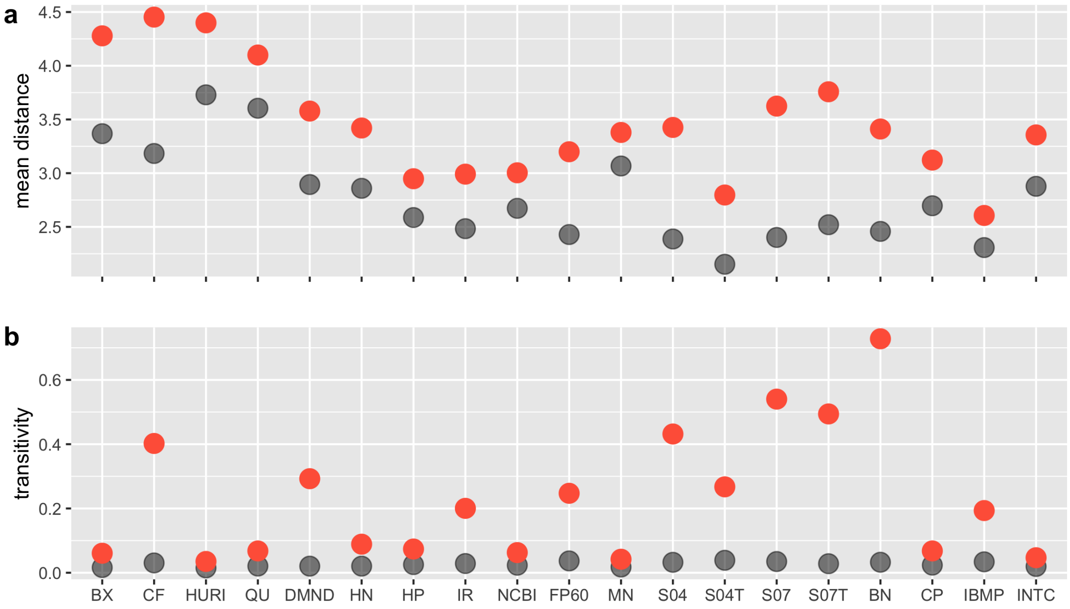


FIGURE 1.19: Comparison of mean distance and global clustering coefficient of interactome reconstructions and a Barbasi-Albert scale free network with the same number of nodes and link density (red dots are interactomes).

Calculation of Null Model Mean

Let us recall the following property of the binomial coefficient:

$$k\binom{K}{k} = K\binom{K-1}{k-1} \tag{1.24}$$

The mean of the distribution p(k) is given by:

$$< k >= \sum_{k=0}^{n_B} p(k)k = \sum_{k=0}^{n_B}\sum_{k_A=0}^{n_A} \frac{\binom{N_C}{k_A}\binom{N_A-N_C}{n_A-k_A}}{\binom{N_A}{n_A}} \frac{\binom{k_A}{k}\binom{N_B-k_A}{N_B-k}}{\binom{N_B}{n_B}} \tag{1.25}$$

Using two times relation 1.24, we have:

$$< k >= \sum_{k=0}^{n_B}\sum_{k_A=0}^{n_A} N_C \frac{\binom{N_C-1}{k_A-1}\binom{N_A-N_C}{n_A-k_A}}{\binom{N_A}{n_A}} \frac{\binom{k_A-1}{k-1}\binom{N_B-k_A}{N_B-k}}{\binom{N_B}{n_B}} \tag{1.26}$$

Applying relation **??** to both binomial coefficients in the denominator:

$$\begin{gathered}
< k >= \sum_{k=0}^{n_B} \sum_{k_A=0}^{n_A} N_C \frac{\binom{N_C-1}{k_A-1}\binom{N_A-N_C}{n_A-k_A}}{\frac{N_A}{n_A}\binom{N_A-1}{n_A-1}} \frac{\binom{k_A-1}{k-1}\binom{N_B-k_A}{N_B-k}}{\frac{N_B}{n_B}\binom{N_B-1}{n_B-1}} = \\
= \frac{N_C n_A n_B}{N_A N_B} \sum_{k=0}^{n_B} \sum_{k_A=0}^{n_A} \frac{\binom{N_C-1}{k_A-1}\binom{(N_A-1)-(N_C-1)}{(n_A-1)-(k_A-1)}}{\binom{N_A-1}{n_A-1}} \frac{\binom{k_A-1}{k-1}\binom{(N_B-1)-(k_A-1)}{(N_B-1)-(k-1)}}{\binom{N_B-1}{n_B-1}}
\end{gathered} \tag{1.27}$$

Noting that the summation is of the form:

$$\sum_{k_A=0}^{n_A} \sum_{k_B=0}^{n_B} p(k_B|k_A)p(k_A) = 1 \tag{1.28}$$

we have:

$$< k >= \frac{N_C n_A n_B}{N_A N_B} \tag{1.29}$$

Global Network Resilience

Global Network Resilience was defined following [92]. Let's call I a network of N nodes. If we remove a fraction f of nodes from I, the network is fragmented in a set of c components of different sizes. Let's call s_i, with $i = 1, \ldots, c$, the number of nodes in the i-th component.

The *Shannon Diversity* for the resulting components set is defined as:

$$S(I_f) = -\frac{1}{log(N)} \sum_{i}^{c} p_i log p_i \tag{1.30}$$

where $p_i = \frac{s_i}{N}$ and I_f denotes the set of componets originating from I when a fraction of nodes f is removed. The factor $\frac{1}{log(N)}$ is introduced to allows the comparison of interactomes of different size [92].

In this definition, each removed node is a component of size $s = 1$ and the corresponding p is $p_{(1)} = 1/N$, so that the contribution to S of the $n = Nf$ removed nodes is given by:

$$S = N\left[-\frac{1}{log(N)} \cdot \frac{1}{N} log\left(\frac{1}{N}\right)\right] f = f \tag{1.31}$$

since the contribution to S is fixed and equal to f, we can define a shifted Shannon Diversity S' as:

$$S'(I_f) = S(I_f) - f \tag{1.32}$$

In a connected networx, S' has minimum, $S' = 0$, at $f = 0$ and $f = 1$. We define the overall network resilience $R(I)$ as:

$$R(I) = 1 - 2\int_0^1 S'(f)df \tag{1.33}$$

so that R takes values in the range $[0, 1]$.

Maximum Resilience

For a network I of N nodes, after removal of $n = Nf$ nodes, the minimum possible shifted Shannon Diversity, S', is obtained when only a giant component of dimension $(N - n)$ is present, in this case, the summation in 1.30 has only one term and $S'(I_f)$ is given by:

$$S'(I^1_{f=\frac{n}{N}}) = -\frac{1}{log(N)}\left[\frac{N-n}{N} log\left(\frac{N-n}{N}\right)\right] \tag{1.34}$$

Substituting $n = Nf$ we have:

$$S(I_1) = -\frac{1}{log(N)}(1-f)log(1-f) \tag{1.35}$$

In figure 1.20 is reported $S(I^1)$ varying N. Fixed f, $S(I_1) \to 0$ for $N \to \infty$.

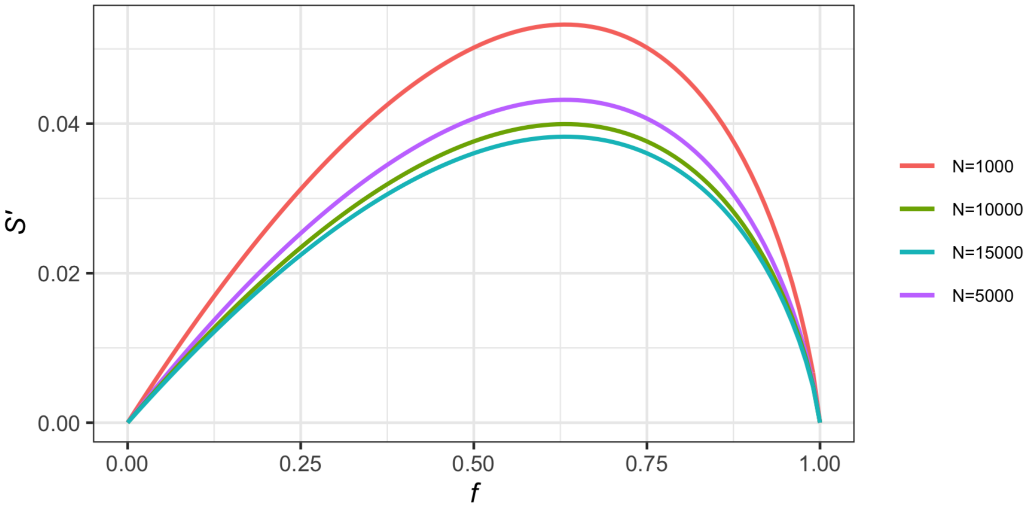


FIGURE 1.20: Minimum Shannon Diversity varying f for different values of N. $N = 1000$ (top line), 5000, 10000, 15000 (bottom line)

Chapter 2

Statistical Methods for Cancer Genome Landscaping

Introduction

Cancer is an evolutionary process where subsequent mutations confer to cancer cells a selective advantage on the others, causing the spreading of the disease in the tissues. Despite its genetic roots, cancer is not the direct outcome of specific gene mutations. It is a complex disease, where different patterns of mutations lead to altered cell functions which, in turn, results in cancer phenotypes [89].

For this reason, tumours show a high mutational heterogeneity and rarely two different patients have the same mutational profile [83].

Mutational heterogeneity is at the basis of different responses to treatment and different survival of patients. In the growing field of personalized medicine [7], finely characterize cancers genomics, stratifying patients based on their genotype and uncovering gene interactions [65] are central tasks to better tailor medical treatments.

To this aim, the choice of clustering method has a key role since the model encodes our hypothesis on the hidden structure structure underlying data.

In this chapter, we introduce bayesian methods for cancer genome landscaping.

We will focus on nonparametric models based on the Dirichlet Process. The advantage of using a nonparametric approach to clustering is that it allows not to specify a priori the unknown number of clusters letting the model to adapt its complexity to data.

We will start by introducing finite mixture models and their nonparametric generalization, the *Dirichlet Process* mixture model. We will then introduce Latent Dirichlet Allocation (LDA), a parametric model originally introduced with the aim of classifying text documents based on their content [16].

LDA generalizes mixtures allowing to model more complex data structures where clusters of features can be shared among samples. As we will see this flexibility is particularly useful in modelling cancer genotypes where we expect some genes categories to take part in different proportions in each genotype. Since LDA is a parametric model, still requiring to fix a priori the number of clusters categories, we introduce its nonparametric extension, the Hierachical Dirichlet process (HDP).

In the second part of the chapter, we apply Hierachical Dirichlet process to clustering of cancer genotypes characterizing a cohort of almost 2000 Myelodispastic syndrome patients.

2.1 Mixture Models

For completeness let's start by recall some basic notion of bayesian statistics.

In bayesian modelling one have to specify a model $m(\boldsymbol{\theta})$ for the data, with a number of free parameters $\boldsymbol{\theta}$ which in general is a vector. If we call D' the observable data, then:

$$P(D'|m, \boldsymbol{\theta}) \tag{2.1}$$

is the *likelihood*, the probability of the data given the model and the parameters.

Since in bayesian statistics parameters are random variable, the model $m(\boldsymbol{\theta})$ is fully specified once a *prior* distribution over the range of parameters, $P(\boldsymbol{\theta})$, is defined. In other words, a fully specified model defines a *joint probability distribution* over the observable data and parameters $P(D', \boldsymbol{\theta}|m) = P(D'|m, \boldsymbol{\theta})P(\boldsymbol{\theta}|m)$.

Once fixed the model we would know how observation data D modify our initial hypothesis on the distribution on parameters space, i.e. how the prior is modified by observations. This is done by computing the *posterior distribution*:

$$P(\boldsymbol{\theta}|D) = \frac{P(D|m, \boldsymbol{\theta})P(\boldsymbol{\theta})}{P(D)} \tag{2.2}$$

A fully specified bayesian model, is a *generative process*. It reflects our hypothesis on how date arose, specifying the joint probability of observed and hidden variable. Computing the posterior allows to 'revert' the process, i.e. allows infer the distributions of hidden variable that likely generated the observed data. Before introducing the finite mixture model as a generative model for the data, we introduce the *Beta* and the *Dirichlet* distributions since their are central in what follows.

Beta and Dirichlet Distributions

The *Beta distribution* is a continuous univariate probability distribution defined on the interval $[0,1]$:

$$Beta(\pi;\alpha,\beta) = \frac{\pi^{\alpha-1}(1-\pi)^{\beta-1}}{B(\alpha,\beta)} \tag{2.3}$$

where, α and β are two shape parameters, and:

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{2.4}$$

The multivariate generalization of the Beta distribution to K dimensions, is the *Dirichlet distribution*, $Dir(\boldsymbol{\pi};\boldsymbol{\alpha})$:

$$Dir(\boldsymbol{\pi};\boldsymbol{\alpha}) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \alpha_i} \prod_{i=1}^{K} \pi_i^{\alpha_i - 1} \tag{2.5}$$

where $\boldsymbol{\alpha}$ is a K dimensional vector of shape parameters, α_i, and $E(\pi_i) = \frac{\alpha_i}{\sum_i \alpha_i}$.

$Dir(\boldsymbol{\pi};\boldsymbol{\alpha})$ is defined on the K dimensional simplex, Δ_K:

$$\Delta_K = \{(\pi_1,\pi_2,\dots,\pi_K) : \pi_i \geq 0; \sum_i \pi_i = 1\} \tag{2.6}$$

We note that, intuitively, a draw from a $Dir(\boldsymbol{\pi};\boldsymbol{\alpha})$ of dimension K can be seen as breaking a stick, of length one, in K parts of random lengths, α_i being the expectation value of the length of part i. The *Beta* is the special case for $K = 2$.

The *Beta* and the *Dirichlet* Distribution will have a central role in what follows since they are the conjugate priors of the *Binomial* and of *Multinomial or Categorical* distributions, respectively. In particular, given the observations, $y_i \in D$, the Dirichlet posterior of a multinomial with Dirichlet prior, $Dir(\boldsymbol{\alpha})$, is a Dirichlet distribution with parameters:

$$\alpha'_j = \alpha_j + \sum_{y_i \in D} y_{ij} \tag{2.7}$$

Moreover, the *Beta* is at the basis of the stick-breaking construction of the Dirichlet Process.

2.1.1 Finite Mixture Model

A finite mixture model assumes assumes that observed data belong to a fixed number, K, of different unobserved clusters and, given a cluster, observations follow a cluster specific distribution, $F(x_i|\boldsymbol{\theta}_j)$ for $j \in \{1, \ldots, K\}$. The probability density has the form:

$$p(x_i|\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{j=1}^{K} \pi_j F(x_i|\boldsymbol{\theta}_j) \tag{2.8}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ and π_j is the weight of cluster j and $\boldsymbol{\theta}_j$ is the vector of parameters of F.

This intuition can be modelled as a generative process.

To fully specify the model we have to put priors on the parameters: the mixture weights $\boldsymbol{\pi}$ and the $\boldsymbol{\theta}_j$. The natural choice for the prior on $\boldsymbol{\pi}$ is the Dirichlet distribution since it is defined on the K-simplex (i.e., $\sum_j \pi_j = 1$) and is the conjugate prior of the categorical distribution. The choice of the prior on $\boldsymbol{\theta}$, H, depends on the specific form of $F(x_i|\boldsymbol{\theta}_j)$.

The complete generative model is defined by the following steps:

- a vector of weights, one for each cluster, is generated randomly according to a Dirichlet distribution with hyperparameter $\boldsymbol{\alpha}$:

$$\boldsymbol{\pi} \sim Dir(\boldsymbol{\alpha}) \tag{2.9}$$

- parameters for each cluster, j, are generated randomly:

$$\boldsymbol{\theta}_j \sim H \tag{2.10}$$

- then, for each observation, x_i:
  - a cluster assignment variable, $z_i \in \{1, \ldots, K\}$, is drawn:

$$z_i \sim Cat(\boldsymbol{\pi}) \tag{2.11}$$

  $Cat(\boldsymbol{\pi})$ is the categorical distribution, i.e. a multinomial, $Mult(\boldsymbol{\pi}, n)$, with $n = 1$.

  - x_i, is drawn from the cluster specific distribution:

$$x_i|z_i \sim F(\boldsymbol{\theta}_{z_i}) \tag{2.12}$$

We note that the first two steps, where $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_j$ for $j \in \{1, \ldots, K\}$ are chosen randomly from their priors, amount to draw a random discrete distribution on the parameter space of the form:

$$G^{(k)} = \sum_{j=1}^{K} \pi_j \delta_{\boldsymbol{\theta}_j} \tag{2.13}$$

where:

$$\delta_{\boldsymbol{\theta}_j}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x} = \boldsymbol{\theta_j} \\ 0 & \text{otherwise} \end{cases}$$

$\delta_{\boldsymbol{\theta}_j}(\boldsymbol{x})$ is called a *probability atom*.

$G^{(k)}$ places K atoms at points $\boldsymbol{\theta}_j$ in the space of the parameters ($\boldsymbol{\theta}_j \sim H$) and associate a mass probabilities, π_j to each of them ($\boldsymbol{\pi} \sim Dir(\boldsymbol{\alpha})$).

In figure 2.1 is reported an illustrative example for G with $K = 3$ and $\boldsymbol{\theta}_j \in \mathbb{R}^2$. This formulation will be useful in the following when introducing the Dirichlet process.

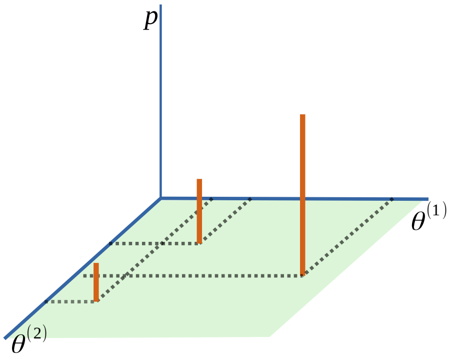


FIGURE 2.1: Random discrete probability distribution generated by randomly choose the vector of cluster weights, $\boldsymbol{\pi}$ (orange sticks length) and the parameters of the cluster specific distributions, $\boldsymbol{\theta}_j$ (locations of atoms is $\mathbb{R}^2$).

Finite Multinomial Mixture for Cancer Genotypes Modelling

As a simple example and to introduce some notation useful in what follows, let's introduce an application of finite mixture model to cluster cancer genotypes.

Suppose we have observations of mutational status of m different genes for n cancer patients. The genotype of patient i is the vector $\mathbf{g}^i = (g_1^i, \ldots, g_m^i)$, where $g_v^i = t$ means that the gene v is mutated t times in sample i. We want to cluster patients based on their pattern of mutations. To to this end, we have to make hypothesis on the structure underlying data.

A finite mixture of multinomials assumes that K different '*genomic*' classes, or gene groups, exist, each being characterized by different probabilities of mutation of the m genes. Moreover, they assume that each patient belong to only one class. An illustrative example of the mixture model data structure is reported in figure 2.2.

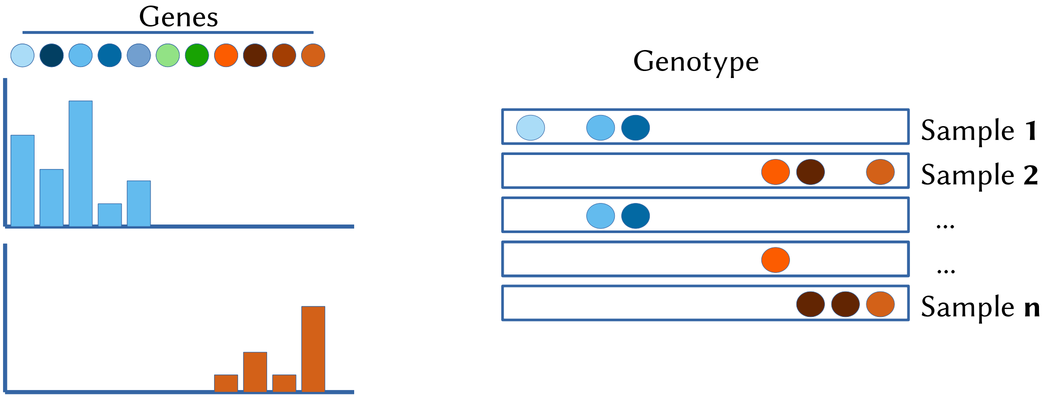


FIGURE 2.2: Structure underlying cancer genotypes encoded in the finite multinomial mixture model. The model reflects the hypothesis that the n genotypes arise from two different classes, each class being characterized by different probabilities of mutation of the m genes (left). Note that, in a genotype, mutations (right) comes from only one component. The parameters of the two multinomial, $\boldsymbol{\theta}_1; \boldsymbol{\theta}_2$, encode the probability of mutation per gene per class. The mixture weights summarize the proportion of classes in the dataset.

If we call $\boldsymbol{\theta}_j = (\boldsymbol{\theta}_{j1}, \ldots, \theta_{jm})$, the vector of gene mutation probabilities of class j, the probability of observing a genotype $\mathbf{g}$, fixed the class, is given by a multinomial distribution:

$$F(\mathbf{g}|\boldsymbol{\theta}_j) \propto \prod_{v=1}^{m} \theta_{jv}^{g_v} \tag{2.14}$$

As other clustering methods, finite mixtures have the drawback that the usually unknown number of clusters, K, has to be fixed a priori and several models with different K has to be compared a posteriori.

Moreover, modelling cancer genotypes we expect the existence of gene groups shared among cancer patients and cancer types. Bayesian non parametric setting allows to overcome the former problem, while Latent Dirichlet Allocation and Hierarchical Dirichlet Process generalize mixture models allowing genotypes to share gene from different groups.

2.1.2 Nonparametric Bayesian Methods

A central problem in bayesian statistics is to select the right model complexity, i.e., how many parameters are needed to capture the important structure of a dataset without overfitting [35].

In parametric methods, model comparison is performed explicitly a posteriori. Given the data, D, the probability of a model m_i is given by the Bayes rule:

$$P(m_i|D) = \frac{P(D|m_i)P(m_i)}{P(D)} \tag{2.15}$$

Assuming a flat prior on models, $P(m_i)$, $P(m_i|D)$ is proportional to the *evidence*, the probability of the data given the model i:

$$P(D|m_i) = \int P(D|\theta, m_i)P(\theta|m_i)d\theta \tag{2.16}$$

Comparing two different models, the model with higher evidence is better.

In other words, if we see the evidence as the probability of generating the dataset D by choose randomly parameters of model i, too complex models are unlikely to generate that particular data set at random since the probability is spread over a large parameter space, on the other side too simple models have low probability of generating that dataset at random.

Even tough many techniques have been developed based on this approach, its main drawback is that the evidence is hard to compute.

On the other side, bayesian nonparametric approach reflects the fact that constraining the number of parameters do not fit our prior beliefs about the data generating process and allows a model to have an infinite number of parameters.

In nonparametric methods, the appropriate model complexity is determined directly from data. Loosely, we can think that they pose a prior distribution on model complexities that is sharpened by computing the posterior.

In clustering methods model complexity is primary related to the number of expected clusters K. In a nonparametric setting, on one side, we allow the a priori number of clusters to be infinite and the other side, computing the posterior, only a finite number of clusters will be associated with non zero weights.

Before introducing in more details the theory of infinite mixture modelling we can sketch briefly the idea underlying them.

Introducing finite mixture models, we have shown that the random choices of mixture weights, $\boldsymbol{\pi}$, from the Dirichlet Prior and of the cluster specific distributions, $\boldsymbol{\theta}_j$ for $j \in \{1, 2, \ldots, K\}$ from the prior H is equivalent to generate a random probability distribution of the form:

$$G_{(k)} = \sum_{i=1}^{K} \pi_i \delta_{\boldsymbol{\theta}_i} \tag{2.17}$$

In an infinite mixture, we have to assign π and $\boldsymbol{\theta}_j$ to a countably infinite number of clusters. The Dirichlet process (DP) is a prior which generates random discrete distribution with a countably infinite number of probability 'atoms' (figure 2.3). In other words, a draw from a DP is a distribution of the form:

$$G_{(\infty)} = \sum_{i=1}^{\infty} \pi_i \delta_{\boldsymbol{\theta}_i} \tag{2.18}$$

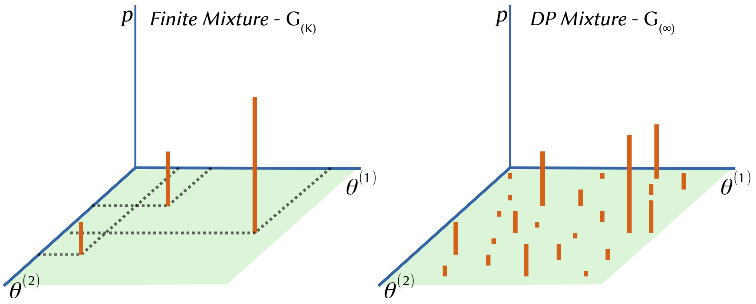


FIGURE 2.3: Intuitive representation of a draw from a Dirichlet process. It generates a discrete probability distribution with a countable infinite number of probability atoms.

2.1.3 The Dirichlet Process

The Dirichlet Process is a stochastic process which generates random discrete distribution with a countably infinite number of probability atoms. Here, we first introduce more formally the notions of probability atoms and random measures and later we introduce the stick-breaking construction of the DP. Let $\mathbb{X}$ be a set and $(A_1, A_2, \ldots)$ a disjoint partition of $\mathbb{X}$. A measure μ on $\mathbb{X}$ is a function from the subsets of $\mathbb{X}$ to $\mathbb{R}_+$ with the property $\mu(\cup_n A_n) = \sum_n \mu(A_n)$. An *atomic measure*, $\mu = \delta_\theta$, places a unit mass at point θ, such that:

$$\delta_{\boldsymbol{\theta}}(A) = \begin{cases} 1 & \text{if } \boldsymbol{\theta} \in A \\ 0 & \text{otherwise} \end{cases}$$

Since linear combinations of measures, with non negative coefficients, are measures, $\delta_{\boldsymbol{\theta}}$ can be used to build more general discrete measures of the form:

$$\mu = \sum_{i=1}^{K} \omega_i \delta_{\boldsymbol{\theta}_i} \tag{2.19}$$

with $\omega_i \geq 0$. Moreover, if $\mu(\mathbb{X}) = 1$ we a have a probability measure. The DP generates random discrete probability measures with a countably infinite number of probability atoms:

$$G_{(\infty)} = \sum_{i=1}^{\infty} \pi_i \delta_{\boldsymbol{\theta}_i} \tag{2.20}$$

Going back to the case of the finite mixture of K components, we saw that the discrete random distribution was generated in two step. First generating the K weights π_i from the Dirichlet distribution, then generating the parameters θ_j for the cluster specific distribution from their prior H. Similarly, a DP process can be built in two step:

- choosing atoms location randomly from a base distribution, usually called *base distribution*, H, with support on the on the parameter space,
- building a process which generates an infinite number of weights that sum to one.

There are several equivalent ways to generate the distribution weights. Below we introduce the Stick-Breaking process and the Chinese Restaurant Process. A schematic representation is reported in figure 2.4.

The Stick-Breaking Construction

The stick-breaking process [1] generates the weights as an infinite collection of fragments of a stick of initial length 1, figure 2.4:

- we start with a stick of length $l_1 = 1$ and we break it at a random point, $\beta_1 \sim Beta(1, \alpha)$:
$$\pi_1 = \beta_1 \tag{2.21}$$
- the remaining part has length $l_2 = 1 - \beta_1$. We repeat the stick-breaking removing a proportion $\beta_2 \sim Beta(1, \alpha)$:
$$\pi_2 = \beta_2 l_2 = \beta_2(1 - \beta_1) \tag{2.22}$$

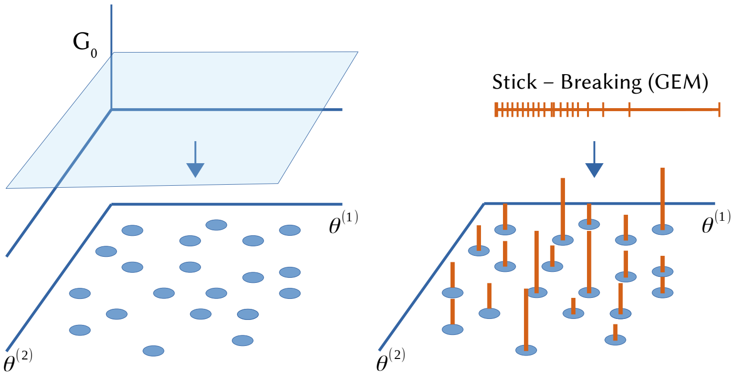


FIGURE 2.4: Stick-breaking representation of the Dirichlet Process. Atoms locations are drawn from the *base* distribution G_0 (here a uniform distribution). Atoms weights are assigned with the stick-breaking construction.

the length of part is $l_3 = (1-\beta_1) - \beta_2(1-\beta_1) = (1-\beta_1)(1-\beta_2)$. We can thus define a recurrence formula for the weights π_i:

$$\begin{cases} \pi_1 = \beta_1 \\ \pi_i = \beta_i \prod_{k=1}^{(c-1)} (1-\beta_k) \end{cases}$$

The stick-breaking process defines a distribution on the infinite vector of weights usually referred to as $GEM(\alpha)$:

$$\boldsymbol{\pi} \sim GEM(\alpha) \tag{2.23}$$

The parameter α of the $Beta(1,\alpha)$ distribution is called *concentration parameter* of the DP. To understand its meaning recall that the expectation value of $Beta(1,\alpha)$ is:

$$E_{Beta} = \frac{1}{1+\alpha} \tag{2.24}$$

Changing α changes the mean fraction of the stick to be removed at each step. Low α correspond to high probability mass in few atoms, as α grows probability is spread over an increasing number of atoms.

The Chinese Restaurant Process

Suppose we have a set of n points, a set of subset of them is called a *partition* [1]. For example, if $n = \{1,2,3,4,5\}$ a partition of n is $\{\{1\},\{2,3\},\{4,5\}\}$. Let's call $\boldsymbol{\pi}_{(k)}$ a partition of n points.

We have seen that stick-breaking process generates an the infinite vector of weights. $\boldsymbol{\pi}$, that sums to one. Once we have the weights, a random partition of n points can be generated by assigning each point to a group i with probability proportional to π_i. Thus the GEM define indirectly a distribution on partitions of the n points.

Differently from stick-breaking, the chinese restaurant process (CRP) generates directly random partitions of n points and thus defines a probability distribution on partitions. As usual, let's introduce the CRP by using the metaphor of a restaurant, where tables represents clusters and customers represent data points, a random partition is generated as follows: suppose we have an infinite number of unoccupied tables. Customers arrive at the restaurant one at a time. The first customer seats at a random table. Following customers choose:

- an unoccupied table with probability proportional to a constant α
- an occupied table with probability proportional to the number of customers sitting at that table.

After k customers have arrived at the restaurant, let's call $|t|$ the number of them sitting at table t. Formally, the probability of $k+1$ costumer to join a table t, $P(k+1 \to t|\boldsymbol{\pi}_{(k)})$, is given by:

$$P(k+1 \to t|\boldsymbol{\pi}_{(k)}) = \begin{cases} \dfrac{|t|}{\alpha+k} & \text{if } t \text{ is already occupied} \\ \dfrac{\alpha}{\alpha+k} & \text{otherwise} \end{cases}$$

When all the n costumers have arrived at the restaurant, the process has generated a random partition of them across the tables. Different realizations of the CRP random process define different random partitions, so we can write:

$$\boldsymbol{\pi}_n \sim CRP(\alpha, n) \tag{2.25}$$

note that $\frac{1}{\alpha}$ has the role of concentration parameter: the higher α the higher the number of different clusters. The explicit expression for the CRP probability of a partition on n points in K clusters $\boldsymbol{\pi}_{(n)}$ is easily obtainable:

$$P_{CRP}(\pi_n) = \frac{\alpha^K}{\alpha(\alpha+1)\dots(\alpha+n-1)} \prod_{t\in\{1,\dots,K\}} (|t|-1)! \tag{2.26}$$

The Posterior Dirichlet Process

Given the generative model:

$$G \sim DP(\alpha, G_0) \tag{2.27}$$

$$\theta_i | G \sim G \ \text{ for } i = 1, \ldots, N \tag{2.28}$$

The posterior Dirichlet process is given by:

$$G|\theta = DP\left(\alpha + N; \frac{\alpha}{\alpha + N} G_0 + \frac{a}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i}\right) \tag{2.29}$$

2.1.4 Dirichlet Process Mixture Model

Once defined the Dirichlet process prior, The DP mixture model has the same generative structure of finite mixtures:

- a random probability distribution with an infinite number of atoms, G is drawn from a DP:

$$G \sim DP(\alpha, G_0) \tag{2.30}$$

- for each observation, x_i a vector of parameters $\boldsymbol{\theta}_i$, is generated randomly from G:

$$\boldsymbol{\theta}_i \sim G \ \text{ for } i = 1, \ldots, N \tag{2.31}$$

- x_i is drawn the cluster specific distribution F:

$$x_i | \boldsymbol{\theta}_i \sim F(\boldsymbol{\theta}_i) \ \text{ for } i = 1, \ldots, N \tag{2.32}$$

2.2 Latent Dirichlet Allocation and Hierarchical DP

Mixture models, applied to genotype clustering, assume that mutated genes in a specific patient come from the same underlying group. In other words, the hypothesis is that there is a one-to-one correspondence between groups of mutated genes and patients genotypes. This hypothesis can be satisfactory when analysing different cancer types, which are characterised, at least partially, by different gene sets. However, it is somewhat restrictive in other situations, when we would allow gene groups to be shared among patients.

To better understand this point, let's do a simplified example (see figure 2.5): suppose a given tumour arises when three different molecular pathways are impaired by mutations, each pathway being characterized by the genes that take part in it. The impairment of a pathway can be caused by mutations in different genes and patients will show different mutational patterns, comprising different mutated genes from the three gene groups. Mixture models would not be able to capture this complexity.

The Latent Dirichlet Allocation (LDA) and its non parametric counterpart, the Hirarchical Dirichlet process (HDP) are a generalization of mixture models that allows gene clusters to be shared among patients.

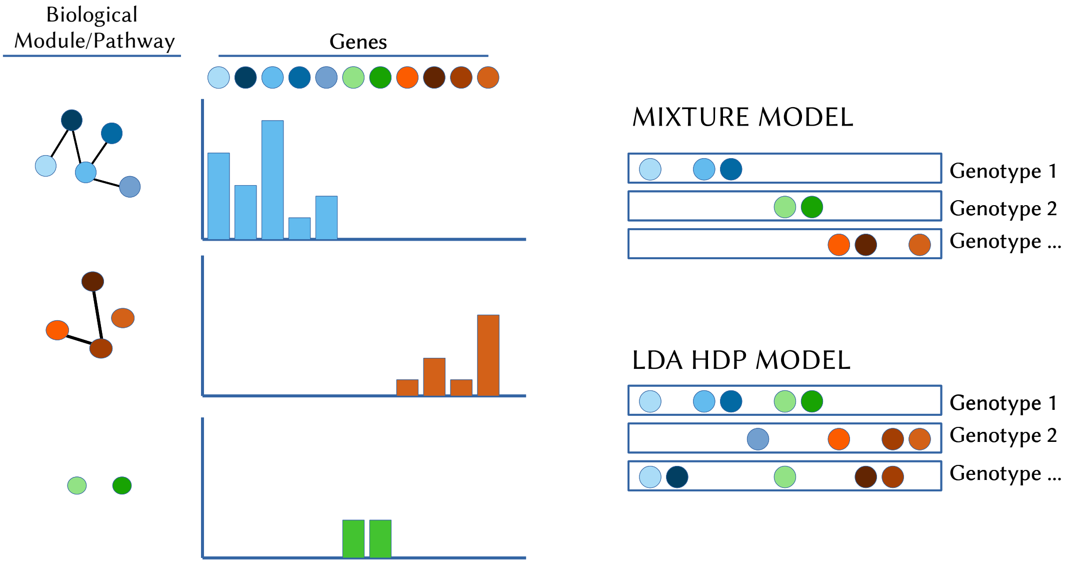


FIGURE 2.5: Schematic representation of the differences between MM and LDA - HDP data structure. While for MM a genotype has mutations from a single cluster, LDA-HDP allows clusters to be shared. This is a more biological plausible hypothesis since we expect a cancer to be originated from from the impairment of different biological pathways/modules.

2.2.1 Latent Dirichlet Allocation

LDA was developed in the context of topic modelling with the aim of cluster text documents based on their content [17]. A document could be, for example, a scientific article. A *corpus* is a set of documents. LDA is built on two hypothesis: each document in a corpus is a mixture of several topics and topics are shared among documents in different proportions. For example, if we assume that our documents are scientific articles in *cancer computational genetics* there will be three main topics: *cancer*, *computer science* and *genetics*. Depending on the specific article there will be more words related to one or the other topic.

Topic modelling assumes documents are *bags of words*, i.e what matters is only the number of occurrences of each word. A dictionary is the set which comprises all the words in the documents of a corpus. Formally, a topic can be defined as a probability distribution over the words in the dictionary. So, for example, the topic *cancer* defines a probability distribution on the dictionary which puts high probability on words about cancer (e.g, tumour, metastasis, diagnosis, etc.).

LDA defines a generative process assuming that each document is generated as follows:

- An a priori fixed number of topics, k, is generated. Formally this means that k distributions over the words in the dictionary are chosen.
- for each document in the corpus, randomly choose a weight for each topic, i.e a distribution over topics
- for each word in the document:
    - randomly draw a topic with probability proportional to the topic weights
    - randomly draw a word with probability given by the topic distribution aver the vocabulary

It is important to note that LDA do not have any prior information about topics structure. They are encoded in the co-occurrences of words in the documents and emerge from model inference.

Let's introduce LDA formally, using the notation introduced above for cancer genotype modelling.

A cancer genotype is given by $\mathbf{g}^i = (g^i_1, g^i_2, \ldots, g^i_m)$ where m is the number of genes considered. If gene v is mutated t times, $g^i_v = t$.

LDA supposes the existence of an a priori fixed number of gene groups K. Each group is characterized by different probability of mutation per gene and each patient genotype is built by choosing genes from some, or all, the K classes in different proportions. Modelling genotype, we assume that the probability of mutation of gene v in group k, is a multinomial with parameters θ_{kv} for $v = 1, \ldots, m$. Then, the LDA generative process is:

- K multinomial distribution over the m genes, with vector of parameters $\boldsymbol{\theta}_j$, are choosen from a Dirichlet distribution prior:

$$\boldsymbol{\theta}_j \sim Dir(\boldsymbol{\alpha}) \text{ for } j = 1, \dots, K \tag{2.33}$$

  where $\alpha \in \mathbb{R}^m_+$.

- for each genotypes i, a vector of weights $\boldsymbol{\pi}_i$ is drawn from a Dirichlet prior :

$$\boldsymbol{\pi}_i \sim Dir(\boldsymbol{\alpha}') \tag{2.34}$$

  where $\alpha' \in \mathbb{R}^K_+$

- for each mutation, k, in the genotype:
  - randomly draw a gene group with probability proportional to π_i:

$$z_k|\boldsymbol{\pi}_i \sim Cat(\boldsymbol{\pi}) \tag{2.35}$$

  - randomly draw a gene to be mutated with probability given by the selected gene group distribution:

$$g_k|z_k \sim Cat(\boldsymbol{\theta}_{z_k}) \tag{2.36}$$

LDA is a parametric model in that it specify a fixed number of parameters and in particular the usually unknown number of topics K. Its non parametric generalization, is the Hierachical Dirichlet Process.

2.2.2 The Hierachical Dirichlet Process

The Hierachical Dirichlet Process is the nonparametric extension of LDA and allows not to fix a priori the number of expected gene groups. As in the DP, K is determined directly from the data. The step from LDA to HDP can be easily understood, as in the case of finite mixtures and DP, in terms of random probability distributions. If we consider the first two step of LDA:

- drawing K group specific distributions over the the genes:

$$\boldsymbol{\theta}_j \sim Dir(\boldsymbol{\alpha}) \text{ for } j = 1, \dots, K \tag{2.37}$$

- for each genotype, draw a the vector of group weights:

$$\boldsymbol{\pi}_i \sim Dir(\boldsymbol{\alpha}') \tag{2.38}$$

we can see that they are equivalent to randomly choose, for each genotype i, a discrete distribution,$G^i_{(K)}$, of the form:

$$G^i_{(K)} = \sum_{k=1}^{K} \pi_k \delta_{\boldsymbol{\theta}_k} \tag{2.39}$$

The first step, common to all documents, choose the atom locations, $\boldsymbol{\theta}_k$ while the second step choose for each document a different vector of weights.

Intuitively, the shift to the nonparametric setting can be done by replacing, for each genotype the finite random measure, $G^i_{(K)}$, with a discrete infinite random measure:

$$G^i_{(\infty)} = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\theta}_k} \tag{2.40}$$

$G^i_{(\infty)}$ can be generated by placing a Dirichlet process prior $DP(\alpha, G_0)$ on it.

However, for a general base distribution G_0, different draws from $DP(\alpha, G_0)$ would have different set of probability atoms, without possibility of shearing clusters between genotypes. The key step is then to enforce the genotype specific distribution, $G^i_{(\infty)}$, to share the same probability atoms. This can be done by choosing them from a discrete base distribution which is itself drawn from a Dirichlet process, $DP(\mu, H)$. With this assumptions, the HDP generative process is:

$$G_0 \sim DP(\mu, H) \tag{2.41}$$

$$G^i \sim DP(\alpha, G_0) \;\; \text{for } i \in \text{genotypes} \tag{2.42}$$

$$\theta_{iv} | G^i \sim G^i \;\; \text{for each muatation } v \text{ in genotype } i \tag{2.43}$$

$$g_{iv} | \theta_{iv} \sim Mult(\theta_{iv}) \;\; \text{for each muatation } v \text{ in genotype } i \tag{2.44}$$

The number of hierarchical steps is not limited, for example if it is known a priori that genotypes belong to different classes (e.g., different tumour type or subtypes) other hierarchical levels can be added.

2.3 Bayesian Networks

Given a set of random variables a *Bayesian Networks* (BN) is a graphical way to represent causal probabilistic dependencies among them, i.e. how the value taken by a given variable influences the probability of the others.

They main hypothesis underlying BN is that the joint probability distribution (j.p.d.) over a set of variables can be represented as a Directed Acyclic Graph (DAG), a directed graph with no loops. DAG nodes represent random variables, with the associated probability distribution, P_i and links represent causal dependences between variable pairs. For example, an arrow from node A to node B is a probabilistic directed dependence of B on A. Directed dependence means that the probability of the *child* node B is influenced by the value taken by its *parent* A. The link is associated to the conditional probability distribution $P(B|A)$.

The assumption of a DAG structure implies that the probability of variable i given its parents, is independent of its non descendent nodes. This independence assumption provide a factorization of the j.p.d.

Let's illustrate BN for modelling mutational interdependencies among gene mutations. Given a set of genes $(g_1, \ldots, g_m)$, we consider the presence of mutations on gene v to be a random variable with $p(g_v = t_v)$ being the probability of observing t_v mutations on v. The probability for a genotype $\mathbf{g}$ is given by the joint probability distribution $P(g_1 = t_1, \ldots, g_m = t_m)$. Moreover, the conditional probability $P(g_v|g_i)$ expresses the mutational probability of gene v when gene i is mutated.

Specifying a directed acyclic graph structure among mutation probabilities amounts to specify probabilistic causal relations among them, building a hierarchy where parent mutation in parent genes constraint probability of child genes.

2.4 Myelodysplastic Syndromes Genomic Landscaping

2.4.1 Introduction

Myelodysplastic syndromes (MDS) are clonal hematopoietic disorders characterized by peripheral blood cytopenias due to ineffective hematopoiesis and increased risk of evolution into acute myeloid leukemia (AML) [2]. Current disease classification provided by the World Health Organization (WHO) mainly uses morphological features to define MDS categories, leading to a clinical overlap between subtypes and to low inter-observer reproducibility in the evaluation of marrow dysplasia [6, 30].

MDS range from indolent conditions to cases rapidly progressing into AML [62]. Therefore, a risk-adapted strategy is needed in such heterogeneous disorders. Currently, individual disease-related risk is assessed by International Prognostic Scoring System (IPSS, later revised as IPSS-R) based on clinical and hematological features [61, 38]. While IPSS/IPSS-R are excellent tools for clinical decision-making, these scoring systems have their own weaknesses and may fail to capture reliable prognostic information at an individual patient level [31].

Biologically, the development of MDS is driven by mutations on genes involved in RNA splicing, DNA methylation, chromatin modification, transcriptional regulation and signal transduction [22, 67, 90, 68]. Many patients have additional mutations that span a wide range of cancer genes, with high patient-to-patient variation. Chromosomal abnormalities (including copy number alterations, chromosomal translocations and complex karyotype) also contribute to MDS pathophysiology [78]. Despite recent progresses in understanding the disease biology, MDS with isolated 5q deletion is the only category defined by a specific genomic abnormality in the current WHO classification2 and only few genotype-phenotype associations have been reported until now, mainly referring to the close relationship between mutations in SF3B1 gene and MDS subtypes with ring sideroblasts.

In myeloid malignancies, a progressive shift is underway, where classifications based on clinical and morphologic criteria are being complemented by introducing genomic features which are closer to the disease biology and better capture clinical-pathological entities [39, 69, 34]. Moreover, as mutations are often responsible for the disease phenotype, they may represent strong predictors of clinical outcomes [39, 69, 34]. Comprehensive analyses of large patient populations are warranted to correctly define specific genotype-phenotype correlations and to estimate the independent effect of each genomic abnormality on clinical outcome. Here, we aimed to define a new genomic classification of MDS and to improve individual prognostic assessment moving from scoring systems based on clinical parameters to models including genomic information.

2.4.2 Material and Methods

Study populations

We studied a retrospective international cohort of 2043 patients affected with MDS according to 2016 WHO criteria, from EuroMDS consortium and an independent cohort of 318 patients prospectively diagnosed at Humanitas Research Hospital, Milan Italy. Overall frequency of mutations and chromosomal abnormalities is reported in figure 2.6. Variant allele frequency (VAF) of mutations related to the main gene functions involved in MDS (drivers) are reported in figure 2.7.

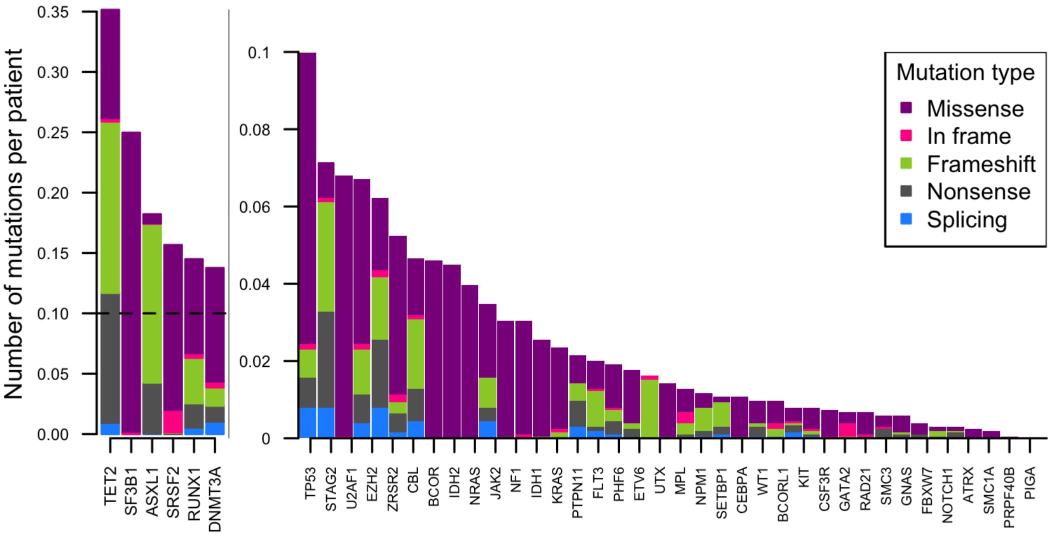


FIGURE 2.6: Frequency of mutations and chromosomal abnormalities in the cohort of 2043 patients, stratified according to type (missense, nonsense, affecting a splice site, or other).

HDP clustering

In order to identify MDS molecular subtypes we carried out Dirichelet Process Clustering (DP) following the current state of the art works [39, 69, 34].

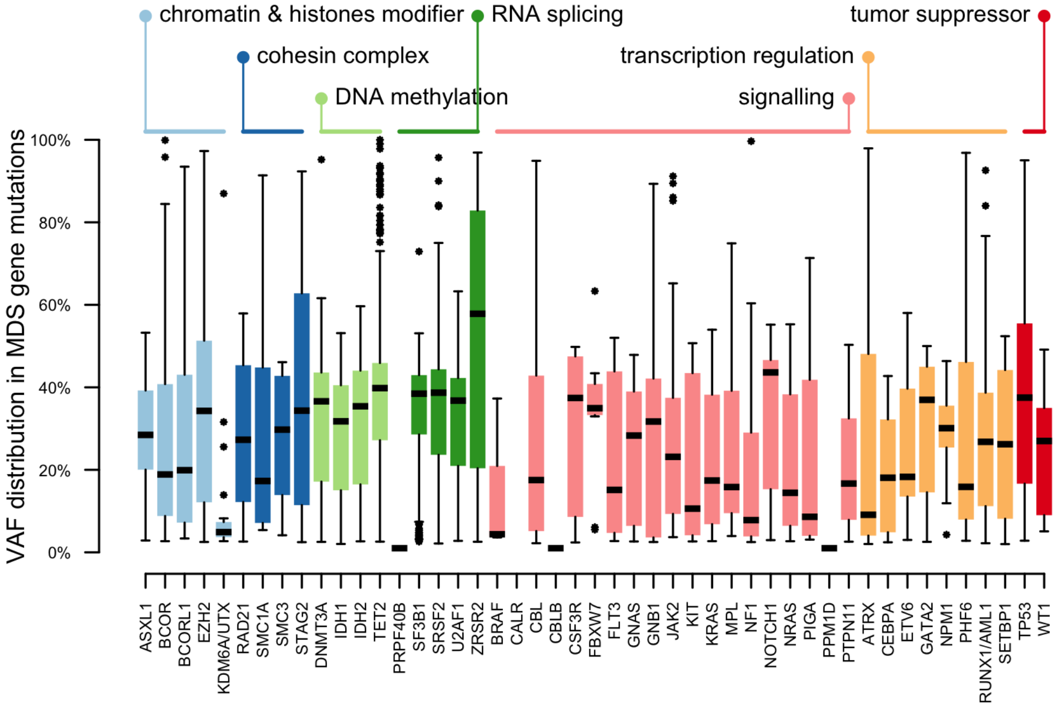


FIGURE 2.7: Driver mutations VAF in the cohort of 2043 patients (25-75 percentiles and ranges).

The HDP infinite multinomial mixture model allows to capture broad dependencies among all gene mutations assuming them to be extracted from a hierarchical mixture of multinomials. The rational underlying the model is that we expect mutations to be clustered together according to the specific molecular mechanism at work in a given tumor. Using an infinite mixture with DP prior, instead of finite mixture, allows not to specifying a priori the number of mutations categories which, instead, is inferred from the data. To carry out the analysis we used the R package HDP available online https://github.com/nicolaroberts/hdp.

The input data consists of a patients by genes binary matrix. The genotype of a patient is a row of the matrix: $mathbf{g} = (g_1, \dots, g_n)$; where n is the number of features per patient, i.e., 12 cytogenetic and 47 genomic variables. G_{ij} is a binary variable which denotes the presence or absence of i-th alteration. Missing data where imputed with several methods with not significant differences on final results.

We carried out Monte Carlo Markov Chain (MCMC) sampling of DP posterior for 4 different initial conditions (n. of different chains). For each chain we discarded the first 3000 iterations and we sampled 4000 realizations at intervals of 20 iterations.

Starting from the raw clusters classes are built by grouping them into components according to the following conditions:

- clusters are merged if their cosine similarity is above a give threshold (0.95 in our case)
- clusters are assigned to component 0 if they have no significant data categories or sample exposure.

Components 1-5 account for the 97% of the data while component 0 accounts for data that cannot be explained by the model. The model found a mixture of 5 components, plus an additional one of unexplained data. Figure 2.8 reports the box plot of the distributions of data items in each component, i.e. the weight of each component in the mixture while figure 2.9 summarizes the mean distribution of data categories for each component (class), i.e. the mean parameters the multinationals forming the data structure.

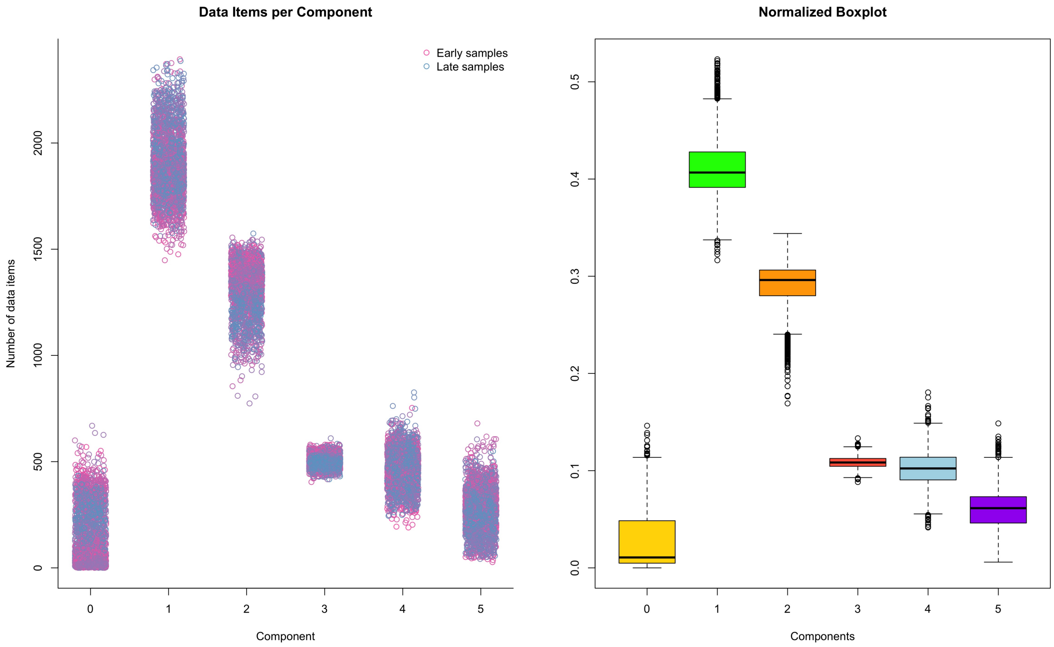


FIGURE 2.8: Left: number of data items (mutations or cytogenetic alterations) in each of the 6 components for each sample of the MC. Right: Components weights boxplot.

Bayesian Network Analysis

We used Bayesian Networks (BN) to define in a more comprehensive way the relationships between genomic abnormalities in MDS. As for HDP clustering, we included gene mutations and cytogenetic abnormalities as random

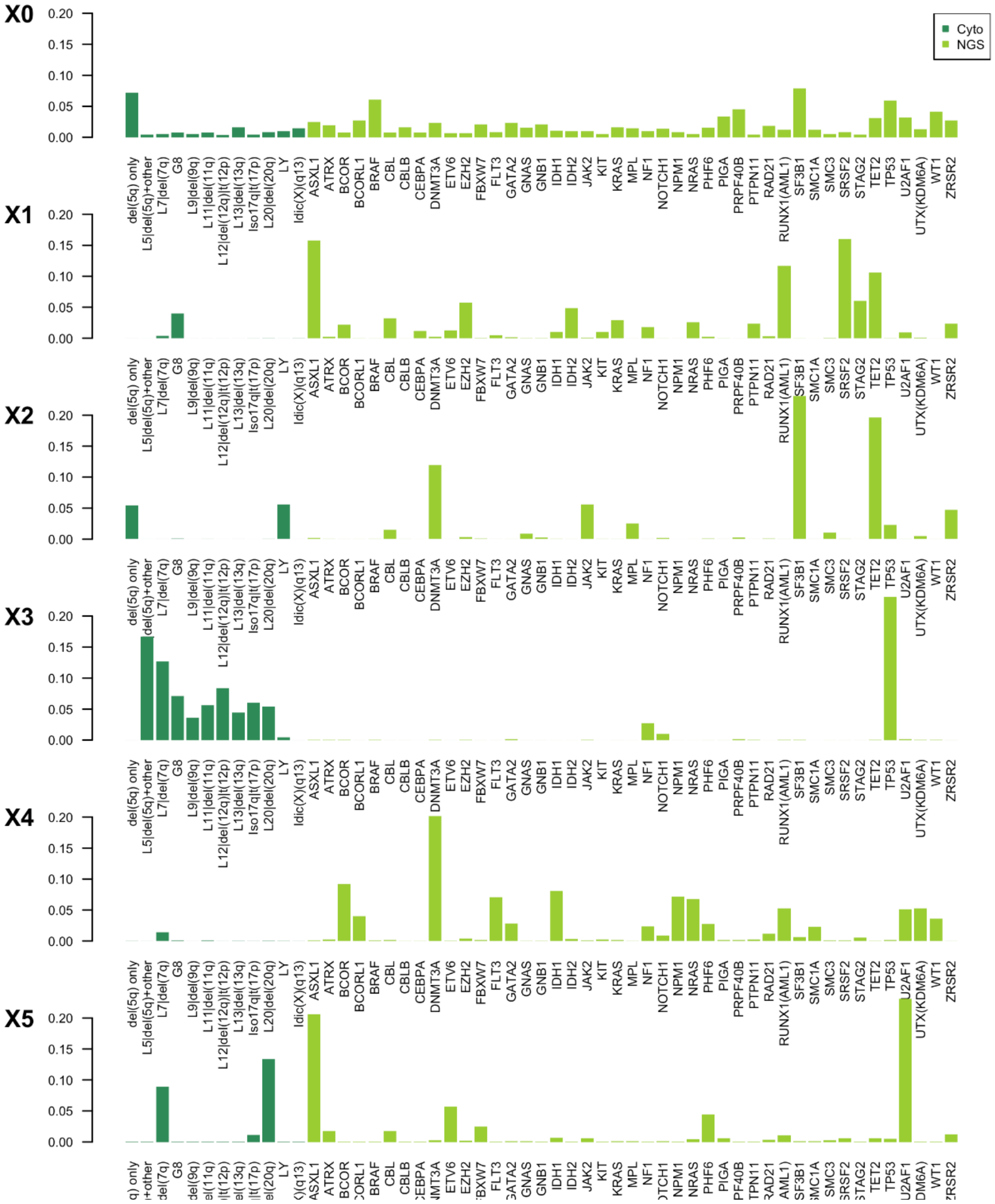


FIGURE 2.9: Mean distribution of data categories for each component resulting from the HDP applied to MDS learning cohort. Each distribution gives the mean parameters of a multinomial of the mixture (component 0 accounts for unexplained data).

variables in the model and we investigated conditional dependency among them. Given the training data we estimated the network structure (S) and the parameters of the JPD in the BN. We inferred the network structure from data using the GOBNILP software [28]. Given a set of random variables, GOBNILP assigns a score (based on data) to each Directed Acyclic Graph and choose the structure which maximizes the score (according to previous literature [39] we set the maximum number of parents to 3).

For each variable in which conditional dependency was found (i.e. a link in the inferred structure is present), the definition of mutually exclusivity was used to define a significant negative dependency, while the definition of co-occurrence was used to define a positive dependency, see figure 2.10.

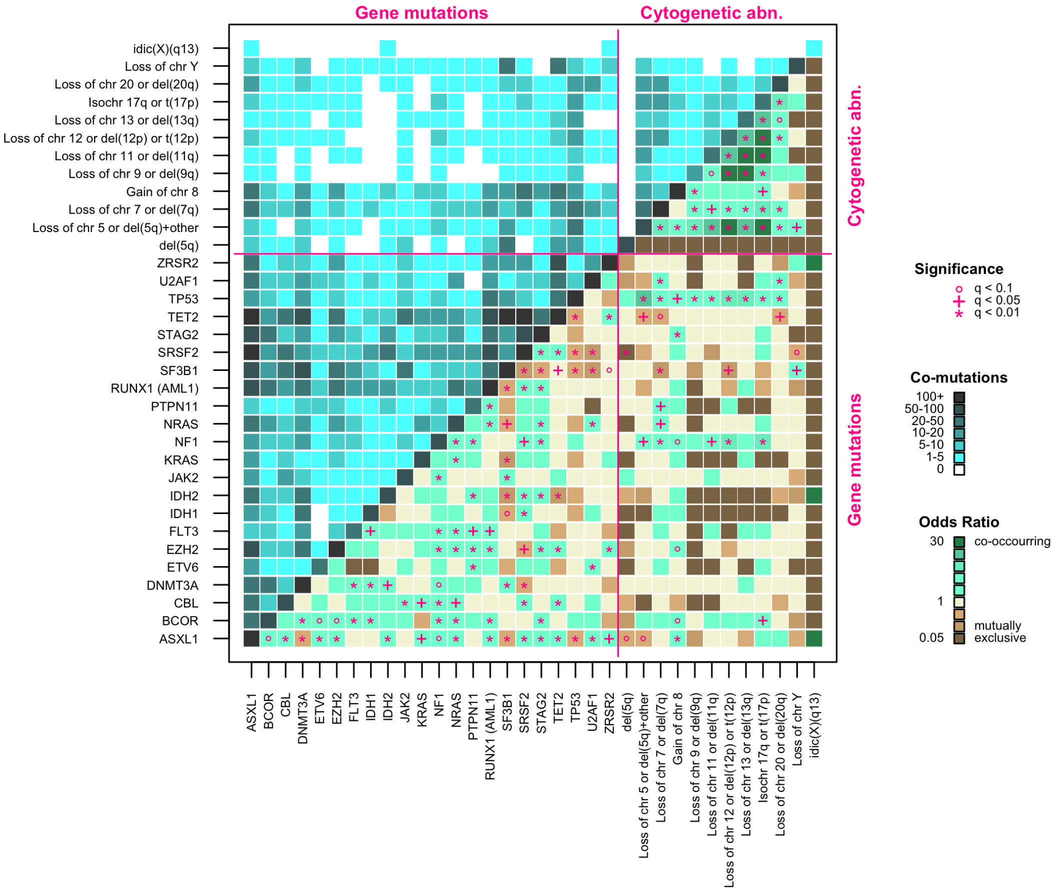


FIGURE 2.10: Co-occurrence (green) and mutual exclusivity (brown) of mutations and cytogenetic abnormalities in the patients cohort.

2.4.3 Definition of a genomic classification of myelodysplastic syndromes

To identify genomic subgroups among MDS we applied Hirerchial Dirichlet processes. We identified six components, each describing a specific distribution of random variables included in the model (i.e., cytogenetic abnormalities and gene mutations, see figure 2.8 and 2.9).

Importantly, genomic features of the 6 components are not (or minimally) overlapping. Each patient was characterized by a weight vector indicating the contribution of each of the 6 components to its genome.

We clustered patients based on the Euclidean distance between weight vectors and performing a hierarchical agglomerative clustering. We obtained 8 groups (clusters) defined according to specific genomic features 2.11 Dominant genomic features of each group were described through Bayesian networks. Multivariate logistic regression analysis was applied to compare clinical and hematological characteristics among different groups, while Kaplan-Meier and Cox multivariable were used to compare clinical outcome (overall survival). Only results showing $p \geq 0.05$ were reported. figure 2.12.One group included patients without specific genomic profiles; strikingly all the remaining groups were deeply characterized by a single (in some cases two) component of Dirichlet processes (figure 2.11).

In many groups dominant genomic features included splicing gene mutations. We identified two groups (1 and 6) in which dominant features were SF3B1 mutations, older age, presence of ring sideroblasts and transfusion-dependent anemia (Figure 2.12 and 2.13).

Group 6 included patients with ring sideroblasts and isolated SF3B1 mutations (except for co-mutation patterns including TET2, DNMT3A and JAK/STAT pathways genes [JAK2, CALR, MPL]), characterized by isolated anemia, normal/high platelet count, single or multilineage dysplasia and very low percentage of bone marrow blasts (median value 2%).

Group 1 included patients with SF3B1 with co-existing mutations in other genes (mainly ASXL1 and RUNX1), characterized by anemia associated with mild neutropenia and thrombocytopenia, multilineage dysplasia and higher bone marrow blast percentage with respect to group 6 (7% vs. 2%, $p < 0.0001$).

In *group 3* and *group 5*, dominant genomic features were represented by SRSF2 mutations. In these groups the most frequently reported chromosomal abnormality was trisomy 8.

Group 3 included patients with SRSF2 and concomitant TET2 mutations. Patients presented single cytopenia (anemia in most cases) and higher monocyte absolute count with respect to the other groups ($p < 0.0001$). Bone marrow features include multilineage dysplasia and excess blasts (median 8%). Group 5 was characterized by SRSF2 mutations with co-existing mutations in other genes (mainly ASXL1, RUNX1, STAG2, IDH2 and EZH2). Patients presented two or more cytopenias, multilineage dysplasia and excess blasts (median 11%, significantly higher with respect to group 3, $p = 0.0031$).

Group 4 dominant features included U2AF1 mutations associated with

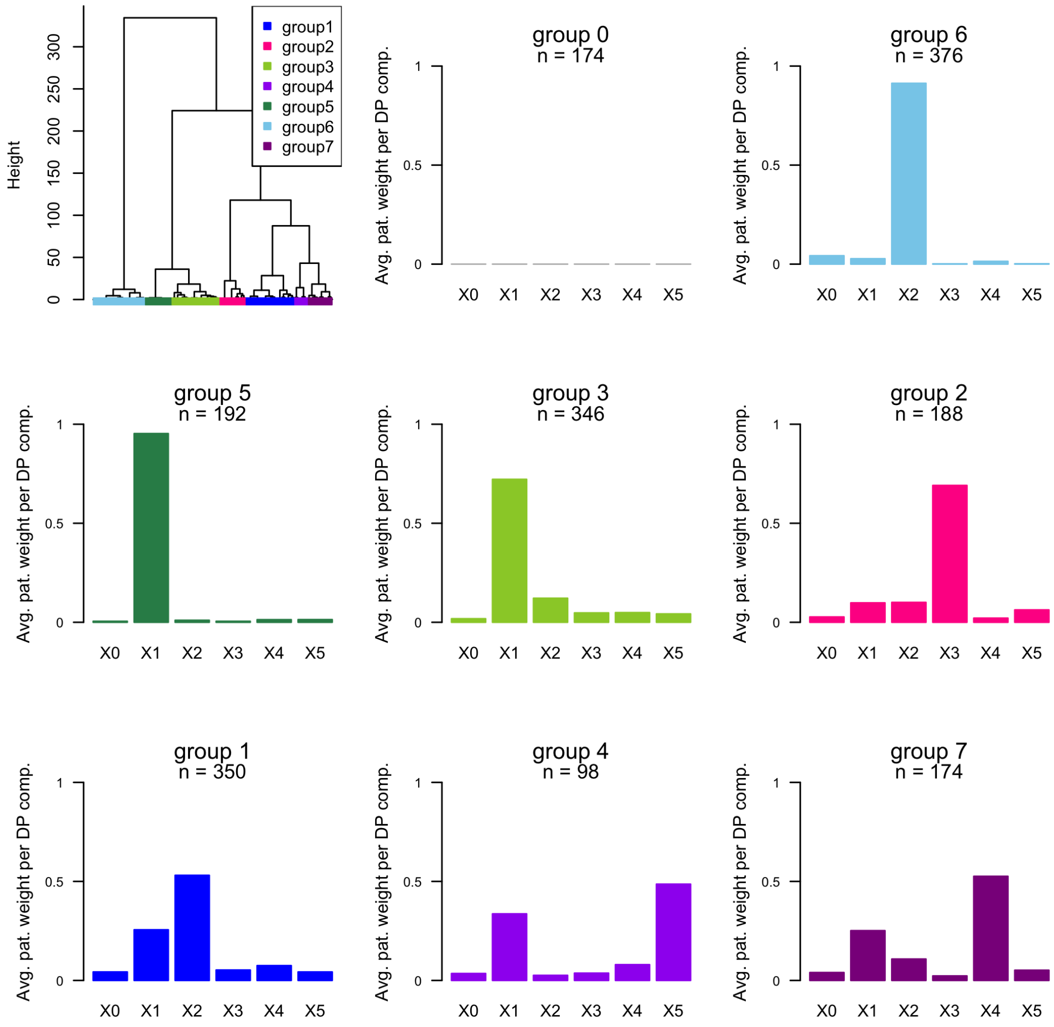


FIGURE 2.11: MDS patient Ward clustering using Euclidean distances of DP outcomes. Top-left panel: the dendogram cut from ward clustering of DP components for the clustering of MDS patients. Remaining panels: Average weight distribution of genomic categories per group. The 8 groups of the 2043 patients are labeled from 0 to 7 (in such fashion group 0 is self-explanatory).

deletion of chromosome 20q and chromosome 7 abnormalities. Patients presented a higher rate of transfusion-dependent anemia with respect to the other groups (p ranging from 0.023 to < 0.0001). Bone marrow features included multilineage dysplasia and excess blasts in most cases.

Group 2 was characterized by TP53 mutations and/or complex karyotype, mainly including abnormalities in chromosome 5, 7, 8, 11, 12, 17 and 20. Most cases presented with two or more cytopenias (with high rate of transfusion-dependency) and excess blasts (Figure 2.12 and2.13).

Group 7 included patients with AML-like mutation patterns (most including DNMT3A, NPM1, FLT3, IDH1 and RUNX1 genes). Patients are characterized by two or more cytopenias (with high rate of transfusion-dependency) and excess blasts, in most cases ranging between 15 and 19%.

Finally, *group 0* included MDS without specific genomic profiles, figure 2.12. These patients were characterized by younger age, isolated anemia with low rate of transfusion-dependency, normal or reduced bone marrow cellularity (with respect to age-adjusted normal ranges), absence of ring sideroblasts and low percentage of marrow blasts (median value 2%)(figure 2.13).

A significantly heterogeneous distribution of WHO 2016 disease subtypes was observed through the new groups defined by specific genomic features ($p < 0.0001$), figure 2.14. Interestingly, this new classification of MDS accounted for genomic heterogeneity of patients stratified according to WHO criteria. This was particularly evident for MDS with isolated 5q deletion. HDP and hierarchical clustering classified these 75 patients into group 1 and group 6: subjects with none or one mutation (mainly including SF3B1gene) were clustered into group 6, while those with 2 or more mutations or TP53 mutations were classified into group 1. MDS with 5q deletion included in group 6 showed lower rate of transfusion-dependency and lower percentage of marrow blasts with respect to patients classified into group 1 ($p = 0.0043$ and < 0.0001, respectively).

Then we focused on the distribution of mutation hotspots of splicing genes. No significantly different distribution of SF3B1 and SRSF2 hotspots was noticed among genomic MDS subgroups; in patients with U2AF1 mutations, p.S34F missense variant was mainly associated with bone marrow blast percentage $> 10\%$ ($p = 0.0199$). These findings provide the proof of concept for a new classification of MDS based on entities defined according to specific genomic features. In figure 2.14 we provided a diagram to classify patients in the appropriate category on the basis of individual genomic profile.

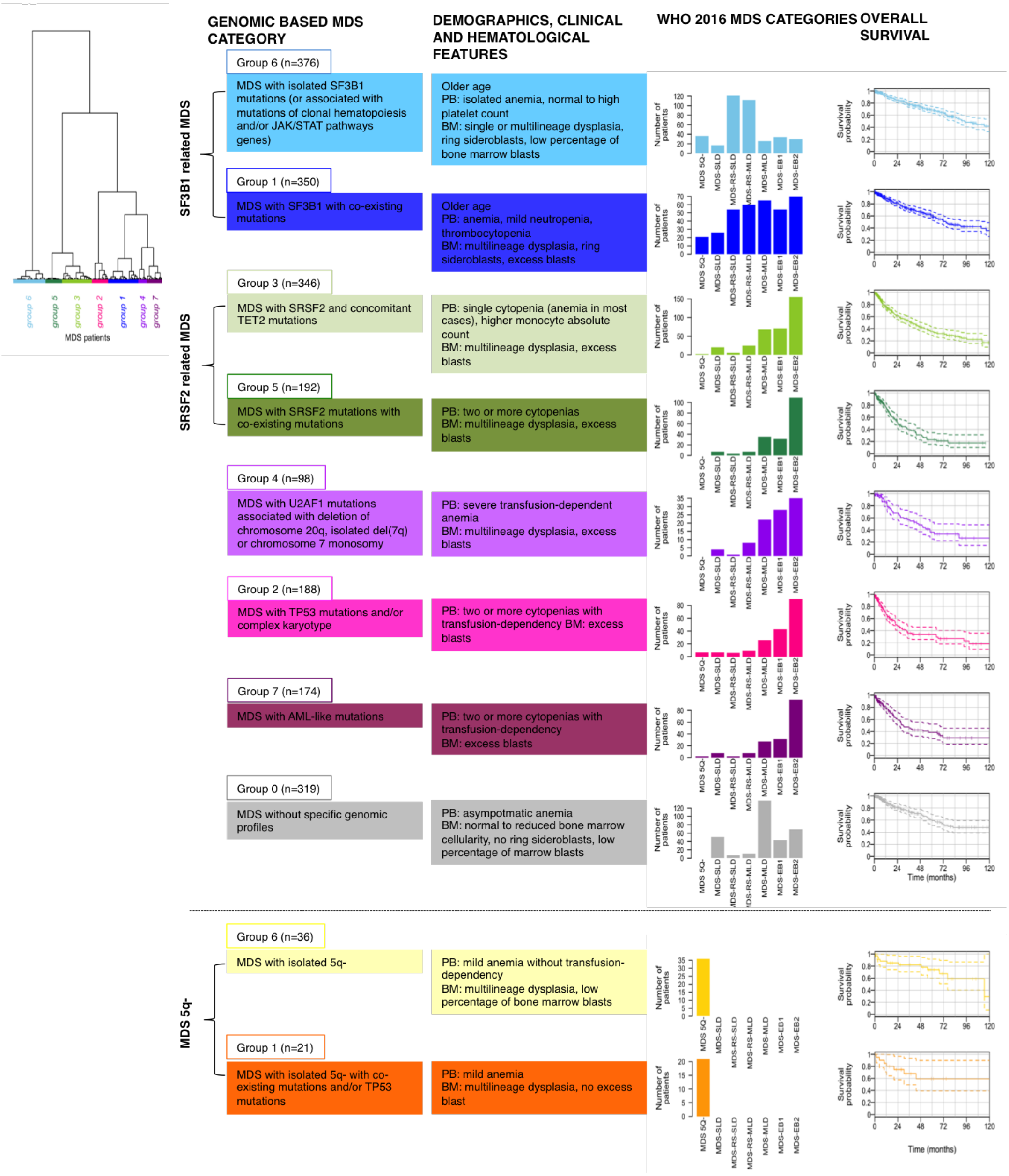


FIGURE 2.12: Genetic groups in Euro MDS cohort and their relationship with WHO category (according to 2016 classification) and overall survival. According to a Dirichelet process clustering algorithm, patients could be classified into eight distinct genetic groups on the basis of the presence or absence of mutations and chromosomal abnormalities: group 0) MDS without recurrent genomic abnormalities; group 1) MDS with ring sideroblasts with SF3B1 coexisting mutations; group 2) MDS with TP53 mutations and complex karyotype; group 3) MDS with SRSF2 mutations and concomitant TET2 mutations; group 4) MDS with U2AF1 mutations and del 20q; group 5) MDS with SRSF2 mutations and concomitant ASXL1, RUNX1 and NRAS pathway mutations; group 6) MDS with ring sideroblasts and SF3B1 isolated mutation; group 7) MDS with eb and AML-like mutations. These genetic MDS groups significantly differ both in WHO MDS categories distribution both in cumulative probability of overall survival.

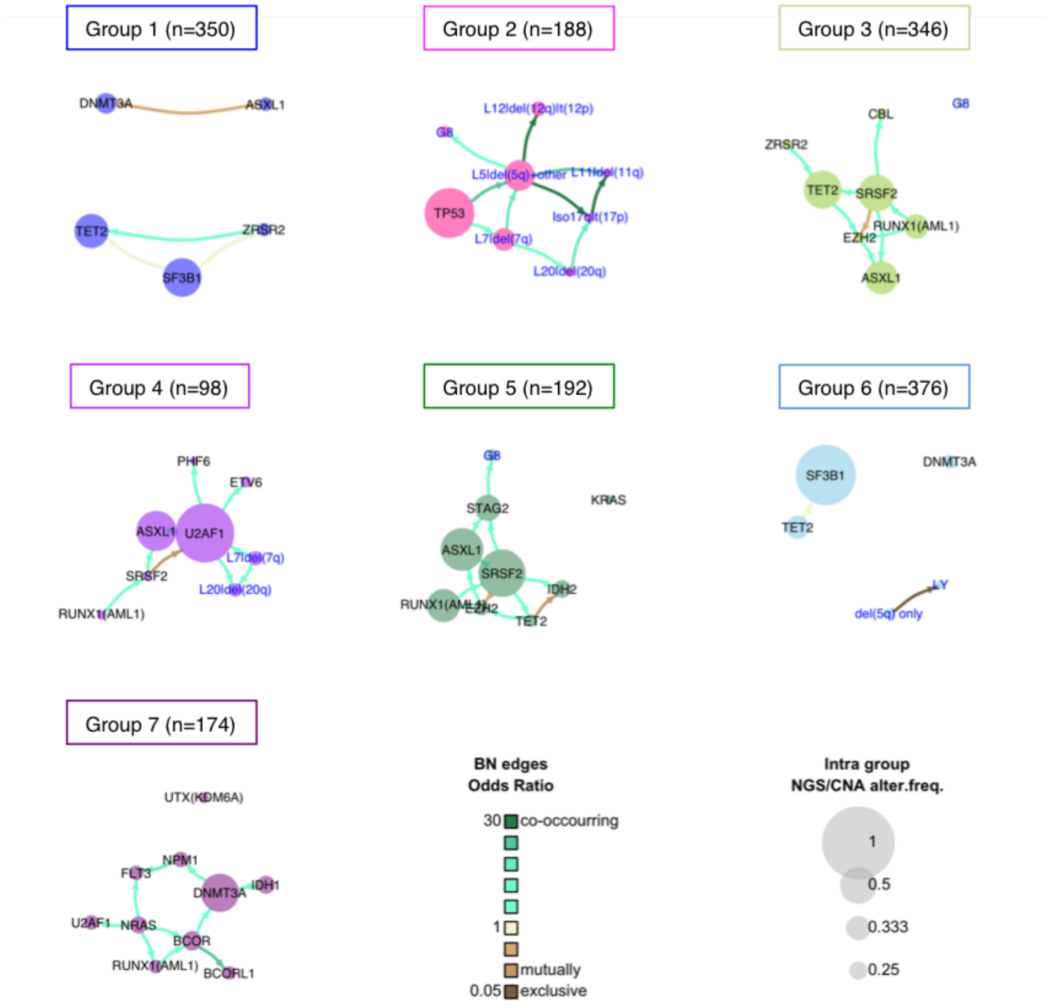
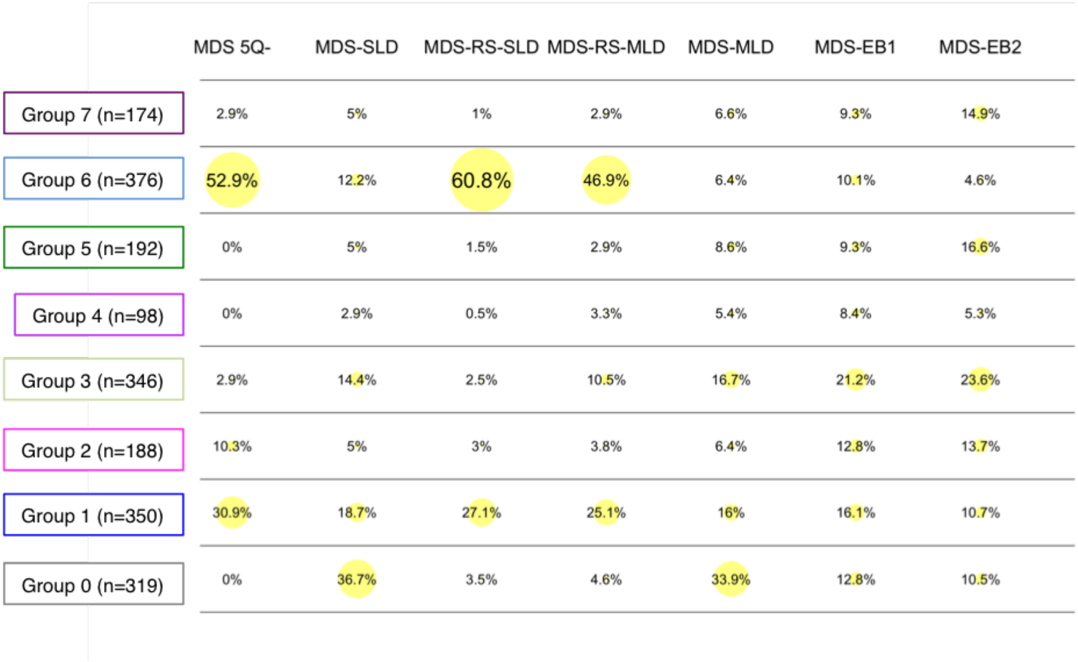


FIGURE 2.13: Extrapolation of genomic landscape through Bayesian Networks of MDS genetic groups. The size of each node accounts for the number of correspondent genomic or cytogenetic alterations. The color of each link reflects odds ratio (shades of brown represent mutual exclusivity while shades of green color degree co-occurrence). The thickness of edges grows with increasing significance of mutual exclusivity/co-occurrence between alterations.



	MDS 5Q-	MDS-SLD	MDS-RS-SLD	MDS-RS-MLD	MDS-MLD	MDS-EB1	MDS-EB2
Group 7 (n=174)	2.9%	5%	1%	2.9%	6.6%	9.3%	14.9%
Group 6 (n=376)	52.9%	12.2%	60.8%	46.9%	6.4%	10.1%	4.6%
Group 5 (n=192)	0%	5%	1.5%	2.9%	8.6%	9.3%	16.6%
Group 4 (n=98)	0%	2.9%	0.5%	3.3%	5.4%	8.4%	5.3%
Group 3 (n=346)	2.9%	14.4%	2.5%	10.5%	16.7%	21.2%	23.6%
Group 2 (n=188)	10.3%	5%	3%	3.8%	6.4%	12.8%	13.7%
Group 1 (n=350)	30.9%	18.7%	27.1%	25.1%	16%	16.1%	10.7%
Group 0 (n=319)	0%	36.7%	3.5%	4.6%	33.9%	12.8%	10.5%

FIGURE 2.14: Wide-ranging genomic heterogeneity of WHO MDS categories according to 2016 update (in relation to genetic MDS groups). D) Diagram to correctly classify patients in the appropriate MDS group on the basis of individual genomic profile..

Chapter 3

Network Laplacian Cell Dynamics Inference

3.1 Cell Gene Expression Dynamics and Cancer Attractors

In the cell, genes mutually regulate each other expressions. The complex structure of cell regulatory interactions is encoded by the *Gene Regulatory network* (GRN) where a link is a regulatory relationship between two genes [78].

If we assign to each node of the GRN a function, $x_i(t)$, which encodes the level of expression of the gene, as a result of mutual regulations, $x_i(t)$ develops in time. This coordinated evolution of gene expression levels finally leads to a stable pattern of expression of all genes, i.e. a *steady* state [45].

Considering the expression state space, the space of all possible gene expression patterns, a stable state is an *attractor* point and, intuitively, can be thought of as the bottom of a potential well. Evolutionary trajectories which pass through its proximity, i.e. its basin of attraction, eventually converge to the attractor.

A complex gene regulatory network can have a great number of such attractor states, each characterized by a basin of attraction, giving rise to a *landscape* in the state space, referred to as *Epigenetic Landscape*. Stable gene expression patterns are valleys while hills represent unstable states. The potential landscape is determined by the structure of the GRN, which is encoded in the genome and it is the same for all the cells in a given organism.

This view, based on the notion of Weddington's epigenetic landscape, explains how the same genome can give rise to different and 'discrete' cell types. The gene expression pattern of a progenitor multipotent stem cell, can be thought of as a point with high potential in the landscape. It evolves in time, driven by regulatory relationships and noise, until it ends up in one of the possible stable or metastable attractors, corresponding to completely or partially differentiated cell types [46].

In other words, each cell type expresses a gene expression pattern characteristic of a particular attractor, or valley in the landscape, and a switching between cell types is a transition between valleys.

3.1.1 Cancers as Attractors in State Space

Cancer is generally viewed as an evolutionary process. In an organism, cells naturally acquire gene mutations. Some mutations lead to cell death while others are inherited by cell progeny and accumulate during the organism lifetime.

While the majority of acquired mutations are *passengers*, i.e. have no effects on the phenotype, some of them, usually referred to as *drivers*, increase the proliferation capability of mutated cells on normal cells. Usually, this reproductive advantage is limited but, when a cell acquires a mutational burden, that allows it to proliferate autonomously, it can invade tissues.

The hypothesis underlying the view of cancer development as an 'accumulation' of mutations is that a *genotype* maps directly to a *phenotype*. In other words, it assumes implicitly that the global effect of a set of mutations is the 'sum' of the effects of single mutations.

Genome wide studies have revealed a great heterogeneity of cancer mutational patterns and rarely two cancers share the same set of mutations. Following the above hypothesis, one would expect the same heterogeneity in cancer gene expression profiles. From gene expression profiling experiments it turns out, however, that cancers expression patterns are much less heterogeneous than mutational profiles and often, given a cancer, transcriptomes allow the identification of few distinct tumour subtypes [46, 47].

These observations lead to the hypothesis of cancer as an *anomalous cell type*, originating from the transition of a cell in an attractor on the gene expression landscape which is normally unused by physiological cell. Since the GRN, which defines the gene expression landscape, is extremely complex, involving thousands of genes and gene interactions, the existence of attractors that are usually inaccessible but still associated with viable phenotypes is reasonable [46].

The GRN architecture is defined by regulatory relationships among genes. In this context, a mutation corresponds to the removal of a network node or link or to a change in the strength of an interaction. When a mutation occurs, the GRN topology is modified and in turn the regulatory landscape changes. Thus, on one hand, mutations could affect the transition probabilities between attractors or even give rise to new, not physiological attractors. On the other hand, one can think that, completely different mutational patterns could facilitate the transition to the same cancer attractor, giving explanation to the existance of well defined cancer gene expression patterns dispite the high mutational heterogeneity [57, 58, 47].

3.1.2 Reconstructing the Epigenetic Landscape

Given these premises, it would be of great interest to reconstruct the epigenetic landscape which regulate differentiation and dynamics of cells in different tissues to test whether cancer cells can actually be associated to abnormal attractors and to study their relation with 'physiological' cell types.

To this aim, current experimental and theoretical efforts point in two directions. On one side, one aims to reconstruct the gene regulatory network by inferring experimentally pairwise gene regulations. However, the complexity of regulatory interactions is overwhelming and current technologies do not allow to test them systematically [78, 40].

As a consequence, while some gene regulatory circuits have been mapped with sufficient accuracy, the mapping of the whole GRN it is still far from completeness [40]. Nevertheless, this approach has been applied with success to model the dynamics of particular gene circuits, for example to study transition to apoptosis in cancer and healty cell [58].

On the the other side, the epigenetic landscape and its dynamics can be inferred indirectly from single cell gene expression data [77]. The rapidly maturing technology of single cell RNA sequencing (scRNA-seq) allows to capture the gene expression status at the level of individual cells. In other words, it enables to measure the abundance of mRNA for all the genes in the cell exome.

Since RNA sequencing is a destructive measurement, it gives no information of the evolution of the gene expression pattern in time. RNA-seq profiles are static snapshot of a cell state. However, since scRNA-seq is a high-throughput technique, it can be applied systematically to thousands of cells providing a gene expression *cell population snapshot* [86].

Intuitively, we can suppose that observing the steady distribution of an ensemble of cells, that are at different time points of their dynamic evolution, is equivalent to observe the evolution of a single cell in time. In other words, a population snapshot is a distribution in the gene expression state space that have to reflect the cell regulatory mechanism. With this assumption, it is possible to infer, at least partially, the underlying epigenetic landscape.

In recent years, several methods for reconstructing cell dynamics from a static large ensemble of single cell snapshots have been developed [77]. Some of them try to tackle the problem by finding an accurate and scalable dimensionality reduction method that allows, for example, to identify bifurcation in cell development or assign a differentiation pseudo-time [63, 23], while others start from a stochastic physical model of cell dynamics [86].

Nevertheless, the two approaches are intimately related. In particular, a fruitful connection between stochastic dynamical system and dimensionality reduction has been pointed out starting from the work of Belkin and Nyogi which establishes a connection between the optimal embedding of graph and smooth manifold and their laplacian operator [11, 10]. Recently, this connection has been made more precise by the works of Ting et Al. [81] and Weinreb et Al. [86] which discovered a tight relation between general diffusion process on a manifold and the random walk laplacian of the network built from a sampling of a probability distribution defined on the manifold.

In the following, we introduce the mathematical framework to study the epigenetic landscape and the concept of quasi-potential. We than introduce the connection between diffusion processes on manifold and the discrete network laplacian of the neighbourhood network built on a sampling of the diffusion stationary distribution. Finally, we apply these concepts to a simple two dimensional toy model.

3.1.3 Cell Dynamics Description

The qualitative description of cell dynamics and epigenetic landscape can be formalized in the framework of stochastic dynamical systems [46, 85].

If we denote the cell state at time t as $\boldsymbol{x}(t) = (x_1(t), \ldots, x_n(t))$, where $x_i(t)$ is the expression level of the i-th gene at time t then, cell dynamics can generally described by the system:

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(\boldsymbol{x}(t)) \tag{3.1}$$

where $\boldsymbol{F}(\boldsymbol{x}(t))$ is the system driving force and encodes gene expression regulatory relationships, i.e.the gene regulatory network. $f_i(\boldsymbol{x}(t))$ for $i =$

$1, 2 \ldots$ expresses the variation in time of the expression of gene i as a function of the expression of the other genes.

Since gene expression is affected by stochastic fluctuations, one can introduce a noise term $\eta(\boldsymbol{x}, t)$ in equation 3.1:

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(\boldsymbol{x}(t)) + \eta(\boldsymbol{x}, t) \tag{3.2}$$

The amplitude of fluctuations is given by the autocorrelation $\langle \eta(\boldsymbol{x}, t)\eta(\boldsymbol{x}, t') \rangle = 2D\boldsymbol{\Sigma}\delta(t - t')$ where D is the diffusion coefficient and $\boldsymbol{\Sigma}$ is the diffusion tensor, if the noise is isotropic $\boldsymbol{\Sigma} = \boldsymbol{I}$.

The expression vector $\boldsymbol{x}(t)$ is a random variable with probability distribution $p(\boldsymbol{x}, t)$ evolving in time according to the Fokker-Planck equation:

$$\frac{\partial p(\boldsymbol{x}, t)}{\partial t} = \boldsymbol{\nabla} \cdot \boldsymbol{j}(\boldsymbol{x}, t) \tag{3.3}$$

where:

$$\boldsymbol{j}(\boldsymbol{x}, t) = \boldsymbol{F}(\boldsymbol{x})p(\boldsymbol{x}, t) - \boldsymbol{\nabla} \cdot (D\boldsymbol{\Sigma}p(\boldsymbol{x}, t)) \tag{3.4}$$

The stationary probability distribution $p_{st}(\boldsymbol{x}, t))$ satisfy:

$$\boldsymbol{\nabla} \cdot \boldsymbol{j}_{st}(\boldsymbol{x}, t) = 0 \tag{3.5}$$

The divergence of the flux is zero in one of the following cases :

- **detailed balance** with $\boldsymbol{j}_{st} = 0$. In this case, the driving force can be written as a gradient of a potential, assuming $\boldsymbol{\Sigma} = \boldsymbol{I}$ for the sake of clarity, we have:

$$F(\boldsymbol{x}) = D\frac{\nabla p_{st}}{p_{st}} = -\nabla U_{st} \tag{3.6}$$

  where the potential is defined as $U_{st}(\boldsymbol{x}) = -D \cdot log(p_{st}(\boldsymbol{x}))$. In other word, at equilibrium, when no net flux is present in the system, the potential defines the steady state probability.

- **steady state** with $\boldsymbol{j}_{st} \neq 0$. In general, in this case $\boldsymbol{F}$ cannot be derived from a potential. However, making use of the potential U defined above we can decompose $\boldsymbol{F}$ as:

$$\boldsymbol{F} = -\nabla U_{st} + \boldsymbol{F}_{st} = D\frac{\nabla p_{st}}{p_{st}} + \frac{\boldsymbol{j}_{st}}{p_{st}} \tag{3.7}$$

where $\boldsymbol{j}_{st}$ is the steady state flux. Note that, since $\nabla \cdot \boldsymbol{j}_{st} = 0$, the non-gradient part of the force has a curl nature (since in general for a vector field A we have: $\nabla \cdot (\nabla \times A) = 0$).

Decomposition of the Force Field

Given a dynamical system of the form of equation 3.1, in general, when the dimension is greater than one, it is not possible to derive the force field from a potential, i.e. $\boldsymbol{F} \neq -\nabla U$. However, it is possible to decompose $\boldsymbol{F}$ as the gradient of a scalar field plus a remainder term:

$$\boldsymbol{F} = -\nabla \tilde{U} + \boldsymbol{F}_r \tag{3.8}$$

Such decomposition in not unique and we can look for a function $\tilde{U}$ which satisfy specific properties. Since we are interested in transitions between steady states of the system, one can look for a decomposition where $\tilde{U}$ has the role of a *quasi-potential*, in the sense that it encodes the transition barriers between the stable states of the system, while $\boldsymbol{F}_r$ does not enter the efforts needed for the transitions.

In particular, $\tilde{U}$ encodes transition barriers if it satisfies the following criteria [91]:

- characterize the stability of the system. In other words, according to the definition of Lyapunov:

$$\begin{cases} \frac{d\tilde{U}}{dt} < 0 & \text{for } \boldsymbol{x} \neq \boldsymbol{x}^s \\ \frac{d\tilde{U}}{dt} = 0 & \text{for } \boldsymbol{x} = \boldsymbol{x}^s \end{cases} \tag{3.9}$$

- is related to the transition rate between two stable points through the Freidlin–Wentzell action.

It turns out [91] that a decomposition with these properties has to be a *normal* decomposition in the sense that:

$$\nabla \tilde{U} \cdot \boldsymbol{F}_r = 0 \tag{3.10}$$

We are interested in how the decomposition based on the steady state distribution of equation 3.7 relates to the normal decomposition. Substituting 3.7 in the Fokker-Planck equation, we have:

$$\frac{\partial}{\partial t}e^{-\frac{U_{st}}{D}} = D\nabla^2 e^{-\frac{U_{st}}{D}} + D\nabla \cdot \left(\nabla U_{st} e^{-\frac{U_{st}}{D}}\right) - \nabla \cdot \left(\boldsymbol{F}_{st} e^{-\frac{U_{st}}{D}}\right) \tag{3.11}$$

which gives:

$$\nabla U \cdot \boldsymbol{F}_{st} = D\nabla \boldsymbol{F}_{st} \tag{3.12}$$

thus, since the two fields are not orthogonal, the decomposition based on the steady state distribution of the Fokker-Planck equation is not a normal decomposition and in general does not satisfy the conditions required by a quasi-potential. However, when D, i.e. the noise, tends to zero this decomposition tends to the normal decomposition and the scalar field U_{st} can be used as a proxy for the normal quasi-potential. One can think such approximation to apply when the noise in small compared to the driving force.

An ensemble of single cell sequencing snapshot is a sampling from the steady distribution, p_{st}. In the following, we explore the link between the nearest neighbours graph built on a sampling from p_{st} and the diffusion-drift process associated with the potential $U_{st}(\boldsymbol{x}) = -D \cdot log(p_{st}(\boldsymbol{x}))$.

We will see that the markov chain associated to the network random walk laplacian approximates, in the high sampling limit, a the continuous diffusion process with drift term given by the gradient of potential U_{st} providing an approximation for computing transition probability between gene expression states.

3.2 Connections between Diffusion on Networks and Manifolds

3.2.1 Network Laplacians

Given a network $G(V, E)$, directed or undirected, with n vertices and m links there are two alternative ways to represent it in a matrix form, the *adjacency matrix* and the *incidence matrix*.

The *adjacency matrix*, A, is a $n \times n$ square matrix that has elements $A_{ij} = 1$ if nodes i and j are neighbours and zero otherwise.

The *incidence matrix*, ∇ is an $m \times n$ matrix whose rows represent links and columns represent nodes. The j-th element of the l-th row is $\nabla_{lj} = 1$ if l is an incoming link of node j, while $\nabla_{lj} = -1$ if l is an out-coming link of node j.

If f_i for $i \in V$ is a function defined on network nodes, the incidence matrix associates to each link the difference of the values of f at the two end-points of the link:

$$\Delta_l = \sum_i \nabla_{li} f_i \tag{3.13}$$

∇ can be considered a *"discrete differential"* operator on the graph and it appears natural to define the analogous of the continuous Laplacian operator on a graph as follow:

$$\partial \cdot \partial f \longleftrightarrow \boldsymbol{\nabla}^T \cdot \boldsymbol{\nabla} \boldsymbol{f} = L\boldsymbol{f} \tag{3.14}$$

L is called *Unnormalized* or *Combinatorial Laplacian* of the network. Note that it is symmetric and that $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$:

$$L = \boldsymbol{\nabla}^T \cdot \boldsymbol{\nabla} = D - A = \begin{bmatrix} d_1 & -a_{12} & \dots \\ -a_{21} & d_2 & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} \tag{3.15}$$

where D is the diagonal degree matrix. Moreover, we have:

$$(Lf)_i = \sum_{j \sim i} (f_i - f_j) \tag{3.16}$$

that is, the laplacian of a function at a node i is the sum of the differences between the function at i and the function at its neighbours [15].

If the network is weighed, with W the matrix if weights, the combinatorial Laplacian it's easily generalized by the *weighed Laplacian*:

$$L^w = D^w - W \tag{3.17}$$

where $D^w_{ii} = \sum_{i \sim j} w_{ij}$. In the following, if the graph is weighed, we will refer to D^w and L^w with the same symbols used for unweighed networks, D and L.

A matrix closely related to the unnormalized Laplacian is the *random walk* or *Diffusive* Laplacian, L^{rw}, representing the average difference between a node i and its neighbours:

$$(L^{rw} f)_i = \frac{1}{d_i} \sum_{j \sim i} (f_i - f_j) \tag{3.18}$$

For a weighed network, L^{rw} is related to the combinatorial laplacian by the relation:

$$L^{rw} = LD^{-1} = (D - W)D^{-1} = I - WD^{-1} \tag{3.19}$$

The random walk laplacian is similar to the *symmetric random walk Laplacian*, L^{sym}, defined as [24]:

$$L^{sym} = \begin{cases} 1 & i = j \\ -\frac{1}{\sqrt{d_i d_j}} & i \sim j \\ 0 & otherwise \end{cases}$$

It is related to L through the relation:

$$L^{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} \tag{3.20}$$

and is similar to L^{rw}:

$$L^{sym} = D^{\frac{1}{2}} L^{rw} D^{-\frac{1}{2}} \tag{3.21}$$

from this last properties follow that L^{rw} has the same eigenvalue of L^{sym} and if $\boldsymbol{v}$ is an eigenvector of L^{rw} with eigenvalue λ, then $\boldsymbol{q} = D^{-1/2}\boldsymbol{v}$ is an eigenvector of L^{sym} with the same eigenvalue. Since L^{sym} is positive semidefinite it has non negative eigenvalues $0 = \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \ldots$.

3.2.2 Network Laplacians and Embedding

A network embedding is a representation of a network in an m dimensional Euclidean space, $\mathbb{R}^m$, such that nodes that are close on the graph are mapped to neighbouring points on $\mathbb{R}^m$.

The simplest examples of graph embeddings are the graphic representations of network on the paper. They are embeddings in $\mathbb{R}^2$ and different plot layout correspond to different embedding maps.

Given a connected weighed graph $G(V, E, w)$ where V is the set of n vertices, E the set of edges and w are the edge weights and an Euclidean space

$\mathbb{R}^m$ an optimal mapping aims to assigns to each vertex i a set of m coordinates $\boldsymbol{y}_i = (y_i^1, \ldots, y_i^m)$ which minimizes the Euclidean distance between neighbouring nodes [10]:

$$min_{\boldsymbol{y}} \quad \sum_{i,j} |\boldsymbol{y}_i - \boldsymbol{y}_j|^2 w_{ij} \tag{3.22}$$

For each node pair the distance is weighed by the link weight, w_{ij}. If $Y^T = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, from the definition of the weighed network laplacian, $L = D - W$, follows that:

$$\sum_{i,j} |\boldsymbol{y}_i - \boldsymbol{y}_j|^2 w_{ij} = tr(Y^T L Y) \tag{3.23}$$

To remove arbitrary scaling factors, minimization is carried out adding the constraint $Y^T D Y = \boldsymbol{I}$. Since L is positive semidefinite, the matrix Y which minimizes the trace is the matrix whose columns are the m eigenvectors with lowest eigenvalues, given by the generalized eigenvalue problem:

$$L\boldsymbol{y} = \lambda D \boldsymbol{y} \tag{3.24}$$

Since the lowest eigenvector, $\boldsymbol{f}_0$, is the constant vector, the graph embedding is given by the matrix $[\boldsymbol{f}_1, \ldots, \boldsymbol{f}_m]$ where the elements of the i-th row are the embedding coordinates of node i.

Embedding for Smooth Manifolds

It can be shown that the same arguments hold for a smooth manifolds [10].

Given a smooth manifold, $\mathcal{S}$, of dimension s, embedded in $\mathbb{R}^t$ one can look, as in the case of networks, for an optimal embedding of $\mathbb{M}$ in a lower dimensional space.

If $f : \mathcal{S} \to \mathbb{R}$ is a map from the manifold to the real line, two points $\boldsymbol{x}$,$\boldsymbol{z}$ on the manifold are mapped to $f(z)$ and $f(z)$ on $\mathbb{R}$.

An optimal map should minimize $|f(z) - f(z)|$. It turn out that f is optimal when:

$$min_f \quad \int_{\mathcal{S}} |\nabla f(x)|^2 = \int_{\mathcal{S}} \mathcal{L}(f) f \tag{3.25}$$

were the second equivalence follows from the Stockes theorem and $\mathcal{L}$ is the Laplace-Beltami operator of the manifold. $\mathcal{L}$ is positive semidefinite with non-negative eigenvalues and the term on the left is minimum when f is an eigenfunction of the Laplace-Beltrami operator. Since the spectrum of $\mathcal{L}$

is discrete, the embedding for $\mathcal{S}$ in an euclidean space of dimension m, is given by the first m non-constant eigenfunctions (i.e., relative to eigenvalues $\lambda_i > 0$):

$$\boldsymbol{x} \to (f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})) \tag{3.26}$$

3.2.3 Graph Approximation of the Laplace-Beltrami Operator on a Manifold

If we have a set of data points $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})$ with $\boldsymbol{x_i} \in \mathbb{R}^t$ and we assume that they lie on a submanifold $\mathcal{S}$ of the space $\mathbb{R}^t$ we could ask how to find an optimal embedding of the submanifold from sampled data. A solution is to build a network whose nodes are data points and such that the network laplacian, $L = D - W$, approximates the Laplace-Beltrami on the manifold.

Belkin and Nyogi [10, 11] proposed the following connection based on the heat equation on a manifold:

$$\frac{\partial p(\boldsymbol{x}, \boldsymbol{t})}{\partial t} + \mathcal{L}p(\boldsymbol{x}, \boldsymbol{t}) = 0 \tag{3.27}$$

assuming $p(\boldsymbol{x}, 0) = f(\boldsymbol{x})$ the solution, can be written in terms of the of the *heat kernel* (Green function for eq. 3.27), $K_t(\boldsymbol{x}, \boldsymbol{y})$:

$$p(\boldsymbol{x}, t) = \int_{\mathcal{S}} K_t(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{y}) \tag{3.28}$$

for small t, K_t can be approximated by a Gaussian on the manifold:

$$K_t(x, y) \approx (4\pi t)^{\frac{-m}{2}} e^{\frac{-\|x-y\|^2}{4t}} \tag{3.29}$$

and:

$$p(\boldsymbol{x}, \boldsymbol{t}) \approx \int_{\mathcal{S}} (4\pi t)^{\frac{-m}{2}} e^{\frac{-\|x-y\|^2}{4t}} f(y) dy \tag{3.30}$$

The Laplace-Beltrami operator can thus be approximated as:

$$\mathcal{L}f(x) = -\frac{\partial p}{\partial t} \approx \frac{1}{t} \left[f(x) - (4\pi t)^{\frac{m}{2}} \int_{\mathcal{S}} e^{\frac{-\|x-y\|^2}{4t}} f(y) dy \right] \tag{3.31}$$

The above equation gives an explicit representation of the action of the laplacian on the function f.

Given the data points,$(\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})$, supposed to be a sampling from the manifold, the above expression can be approximated around the point $\boldsymbol{x_i}$ by approximating the integral on $\mathcal{S}$ with the 'average' on the first k nearest neighbours of $\boldsymbol{x_i}$ (note that the contribution of points far from $\boldsymbol{x_i}$ decrease exponentially) :

$$\mathcal{L}f(\boldsymbol{x_i}) \approx \frac{1}{t}\left[f(\boldsymbol{x_i}) - \frac{1}{k}(4\pi t)^{\frac{m}{2}} \sum_{x_j \in [k]_i} e^{\frac{-\|x_i - x_j\|^2}{4t}} f(\boldsymbol{x_j})\right] \tag{3.32}$$

where the sum is on the set of nearest neighbours of $\boldsymbol{x_i}$, denoted $[k]_i$.

Since f can be any function, we can take $f(\boldsymbol{x_i}) = Const$ such that the Laplacian is zero. Putting $\frac{1}{k}(4\pi t)^{\frac{m}{2}} = \alpha$, from equation 3.32 we have:

$$\alpha = \left[\sum_{x_j \in [k]_i} e^{\frac{-\|x_i - x_j\|^2}{4t}}\right]^{-1} \tag{3.33}$$

If we consider now the weighed kNN network of data points with W being the weight matrix and $f(\boldsymbol{x_i}) = f_i$ the function on the i-th node, such that $\boldsymbol{f} = (f(\boldsymbol{x}_1), \ldots, (f\boldsymbol{x}_n))$, the heat equation on the network, reads:

$$\frac{d\boldsymbol{f}}{dt} = -L\boldsymbol{f} \tag{3.34}$$

where:

$$(L\boldsymbol{f})_i = f_i - \sum_j w_{ij} f_j \tag{3.35}$$

where $(Lf)_i$ is the i-th elements of the image of f through L, the network Laplacian. We can see the following correspondence:

$$w_{ij} \iff \frac{1}{\alpha} e^{\frac{-\|x_i - x_j\|^2}{4t}} \tag{3.36}$$

3.2.4 Connection Between Laplace-Beltrami and L^{rw}

Random Walk on Networks

A random walk on a network is defined as follow. Let $G = (V, E, W)$ be a network where W is the weight matrix. If at the t-th time step a particle is at at a node v, it moves to a neighbouring site k of v with probability $\pi_{ij} = \frac{w_{vk}}{d_v}$, where $d_v = \sum_j w_{vj}$ is the degree of the node.

We denote by p_i^t the probability of being at i at time t. The matrix of transition probabilities (π^T) is given by:

$$\pi_{ij}^T = \begin{cases} \frac{w_{ij}}{d_i} & i \sim j \\ 0 & : otherwise \end{cases}$$

If D is the diagonal degree matrix and W is symmetric, the transition matrix $\Pi = WD^{-1}$. Note that Π is a stochastic matrix:

$$\sum_j \pi_{ij} = 1 \quad \text{and} \quad \pi_{ij} \geq 0 \tag{3.37}$$

Defining $\boldsymbol{p}^t = (p_1^t, \ldots, p_n^t)$ the vector of node probabilities at time t, its the evolution is given by:

$$\boldsymbol{p}(t + \Delta t) = \Pi \boldsymbol{p}(t) \tag{3.38}$$

Starting from the initial distribution $\boldsymbol{p}^0$, after a time $t = k\Delta t$ we have:

$$\boldsymbol{p}(t = k\Delta t) = \Pi^k \boldsymbol{p}^0 \tag{3.39}$$

with the stationary distribution:

$$\boldsymbol{p}_{st} = \frac{\boldsymbol{d}}{2\sum_{ij} w_{ij}} \tag{3.40}$$

$\boldsymbol{p}_{st}$ is proportional to the degree vector, $\boldsymbol{d} = (d_1, \ldots, d_n)$, and does not depend on the initial distribution $\boldsymbol{p}^0$.

Lazy Random Walk

To derive the continuous time random equation, it is convenient to consider a variation of random walk sometimes referred to as *lazy random walk*. At each step, with probability $1 - \beta$, a particle stays at the current vertex and with probability β it jumps to another node. The evolution of the probability is given by:

$$\boldsymbol{p}(t + \Delta t) = \beta \Pi \boldsymbol{p}(t) + (1 - \beta) \boldsymbol{I} \boldsymbol{p}(t) = \tilde{\Pi} \boldsymbol{p}(t) \tag{3.41}$$

with:

$$\tilde{\pi}_{ij} = \begin{cases} 1 - \beta & i = j \\ \pi_{ij}\beta & i \neq j \end{cases}$$

It is possible to show that equations 3.38 and 3.41 have the same stationary distribution.

Continuous Time Random Walk

Let's derive the continuous time version of the evolution equation. In the limit $\Delta t \to 0$ we expect that the probability for a particle of changing state tends to zero. Starting from eq. 3.41 and assuming the expansion for the coefficient $\beta = \beta_0 \Delta t + o(\Delta t)$, the transition probabilities become:

$$\tilde{\pi}_{ij}(\Delta t \to 0) = \begin{cases} 1 - \beta_0 \Delta t + o(\Delta t) & i = j \\ \pi_{ij}\beta_0 \Delta t + o(\Delta t) & i \neq j \end{cases}$$

with the condition:

$$\sum_i \tilde{\pi}_{ij}(t) = 1 \tag{3.42}$$

The probability of being at node i at time $t + \Delta t$ can be written as:

$$p_i(t + \Delta t) = \sum_j p_j(t)\pi_{ij}\beta_0 \Delta t + (1 - \beta_0 \Delta t)p_i(t) \tag{3.43}$$

which gives:

$$\frac{p_i(t + \Delta t) - p_i(t)}{\Delta t} = \beta_0 \sum_j p_j \pi_{ij} - p_i = \beta_0 \sum_j \left[\pi_{ij} - \delta_{ij}\right] p_j(t) \tag{3.44}$$

Recalling the definition of random walk Laplacian of the network:

$$L^{rw} = (D - W)D^{-1} = \boldsymbol{I} - \Pi \to l^{rw}_{ij} = \delta_{ij} - \pi_{ij} \tag{3.45}$$

Connection to Manifold Laplacian

Introducing the vectorial notation $\boldsymbol{p}(t) = (p_1, \ldots, p_n)$, in the limit $\Delta t \to 0$, we have:

$$\frac{d\boldsymbol{p}(t)}{dt} = -\beta_0 L^{rw} \boldsymbol{p}(t) \tag{3.46}$$

The action of L^{rw} on a vector $\boldsymbol{f}$ is:

$$(L^{rw}\boldsymbol{f})_i = f_i - \sum_j \pi_{ij} f_j \tag{3.47}$$

where $(L^{rw}f)_i$ is the i-th elements of the image of f through L^{rw}. Recalling the action of the Laplacian of a manifold on a function f for $t \to 0$:

$$\mathcal{L}f(\boldsymbol{x_i}) \approx \frac{1}{t}\left[f(\boldsymbol{x_i}) - \frac{\sum_{x_j \in [k]_i} e^{\frac{-\|x_i - x_j\|^2}{4t}} f(\boldsymbol{x_j})}{\alpha_i} \right] \tag{3.48}$$

where:

$$\alpha_i = \left[\sum_{x_j \in [k]_i} e^{\frac{-\|x_i - x_j\|^2}{4t}} \right]^{-1} \tag{3.49}$$

Comparing equations 3.47 and 3.48 we can put:

$$\pi_{ij} = \frac{e^{\frac{-\|x_i - x_j\|^2}{4t}}}{\alpha_i} \tag{3.50}$$

It turn out that the right hand side satisfies the properties required from the transition matrix of being time independent and stochastic. From the definition of α_i we have:

- time independence,

$$\pi_{ij} = \frac{e^{\frac{1}{4t}} e^{-\|x_i - x_j\|^2}}{e^{\frac{1}{4t}} \sum_{x_j \in [k]_i} e^{-\|x_i - x_j\|^2}} = \frac{e^{-\|x_i - x_j\|^2}}{\sum_{x_j \in [k]_i} e^{-\|x_i - x_j\|^2}} \tag{3.51}$$

- column normalization (stochastic matrix),

$$\sum_j \pi_{ij} = \frac{1}{\sum_{x_j \in [k]_i} e^{-\|x_i - x_j\|^2}} \sum_{x_j \in [k]_i} e^{-\|x_i - x_j\|^2} = 1 \tag{3.52}$$

Moreover, recalling that $\pi_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}$, where w_{ij} are the is the link weights we have:

$$w_{ij} = e^{-\|x_i - x_j\|^2} \tag{3.53}$$

This observations highlights that, given n points $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ from a manifold $\mathcal{S}$ embedded in $\mathbb{R}^t$ and the kNN network of the points with link weights given by equation 3.53 the *diffusive laplacian*, L^{rw}, of the network approximate the Laplace-Beltrami operator on the manifold.

Decomposition along Diffusion Eigenfunction

The solution of the diffusion equation 3.46 is given by:

$$\boldsymbol{p}(t) = e^{L^{rw}t}\boldsymbol{p}(0) \tag{3.54}$$

where we put $\beta_0 = 1$ for convenience. Considering the eigenvalue problem for L^{rw}:

$$L^{rw}\boldsymbol{v} = \lambda\boldsymbol{v} \tag{3.55}$$

since L^{rw} is similar to L^{sym} which is positive semidefinite, it has non negative eigenvalues $0 = \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \dots$

If we write $\boldsymbol{p}(t)$ and $\boldsymbol{p}(0)$ as linear combinations of the eigenvectors of L^{rw} we can decompose the solution of 3.54 as:

$$\boldsymbol{p}(t) = \sum_i a_i(t)\boldsymbol{v}_i \tag{3.56}$$

$$\boldsymbol{p}(0) = \sum_i a_i(0)\boldsymbol{v}_i \tag{3.57}$$

From equation 3.54 and 3.56 we have:

$$\frac{da_i(t)}{dt} = -\lambda_i a_i(t) \tag{3.58}$$

which has the solution:

$$a_i(t) = e^{-\lambda_i t}a_i(0) \tag{3.59}$$

Equation 3.56 can be rewritten as:

$$\boldsymbol{p}(t) = a_0(0)e^{-\lambda_0 t}\boldsymbol{v}_0 + a_1(0)e^{-\lambda_1 t}\boldsymbol{v}_1 + \dots \tag{3.60}$$

If the graph is connected then $\lambda_i > 0$ for all $i > 0$ and the stationary solution of equation 3.60 for $t \to \infty$ is unique and is proportional to the eigenvector with null eigenvalue.

The components of the solution along the eigenvectors with positive eigenvalues are *transient states* since they tend to zero as t approaches infinity. The decay time constant of the component along the eigenvector $\boldsymbol{v}_i$ is proportional to the inverse of λ_i.

3.2.5 Deeper Relation Between Random Walk Laplacian and Diffusion on Manifold

A deeper relation between the random walks on network and diffusive processes on a manifold $\mathcal{S}$ has been recently established by Ting et Al. [81]. In particular, they show that the random walk laplacian of a neighbourhood graph built on points sampled from a distribution defined on a smooth manifold, converges, in the high sampling limit, to the generator of a specific diffusion process on the manifold. Moreover, the establish a connection between the form of the kernel function used to built the graph and the drift and diffusion term of the continuous diffusion process. We resume briefly their results below.

Let $\mathcal{S}$ be a smooth m-dimensional manifold in $\mathbb{R}^b$ and p a density defined on the manifold. Given a set of n points, $\{x_i\}_n$, sampled i.i.d. from p, one can build a neighbourhood graph assuming a general kernel function, $K_n(x, y)$:

$$K_n(x,y) = \omega_x^{(n)}(y) K_0 \left(\frac{\|y - x\|}{h r_x^{(n)}(y)} \right) \tag{3.61}$$

where $\omega_x(y)$ assigns the weight to the link between points located at x and y and K_0 is in general a non-smooth kernel function. $r_x(y)$ is a bandwidth function, which defines the maximum distance for two nodes to be neighbours in the network. For example, for a kNN graph $r_x(y)$ is the distance to the k-th neighbour.

The random walk laplacian, L^{rw}, of the resulting neighbourhood graph tends, in the high sampling limit, to the generator, G, of diffusion process on the manifold:

$$\begin{gathered} dX_t = \mu(\boldsymbol{x})dt + \sigma(X_t)dW_t \\ -L_n^{rw} f \xrightarrow[n \to \infty]{} Gf \end{gathered}$$

where the diffusion and drift terms of the process are given by:

$$\mu(x) = r_x(x)^2 \left(\frac{\nabla p(x)}{p(x)} + \frac{\nabla \omega(x)}{\omega(x)} + (m+2) \frac{\dot{r}_x(x)}{r_x(x)} \right) \tag{3.62}$$

$$\sigma(x)\sigma(x)^T = r_x(x)^2 \boldsymbol{I} \tag{3.63}$$

In other words, the continuous time Markov chain associated to the network random walk laplacian (see equation 3.45) approximates, when the number of network nodes grows, to the above diffusion process.

It is of interest to apply these results to specific network constructions, in particular kNN and r-neighbourhood network:

- in r-neighbourhood network $K_0 = I(|x| < r)$. The bandwidth function is thus $r_x(y) = r = Const$ and $\dot{r}_x(y) = 0$, Moreover, assuming a constant weight function $\nabla\omega(x) = 0$ and we have:

$$\mu(x) = \frac{\nabla p(x)}{p(x)} \tag{3.64}$$

$$\sigma(x)\sigma^T(x) = \boldsymbol{I} \tag{3.65}$$

3.3 Inference of Cell Regulatory Landscape

We have seen in section 3.1.3 that cell regulatory dynamics can be modelled as a dynamical system with noise:

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(\boldsymbol{x}(t)) + \eta(\boldsymbol{x}, t) \tag{3.66}$$

If p_{st} is the stationary distribution of the Fokker-Plank equation associated with the above system, than the potential:

$$U_{st} = -D \cdot log(p_{st}) \tag{3.67}$$

is, in the zero noise limit, a quasi-potential, which encodes the transition probability between the attractor valleys of the landscape defined by $\boldsymbol{F}$.

An ensemble of single cell measurements, can be viewed as a sampling from the stationary distribution p_{st}. As shown in equation 3.65, the Markov chain associated to the random walk laplacian of the r-neighbourhood graph of data points approximates a diffusion process on the potential $U_{st} = -D \cdot log(p_{st})$, in the high sampling limit.

Recalling equation 3.60, the eigenvector of L_{rw} with zero eigenvalue is the stationary distribution of the Markov chain defined by L_{rw}, while eigenvectors with non-zero eigenvalue are associated to diffusive transient states. Studying the eigenvalue problem associated to L_{rw} is thus a way to characterize the regulatory landscape from a 'single cell' population snapshot.

Below we illustrate these results with a toy model example.

3.3.1 Toy Model

We studied a two dimensional toy model of the regulatory landscape, described by the following equations:

$$\frac{dx_1}{dt} = f_1 + \eta_1 = -\Big(2x^3 + 2y^3 - 9(x+y) - 1\Big) + \eta_1 \tag{3.68}$$
$$\frac{dx_2}{dt} = f_2 + \eta_2 = -\Big(2x^3 - 2y^3 - 11(x-y) + 1\Big) + \eta_2 \tag{3.69}$$

where η_i are the noise terms. The normal decomposition of the system is given by [45]:

$$U^{norm} = \left(-5(x^2+y^2) + \frac{1}{2}(x^4+y^4) + xy + x\right) \tag{3.70}$$
$$F_x^{norm} = 10x - 2y^3 - x \tag{3.71}$$
$$F_y^{norm} = -10x + 2y^3 + y + 1 \tag{3.72}$$

In figure 3.1 is reported the potential $U^{norm}(x,y)$. The system has four potential wells of different depth separated by four saddle points, while the remainder force is, at each point, perpendicular to the gradient of the potential.

Simulation and Landscape Inference

Since we would to infer the property of the potential from a sampling of the stationary distribution, we simulated the stochastic evolution of the system 3.69 with a simple stochastic Euler scheme. We carried out 100 simulations ($2 \cdot 10^6$ time steps) with random starting points in the range $x = [-3,3], y = [-3,3]$. For each simulation we discarded the first 10^5 time steps.

We sampled $n = 600$ random points from the simulated stationary distribution and we built the r-neighbourhood graph (the choice of the bandwidth r does not affect results significantly).

The spectrum of the random walk laplacian (figure 3.2) highlights a spectral gap after the lowest four eigenvalues. Since the magnitude of an eigenvalue gives the exponential decay time constant of the relative eigenvector, eigenvector with lowest eigenvalues are related to stable ($\lambda_0 = 0$) and metastable states ($\lambda_1, \lambda_2, \lambda_3$) of the system.

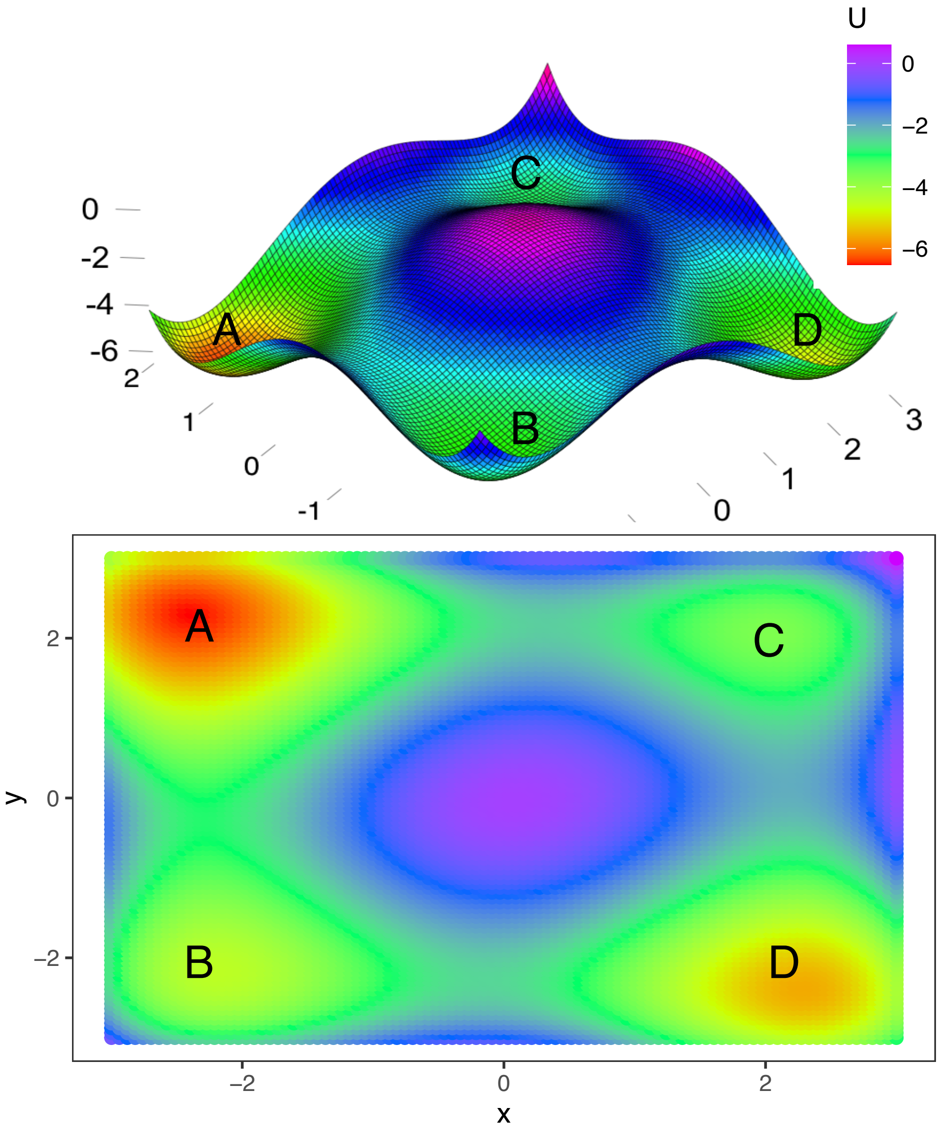


FIGURE 3.1: Quasi-potential U^{norm} resulting from the normal decomposition of the the toy model system (eq. 3.69). It has four attractors located approximately at $A = (-2, 2)$,$B = (-2, -2)$,$C = (2, -2)$,$D = (2, 2)$ and four saddle points between them.

From figure 3.3, we see that for $\lambda = 0$ the eigenvector retrieves the potential U^{norm}. Moreover, the eigenvectors relative to the lowest non zero eigenvalues are related to the four landscape valleys. In particular, the two deeper valleys (more stable states) are associated to a lower eigenvalue (λ_1).

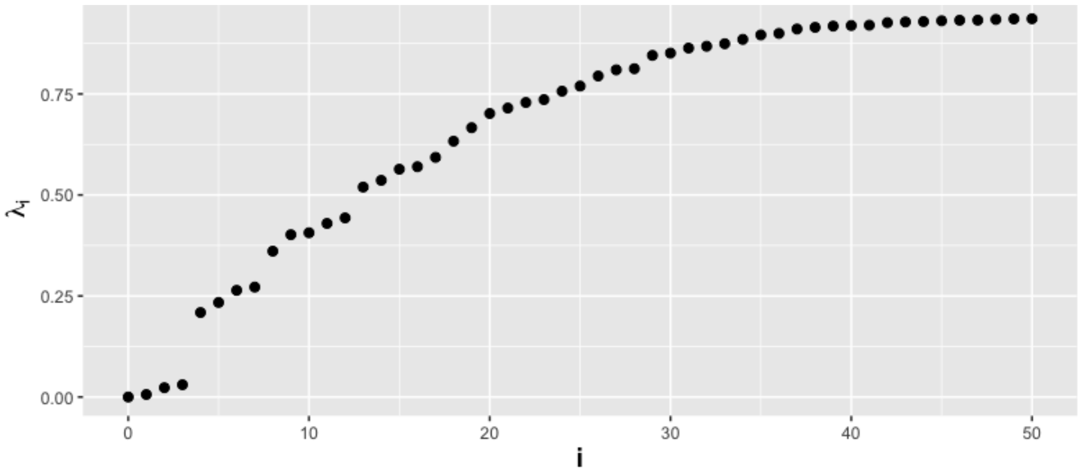


FIGURE 3.2: Spectrum of the random walk laplacian L_{rw} of the r-neighbourhood network of points sampled from the simulated stationary distribution of the toy model system.

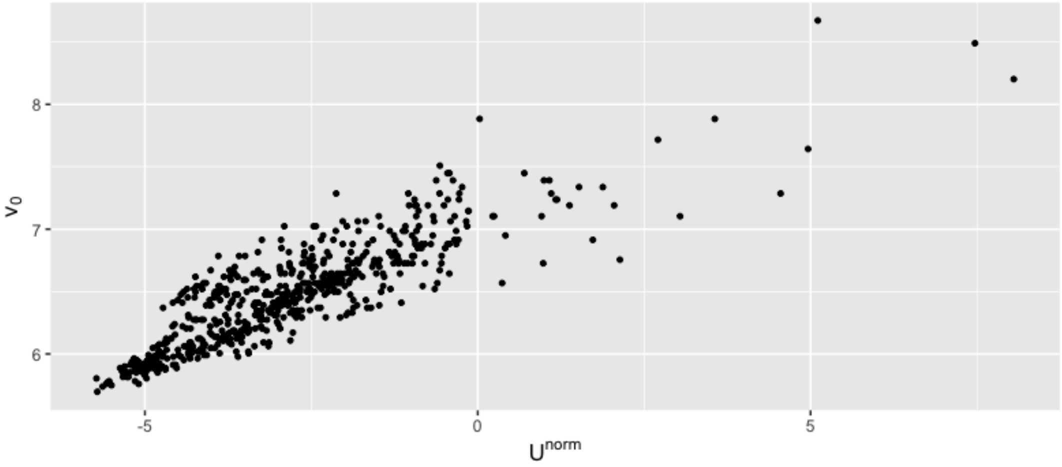


FIGURE 3.3: Logarithm of the eigenvector with zero eigenvalue versus the driving potential U^{norm}.

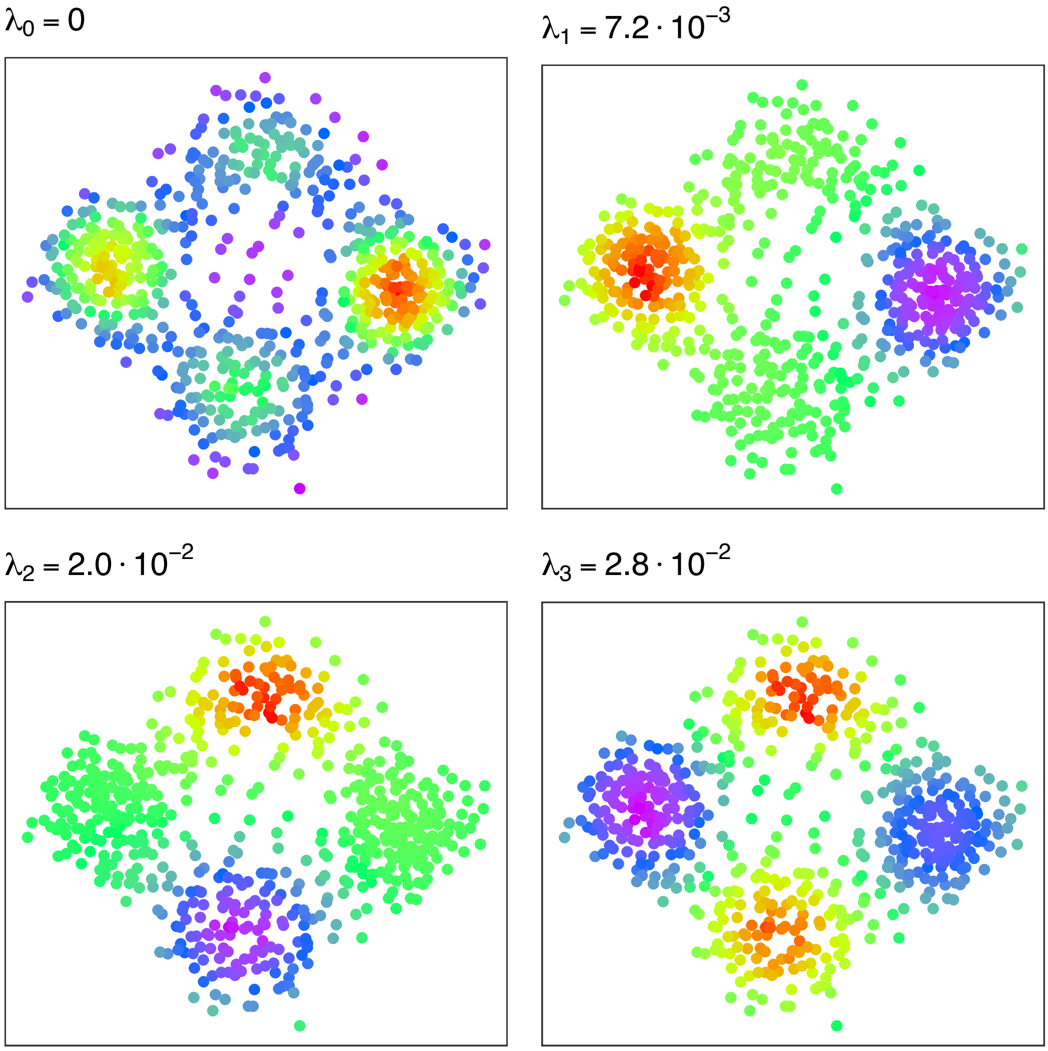


FIGURE 3.4: Eigenvectors with lowest eigenvalues of the random walk laplacian of the *r*-neighbourhood network. The color of each node of the network gives the eigenvector elements for that node (red high, purple low). Network was plotted with Fruchterman-Reingold force-directed layout.

Conclusions

In this work we presented some methods to address the problem of cancer mutational heterogeneity and its relation with the genotype-phenotype mapping form different perspectives.

In the first part, we tackled the problem of relating cancer mutational patterns to higher level biological functions. To this end, in the first chapter, we studied the resilience of the cell macromolecular interaction network treating mutations as network perturbations. We found that all the interactome reconstructions considered, are much more brittle to tumour mutations in comparison to random gene mutations. Even though the interactome is far from capturing many expect of cell dynamics, this result highlights the 'cooperative' nature of cancer mutations and the relevance of their characterization with respect to higher biological levels.

In the second chapter, we presented non parametric models in the context of unsupervised detection of cancer molecular subtypes, highlighting their flexibility in modelling complex data structures. As a case study, we applied these methods to a cohort of 2043 patients affected by Myelodysplastic Syndromes finding a meaningful correspondence between the unsupervised mutational classes and biological functions relevant to the development of the the disease.

The second part of the work was inspired by the work of Huang and Kauffman [46, 47] in which they point out the similarities between the development of cancer and normal cells suggesting the hypothesis of cancer as an abnormal cell type. In there view, mutations act by modifying the cell epigenetic landscape and facilitating the transition of a cell to a cancer attractor. The advent of single cell sequencing made this hypothesis testable by experiments. To this end, in the last chapter we reviewed the relations between mathematical models of the epigenetic landscape and network based dimensionality reduction methods and we show the ability of laplacian based methods to infer the quasi potential of two dimensional toy model of the epigenetic landscape. Since these methods are computationally scalable, future work will be their application to real high dimensional single cell sequencing datasets of both normal and cancer cell to test whether cancer can effectively be associated to abnormal attractor of the cell regulatory landscapes.

Bibliography

[1] *A Gentle Introduction to the Dirichlet Process, the Beta Process and Bayesian Nonparametrics*. Lecture Notes, 2015.

[2] Lionel Adès, Raphael Itzykson, and Pierre Fenaux. "Myelodysplastic syndromes". In: *The Lancet* 383.9936 (2014), pp. 2239–2252. DOI: 10.1016/S0140-6736(13)61901-7. URL: https://doi.org/10.1016/S0140-6736(13)61901-7.

[3] Joaquim Aguirre-Plans et al. "GUILDify v2.0: A Tool to Identify Molecular Networks Underlying Human Diseases, Their Comorbidities and Their Druggable Targets". In: *Journal of Molecular Biology* 431.13 (2019), pp. 2477 –2484. ISSN: 0022-2836. DOI: https://doi.org/10.1016/j.jmb.2019.02.027. URL: http://www.sciencedirect.com/science/article/pii/S0022283619301172.

[4] S. E. Ahnert. "Structural properties of genotype-phenotype maps". In: *Journal of The Royal Society Interface* 14.132 (2017), p. 20170275. DOI: 10.1098/rsif.2017.0275. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2017.0275.

[5] Gregorio Alanis-Lobato, Miguel A. Andrade-Navarro, and Martin H. Schaefer. "HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks". In: *Nucleic Acids Research* 45.D1 (2016), pp. D408–D414. ISSN: 0305-1048. DOI: 10.1093/nar/gkw985. URL: https://doi.org/10.1093/nar/gkw985.

[6] Daniel A. Arber et al. "The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia". In: *Blood* 127.20 (May 2016), pp. 2391–2405. ISSN: 0006-4971. DOI: 10.1182/blood-2016-03-643544. URL: https://doi.org/10.1182/blood-2016-03-643544.

[7] Euan A. Ashley. "Towards precision medicine". In: *Nature Reviews Genetics* 17.9 (2016), pp. 507–522. DOI: 10.1038/nrg.2016.86. URL: https://doi.org/10.1038/nrg.2016.86.

[8] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. "Network medicine: a network-based approach to human disease". In: *Nature Reviews Genetics* 12.1 (2011), pp. 56–68. DOI: 10.1038/nrg2918. URL: https://doi.org/10.1038/nrg2918.

[9] Barabási, Albert-László and Albert, Réka. "Emergence of Scaling in Random Networks". In: *Science* 286.5439 (1999), pp. 509–512. DOI: 10.1126/science.286.5439.509. URL: https://science.sciencemag.org/content/286/5439/509.

[10] M. Belkin and P. Niyogi. "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation". In: *Neural Computation* 15.6 (2003), pp. 1373–1396. ISSN: 0899-7667. DOI: 10.1162/089976603321780317.

[11] Mikhail Belkin and Partha Niyogi. "Towards a theoretical foundation for Laplacian-based manifold methods". In: *Journal of Computer and System Sciences* 74.8 (2008), pp. 1289 –1308. ISSN: 0022-0000. DOI: https://doi.org/10.1016/j.jcss.2007.08.006. URL: http://www.sciencedirect.com/science/article/pii/S0022000007001274.

[12] Matteo Bersanelli et al. "Frailness and resilience of gene networks predicted by detection of co-occurring mutations via a stochastic perturbative approach". In: *Scientific Reports* 10.1 (2020), p. 2643. DOI: 10.1038/s41598-020-59036-w. URL: https://doi.org/10.1038/s41598-020-59036-w.

[13] Matteo Bersanelli et al. "Methods for the integration of multi-omics data: mathematical aspects". In: *BMC Bioinformatics* 17.2 (2016), S15. DOI: 10.1186/s12859-015-0857-9. URL: https://doi.org/10.1186/s12859-015-0857-9.

[14] Matteo Bersanelli et al. "Network diffusion-based analysis of high - throughput data for the detection of differentially enriched modules". In: *Scientific Reports* 6.1 (2016), p. 34841. DOI: 10.1038/srep34841. URL: https://doi.org/10.1038/srep34841.

[15] Stadler Peter F. Biyikoglu Turker Leydold Josef. *Laplacian Eigenvectors of Graphs*. Lecture Notes in Mathematics. Springer, 2007.

[16] David M. Blei. "Probabilistic Topic Models". In: *Communications of the ACM* 55 (2012). DOI: 10.1145/2133806.2133826.

[17] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3 (2003), pp. 993–1022. ISSN: 1532-4435.

[18] Anna D. Broido and Aaron Clauset. "Scale-free networks are rare". In: *Nature Communications* 10.1 (2019), p. 1017. DOI: 10.1038/s41467-019-08746-5. URL: https://doi.org/10.1038/s41467-019-08746-5.

[19] Garth R. Brown et al. "Gene: a gene-centered information resource at NCBI". In: *Nucleic Acids Research* 43 (2014), pp. D36–D42. ISSN: 0305-1048. DOI: 10.1093/nar/gku1055. URL: https://doi.org/10.1093/nar/gku1055.

[20] Rebecca A. Burrell et al. "The causes and consequences of genetic heterogeneity in cancer evolution". In: *Nature* 501.7467 (2013), pp. 338–345. DOI: 10.1038/nature12625. URL: https://doi.org/10.1038/nature12625.

[21] Hannah Carter, Matan Hofree, and Trey Ideker. "Genotype to phenotype via network analysis." In: *Curr Opin Genet Dev* 23.6 (2013), pp. 611–621. DOI: 10.1016/j.gde.2013.10.003.

[22] Mario Cazzola, Matteo G. Della Porta, and Luca Malcovati. "The genetic basis of myelodysplasia and its clinical relevance". In: *Blood* 122.25 (2013), pp. 4021–4034. ISSN: 0006-4971. DOI: 10.1182/blood-2013-09-381665. URL: https://doi.org/10.1182/blood-2013-09-381665.

[23] Huidong Chen et al. "Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM". In: *Nature Communications* 10.1 (2019), p. 1903. DOI: 10.1038/s41467-019-09670-4. URL: https://doi.org/10.1038/s41467-019-09670-4.

[24] Fan R. K. Chung. *Spectral Graph Theory*. CBMS Regional Conference Series in Mathematics. AMS, 1997.

[25] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. "Power-Law Distributions in Empirical Data". In: *SIAM Review* 51.4 (2009), pp. 661–703. DOI: 10.1137/070710111. eprint: https://doi.org/10.1137/070710111. URL: https://doi.org/10.1137/070710111.

[26] Reuven Cohen et al. "Resilience of the Internet to Random Breakdowns". In: *Phys. Rev. Lett.* 85 (21 2000), pp. 4626–4628. DOI: 10.1103/PhysRevLett.85.4626. URL: https://link.aps.org/doi/10.1103/PhysRevLett.85.4626.

[27] Lenore Cowen et al. "Network propagation: a universal amplifier of genetic associations". In: *Nature Reviews Genetics* 18.9 (2017), pp. 551–562. DOI: 10.1038/nrg.2017.38. URL: https://doi.org/10.1038/nrg.2017.38.

[28] James Cussens and Mark Bartlett. *GOBNILP*.

[29] Jishnu Das and Haiyuan Yu. “HINT: High-quality protein interactomes and their applications in understanding human disease”. In: *BMC Systems Biology* 6.1 (2012), p. 92. DOI: 10.1186/1752-0509-6-92. URL: https://doi.org/10.1186/1752-0509-6-92.

[30] M G Della Porta et al. “Minimal morphological criteria for defining bone marrow dysplasia: a basis for clinical implementation of WHO classification of myelodysplastic syndromes”. In: *Leukemia* 29.1 (2015), pp. 66–75. DOI: 10.1038/leu.2014.161. URL: https://doi.org/10.1038/leu.2014.161.

[31] M G Della Porta et al. “Validation of WHO classification-based Prognostic Scoring System (WPSS) for myelodysplastic syndromes and comparison with the revised International Prognostic Scoring System (IPSS-R).” In: *Leukemia* 29.7 (2015), pp. 1502–1513. DOI: 10.1038/leu.2015.55. URL: https://doi.org/10.1038/leu.2015.55.

[32] Linton C. Freeman. “A Set of Measures of Centrality Based on Betweenness”. In: *Sociometry* 40.1 (1977), pp. 35–41. ISSN: 00380431. URL: http://www.jstor.org/stable/3033543.

[33] Javier Garcia-Garcia et al. “Biana: a software framework for compiling biological interactions and analyzing networks”. In: *BMC Bioinformatics* 11.1 (2010), p. 56. DOI: 10.1186/1471-2105-11-56. URL: https://doi.org/10.1186/1471-2105-11-56.

[34] Moritz Gerstung et al. “Precision oncology for acute myeloid leukemia using a knowledge bank approach”. In: *Nature Genetics* 49.3 (2017), pp. 332–340. DOI: 10.1038/ng.3756. URL: https://doi.org/10.1038/ng.3756.

[35] Zoubin Ghahramani. “Bayesian non-parametrics and the probabilistic approach to modelling”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371.1984 (2013). DOI: 10.1098/rsta.2011.0553. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2011.0553.

[36] Susan Dina Ghiassian, Jörg Menche, and Albert-László Barabási. “A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome”. In: *PLOS Computational Biology* 11.4 (2015),

pp. 1–21. DOI: 10.1371/journal.pcbi.1004120. URL: https://doi.org/10.1371/journal.pcbi.1004120.

[37] Colin S. Gillespie. "Fitting Heavy Tailed Distributions: The poweRlaw Package". In: *Journal of Statistical Software* 64.2 (2015), pp. 1–16.

[38] Peter L. Greenberg et al. "Revised International Prognostic Scoring System for Myelodysplastic Syndromes". In: *Blood* 120.12 (2012), pp. 2454–2465. ISSN: 0006-4971. DOI: 10.1182/blood-2012-03-420489. URL: https://doi.org/10.1182/blood-2012-03-420489.

[39] Jacob Grinfeld et al. "Classification and Personalized Prognosis in Myeloproliferative Neoplasms". In: *New England Journal of Medicine* 379.15 (2018), pp. 1416–1430. DOI: 10.1056/NEJMoa1716614. URL: https://doi.org/10.1056/NEJMoa1716614.

[40] Heonjong Han et al. "TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions". In: *Nucleic Acids Research* 46.D1 (2017), pp. D380–D386. ISSN: 0305-1048. DOI: 10.1093/nar/gkx1013. URL: https://doi.org/10.1093/nar/gkx1013.

[41] Pierre C. Havugimana et al. "A Census of Human Soluble Protein Complexes". In: *Cell* 150.5 (2012). DOI: 10.1016/j.cell.2012.08.011. URL: https://doi.org/10.1016/j.cell.2012.08.011.

[42] Marco Y. Hein et al. "A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances". In: *Cell* 163.3 (2015), pp. 712 –723. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2015.09.053. URL: http://www.sciencedirect.com/science/article/pii/S0092867415012702.

[43] Matan Hofree et al. "Network-based stratification of tumor mutations". In: *Nature Methods* 10.11 (2013), pp. 1108–1115. DOI: 10.1038/nmeth.2651. URL: https://doi.org/10.1038/nmeth.2651.

[44] Justin K. Huang et al. "Systematic Evaluation of Molecular Networks for Discovery of Disease Genes". In: *Cell Systems* 6.4 (2018), pp. 484 –495. ISSN: 2405-4712. DOI: https://doi.org/10.1016/j.cels.2018.03.001. URL: http://www.sciencedirect.com/science/article/pii/S2405471218300954.

[45] Sui Huang. "The molecular and mathematical basis of Waddington's epigenetic landscape: A framework for post-Darwinian biology?" In: *BioEssays* 34.2 (2012), pp. 149–157. DOI: https://doi.org/10.1002/

bies.201100031. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.201100031.

[46] Sui Huang, Ingemar Ernberg, and Stuart Kauffman. "Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective". In: *Seminars in Cell & Developmental Biology* 20.7 (2009), pp. 869 –876. DOI: https://doi.org/10.1016/j.semcdb.2009.07.003.

[47] Sui Huang and Stuart Kauffman. "How to escape the cancer attractor: Rationale and limitations of multi-target drugs". In: *Seminars in Cancer Biology* 23.4 (2013), pp. 270 –278. DOI: https://doi.org/10.1016/j.semcancer.2013.06.003. URL: http://www.sciencedirect.com/science/article/pii/S1044579X13000540.

[48] Edward L. Huttlin et al. "The BioPlex Network: A Systematic Exploration of the Human Interactome". In: *Cell* 162.2 (2015), pp. 425–440. DOI: 10.1016/j.cell.2015.06.043. URL: https://doi.org/10.1016/j.cell.2015.06.043.

[49] I. Ispolatov, P. L. Krapivsky, and A. Yuryev. "Duplication-divergence model of protein interaction network". In: *Phys. Rev. E* 71 (6 2005), p. 061911. DOI: 10.1103/PhysRevE.71.061911. URL: https://link.aps.org/doi/10.1103/PhysRevE.71.061911.

[50] Harry C. Jubb et al. "Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health". In: *Progress in Biophysics and Molecular Biology* 128 (2017), pp. 3 –13. ISSN: 0079-6107. DOI: https://doi.org/10.1016/j.pbiomolbio.2016.10.002. URL: http://www.sciencedirect.com/science/article/pii/S0079610716300311.

[51] Ekta Khurana et al. "Interpretation of Genomic Variants Using a Unified Biological Network Approach". In: *PLOS Computational Biology* 9.3 (2013), pp. 1–9. DOI: 10.1371/journal.pcbi.1002886. URL: https://doi.org/10.1371/journal.pcbi.1002886.

[52] Yoo-Ah Kim, Dong-Yeon Cho, and Teresa M. Przytycka. "Understanding Genotype-Phenotype Effects in Cancer via Network Approaches". In: *PLOS Computational Biology* 12.3 (2016), pp. 1–15. DOI: 10.1371/journal.pcbi.1004747. URL: https://doi.org/10.1371/journal.pcbi.1004747.

[53] Brennan Klein et al. "Resilience and evolvability of protein-protein interaction networks". In: *bioRxiv* (2020). DOI: 10.1101/2020.07.02.184325. eprint: https://www.biorxiv.org/content/early/2020/07/02/2020.07.02.184325.full.pdf. URL: https://www.biorxiv.org/content/early/2020/07/02/2020.07.02.184325.

[54] Max Kotlyar et al. "In silico prediction of physical protein interactions and characterization of interactome orphans". In: *Nature Methods* 12.1 (2015), pp. 79–84. DOI: 10.1038/nmeth.3178. URL: https://doi.org/10.1038/nmeth.3178.

[55] Vessela N. Kristensen et al. "Principles and methods of integrative genomic analyses in cancer". In: *Nature Reviews Cancer* 14.5 (2014), pp. 299–313. DOI: 10.1038/nrc3721. URL: https://doi.org/10.1038/nrc3721.

[56] Mark D M Leiserson et al. "Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes." In: *Nat Genet* 47.2 (2015), pp. 106–114. DOI: 10.1038/ng.3168.

[57] Chunhe Li and Jin Wang. "Quantifying Cell Fate Decisions for Differentiation and Reprogramming of a Human Stem Cell Network: Landscape and Biological Paths". In: *PLOS Computational Biology* 9.8 (2013), pp. 1–14. DOI: 10.1371/journal.pcbi.1003165. URL: https://doi.org/10.1371/journal.pcbi.1003165.

[58] Chunhe Li and Jin Wang. "Quantifying the underlying landscape and paths of cancer". In: *Journal of The Royal Society Interface* 11.100 (2014), p. 20140774. DOI: 10.1098/rsif.2014.0774. URL: royalsocietyorg/doi/abs/10.1098/rsif.2014.0774.

[59] Katja Luck et al. "A reference map of the human binary protein interactome". In: *Nature* 580.7803 (2020), pp. 402–408. DOI: 10.1038/s41586-020-2188-x. URL: https://doi.org/10.1038/s41586-020-2188-x.

[60] Katja Luck et al. "Proteome-Scale Human Interactomics". In: *Trends in Biochemical Sciences* 42.5 (2017), pp. 342–354. DOI: 10.1016/j.tibs.2017.02.006. URL: https://doi.org/10.1016/j.tibs.2017.02.006.

[61] Luca Malcovati et al. "Diagnosis and treatment of primary myelodysplastic syndromes in adults: recommendations from LeukemiaNet". In: *Blood* 122.17 (2013), pp. 2943–2964. ISSN: 0006-4971. DOI: 10.1182/blood-2013-03-492884. URL: https://doi.org/10.1182/blood-2013-03-492884.

[62] Luca Malcovati et al. "Prognostic Factors and Life Expectancy in Myelo dysplastic Syndromes Classified According to WHO Criteria: A Basis for Clinical Decision Making". In: *Journal of Clinical Oncology* 23.30 (2005), pp. 7594–7603. DOI: 10.1200/JCO.2005.01.7038. URL: https://doi.org/10.1200/JCO.2005.01.7038.

[63] Eugenio Marco et al. "Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape". In: *Proceedings of the National Academy of Sciences* 111.52 (2014), E5643–E5650. ISSN: 0027-8424. DOI: 10.1073/pnas.1408993111. URL: https://www.pnas.org/content/111/52/E5643.

[64] Naoki Masuda, Mason A. Porter, and Renaud Lambiotte. "Random walks and diffusion on networks". In: *Physics Reports* 716-717 (2017). Random walks and diffusion on networks, pp. 1 –58. ISSN: 0370-1573. DOI: https://doi.org/10.1016/j.physrep.2017.07.007. URL: http://www.sciencedirect.com/science/article/pii/S0370157317302946.

[65] Nigel J. O'Neil, Melanie L. Bailey, and Philip Hieter. "Synthetic lethality and cancer". In: *Nature Reviews Genetics* 18.10 (2017), pp. 613–623. DOI: 10.1038/nrg.2017.47. URL: https://doi.org/10.1038/nrg.2017.47.

[66] Sandra Orchard et al. "The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases". In: *Nucleic Acids Research* 42.D1 (2013), pp. D358–D363. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1115. URL: https://doi.org/10.1093/nar/gkt1115.

[67] E. Papaemmanuil et al. "Somatic SF3B1 Mutation in Myelodysplasia with Ring Sideroblasts". In: *New England Journal of Medicine* 365.15 (2011), pp. 1384–1395. DOI: 10.1056/NEJMoa1103283. URL: https://doi.org/10.1056/NEJMoa1103283.

[68] Elli Papaemmanuil et al. "Clinical and biological implications of driver mutations in myelodysplastic syndromes". In: *Blood* 122.22 (2013). ISSN: 0006-4971. DOI: 10.1182/blood-2013-08-518886. URL: https://doi.org/10.1182/blood-2013-08-518886.

[69] Elli Papaemmanuil et al. "Genomic Classification and Prognosis in Acute Myeloid Leukemia". In: *New England Journal of Medicine* 374.23 (2016). PMID: 27276561, pp. 2209–2221. DOI: 10.1056/NEJMoa1516192. eprint: https://doi.org/10.1056/NEJMoa1516192. URL: https://doi.org/10.1056/NEJMoa1516192.

[70] Romualdo Pastor-Satorras, Eric Smith, and Ricard V. Solé. "Evolving protein interaction networks through gene duplication". In: *Journal of Theoretical Biology* 222.2 (2003), pp. 199 –210. ISSN: 0022-5193. DOI: https://doi.org/10.1016/S0022-5193(03)00028-6. URL: http://www.sciencedirect.com/science/article/pii/S0022519303000286.

[71] Scott D. Pauls and Daniel Remondini. "Measures of centrality based on the spectrum of the Laplacian". In: *Phys. Rev. E* 85 (6 2012), p. 066127. DOI: 10.1103/PhysRevE.85.066127. URL: https://link.aps.org/doi/10.1103/PhysRevE.85.066127.

[72] Sara Rahmati et al. "pathDIP: an annotated resource for known and predicted human gene-pathway associations and pathway enrichment analysis". In: *Nucleic Acids Research* 45.D1 (2016), pp. D419–D426. ISSN: 0305-1048. DOI: 10.1093/nar/gkw1082. URL: https://doi.org/10.1093/nar/gkw1082.

[73] Sabry Razick, George Magklaras, and Ian M. Donaldson. "iRefIndex: A consolidated protein interaction database with provenance". In: *BMC Bioinformatics* 9.1 (2008), p. 405. DOI: 10.1186/1471-2105-9-405. URL: https://doi.org/10.1186/1471-2105-9-405.

[74] Francisco A. Rodrigues et al. "The Kuramoto model in complex networks". In: *Physics Reports* 610 (2016). The Kuramoto model in complex networks, pp. 1 –98. ISSN: 0370-1573. DOI: https://doi.org/10.1016/j.physrep.2015.10.008. URL: http://www.sciencedirect.com/science/article/pii/S0370157315004408.

[75] Thomas Rolland et al. "A Proteome-Scale Map of the Human Interactome Network". In: *Cell* 159.5 (2014), pp. 1212–1226. DOI: 10.1016/j.cell.2014.10.050. URL: https://doi.org/10.1016/j.cell.2014.10.050.

[76] Gert Sabidussi. "The centrality index of a graph". In: *Psychometrika* 31.4 (1966), pp. 581–603. DOI: 10.1007/BF02289527. URL: https://doi.org/10.1007/BF02289527.

[77] Wouter Saelens et al. "A comparison of single-cell trajectory inference methods". In: *Nature Biotechnology* 37.5 (2019), pp. 547–554. DOI: 10.1038/s41587-019-0071-9. URL: https://doi.org/10.1038/s41587-019-0071-9.

[78] Julie Schanz et al. “New Comprehensive Cytogenetic Scoring System for Primary Myelodysplastic Syndromes (MDS) and Oligoblastic Acute Myeloid Leukemia After MDS Derived From an International Database Merge”. In: *Journal of Clinical Oncology* 30.8 (2012), pp. 820–829. DOI: 10.1200/JCO.2011.35.6394. URL: https://doi.org/10.1200/JCO.2011.35.6394.

[79] Michael R. Stratton, Peter J. Campbell, and P. Andrew Futreal. “The cancer genome”. In: *Nature* 458.7239 (2009), pp. 719–724. DOI: 10.1038/nature07943. URL: https://doi.org/10.1038/nature07943.

[80] Damian Szklarczyk et al. “STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets”. In: *Nucleic Acids Research* 47.D1 (2018), pp. D607–D613. ISSN: 0305-1048. DOI: 10.1093/nar/gky1131. URL: https://doi.org/10.1093/nar/gky1131.

[81] Daniel Ting, Ling Huang, and Michael I. Jordan. “An Analysis of the Convergence of Graph Laplacians”. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. Omnipress, 2010, pp. 1079–1086. ISBN: 9781605589077.

[82] Ítalo Faria do Valle et al. “Network integration of multi-tumour omics data suggests novel targeting strategies”. In: *Nature Communications* 9.1 (2018), p. 4514. DOI: 10.1038/s41467-018-06992-7. URL: https://doi.org/10.1038/s41467-018-06992-7.

[83] Bert Vogelstein et al. “Cancer Genome Landscapes”. In: *Science* 339.6127 (2013), pp. 1546–1558. ISSN: 0036-8075. DOI: 10.1126/science.1235122. eprint: https://science.sciencemag.org/content/339/6127/1546.full.pdf. URL: https://science.sciencemag.org/content/339/6127/1546.

[84] Cuihong Wan et al. “Panorama of ancient metazoan macromolecular complexes”. In: *Nature* 525.7569 (2015), pp. 339–344. DOI: 10.1038/nature14877. URL: https://doi.org/10.1038/nature14877.

[85] Jin Wang. “Landscape and flux theory of non-equilibrium dynamical systems with application to biology”. In: *Advances in Physics* 64.1 (2015), pp. 1–137. DOI: 10.1080/00018732.2015.1037068. eprint: https://doi.org/10.1080/00018732.2015.1037068. URL: https://doi.org/10.1080/00018732.2015.1037068.

[86] Caleb Weinreb et al. "Fundamental limits on dynamic inference from single-cell snapshots". In: *Proceedings of the National Academy of Sciences* 115.10 (2018). ISSN: 0027-8424. DOI: 10.1073/pnas.1714723115. URL: https://www.pnas.org/content/115/10/E2467.

[87] Jingwen Yan et al. "Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data". In: *Briefings in Bioinformatics* 19.6 (2017), pp. 1370–1381. ISSN: 1477-4054. DOI: 10.1093/bib/bbx066. eprint: https://academic.oup.com/bib/article-pdf/19/6/1370/27119423/bbx066.pdf. URL: https://doi.org/10.1093/bib/bbx066.

[88] Xinping Yang et al. "Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing". In: *Cell* 164.4 (2016), pp. 805–817. DOI: 10.1016/j.cell.2016.01.029. URL: https://doi.org/10.1016/j.cell.2016.01.029.

[89] Song Yi et al. "Functional variomics and network perturbation: connecting genotype to phenotype in cancer". In: *Nature Reviews Genetics* 18.7 (2017), pp. 395–410. DOI: 10.1038/nrg.2017.8. URL: https://doi.org/10.1038/nrg.2017.8.

[90] Kenichi Yoshida et al. "Frequent pathway mutations of splicing machinery in myelodysplasia". In: *Nature* 478.7367 (2011), pp. 64–69. DOI: 10.1038/nature10496. URL: https://doi.org/10.1038/nature10496.

[91] Joseph Xu Zhou et al. "Quasi-potential landscape in complex multistable systems". In: *Journal of The Royal Society Interface* 9.77 (2012), pp. 3539–3553. DOI: 10.1098/rsif.2012.0434. URL: https://royalsociety.org/doi/abs/10.1098/rsif.2012.0434.

[92] Marinka Zitnik et al. "Evolution of resilience in protein interactomes across the tree of life". In: *Proceedings of the National Academy of Sciences* 116.10 (2019), pp. 4426–4433. ISSN: 0027-8424. DOI: 10.1073/pnas.1818013116. eprint: https://www.pnas.org/content/116/10/4426.full.pdf. URL: https://www.pnas.org/content/116/10/4426.