UNIVERSITÀ DEGLI STUDI DI MILANO
FACOLTÀ DI SCIENZE MATEMATICHE,
FISICHE E NATURALI

CORSO DI LAUREA MAGISTRALE IN FISICA

STATISTICAL METHODS IN CELL CLUSTER ANALYSIS

Relatore: Prof. S. Caracciolo
Correlatore: Dott. S. Zapperi
Correlatore: Dott.ssa C. La Porta

Tesi di Laurea di
Massimiliano Maria Baraldi
Matr. 772016
Codice PACS 87.17.-d

Anno Accademico 2010-2011

Contents

Introduction

The growth of large structures from smaller units is a very common phenomenon in many different areas of science and technology. It has been recognized only relatively recently that many of the large scale structural properties do only depend on the general features of the growth process [1]. This is much like the properties of phase transformations that are determined by very general considerations such as dimension and symmetries. The formation of aggregates is important in many areas of science and has important applications in areas such as air pollution, water pollution and purification, and in many branches of condensed matter, that is polymer physics, percolation theory, coating systems and nanostructure fabrication.

Colloids and polymers [2, 3] (see figure 1) are interesting models for aggregation phenomena. Such systems cover a broad range of particle sizes and interactions. Forces with different length scales (electrostatic, van der Waals, adhesion forces) become relevant for the behavior of individual particles and their collective behavior depending on particle size and on their local environment (particle concentration, confining geometry, properties of the continuous matrix). The time scales on which single particles, groups of particles or the entire sample react to a new situation cover several orders of magnitude.

Colloidal aggregation and the kinetics of colloidal aggregation has been studied for many years. However, emphasis has been focused more on the inter-particle interactions and the kinetics of aggregation processes than on the structure of the aggregates. One of the reasons for the neglect of aggregate structure was the difficulty of characterizing their complex disorderly structures in quantitative terms.

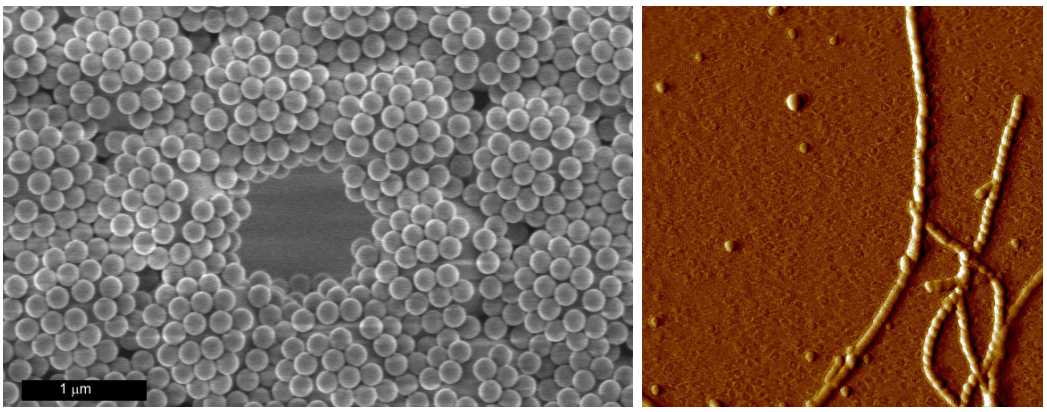

Figure 1: Left: The image shows a colloidal monolayer of 180 nm polystyrene particles on top of a monolayer of 1100 nm colloidal particles. Right: Atomic force microscopy image of Aβ fibrils. Aβ is a 39- to 43-residue peptide that is formed by proteolytic processing of a 770-residue trans-membrane protein and deposited as amyloid fibrils in Alzheimer's disease.

Interest in fractal structures formed by aggregation processes grew rapidly following the fractal analysis of iron particle aggregates by Forrest and Witten [4], the introduction of the DLA (diffusion limited particle-cluster aggregation, figure 2 shows DLA structures found in nature) model by Witten and Sander[5] and the realization that the structures generated by other simple aggregation models and aggregation processes could be described quite well in terms of the concepts of fractal geometry [6].

These advances took place at a time when the basic concepts of fractal geometry had recently become widely disseminated. The intense interest in fractal geometry, at that time, stimulated research on a wide variety of growth and aggregation models, which continues to this day. Interest in this area has been sustained by a strong synergy between computer modeling and experimental work [7].

The most important features of simple aggregation models, in which clusters or aggregates are assembled from a large number of single particles, are the volume distribution of the aggregating clusters, the nature of the relative trajectories of the aggregating clusters, the dimensionality of the space in which the aggregation process takes place and the concentration of particles (the fraction of the space

occupied by particles). In most simple models, the particles are represented by spheres (or hyperspheres) in a continuous space or by filled sites on a lattice. In either case, the distribution of particle sizes is usually neglected.

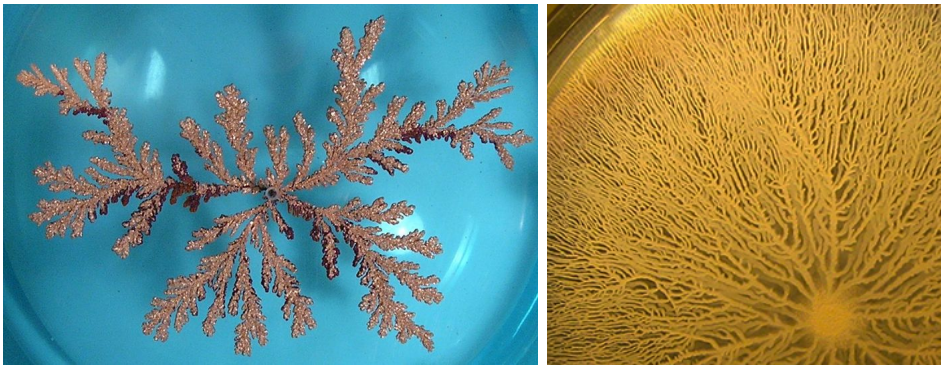

Figure 2: Left: the image represents a DLA cluster grown from a copper sulfate solution in an electrodeposition cell. Right: bacterial colonies grown on a plate.

Aggregation phenomena play an important role in Biology. Bacteria colonies display DLA-like patterns (right panel of figure 2) as those observed in many systems such as electrodeposition (left panel of figure 2), mineral deposits, and dielectric breakdown. The cluster structure observed falls into a universality class according to the growth mechanisms, with its characteristic properties. Just as is known from the field of critical phenomena, the scaling features of these models are universal, i.e. they do not depend on microscopic details. As a consequence, physical concepts developed in Statistical Mechanics clarify the dynamics of biological processes.

This work is focused on clusters of cancer cell that form *in vitro* and motivated by the results obtained in the comprehension of aggregation phenomena. Recent papers have elucidated processes that happen in biological cell systems. Stochastic models of cell division and differentiation have been successful in the comprehension of the maintenance of adult murine tail skin [8, 9].

One of the main goals in Biomedicine is to understand the evolution process of tumors. Since few decades ago, the prevailing theory, pioneered by Robert Weinberg, suggested that all tumor cells are indistinguishable and tumorigenic, that is, all the cells are responsible of tumor growth [10].

In this context, great interest deserved a paper published by John Dick in 1997, in which it was shown that leukemic cells have a hierarchic structure and are originated from a primitive hematopoietic cell [11]. This paper opened the way for many later studies, which suggested that a similar structure existed for solid tumor [12].

Subsequent researches showed the existence of a set of cells, later called cancer stem cells (CSCs), located at the top of the hierarchic pyramid and endowed with the same features of stem cells [13]. Indeed, like normal stem cells, they can self-renew to produce more stem cells and are able to divide (through mitosis) and differentiate into diverse specialized cell types. The ability to self-renew assure the survival of the stem population (that is why they are said to be immortal). Moreover, they can differentiate into diverse progeny with limited proliferative potential or form non-tumorigenic cancer cells that compose the bulk of cells in a tumor.

Figure 3 shows the difference between a non-hierarchic traditional theory and a hierarchic CSC model. This figure shows possible cell division processes. In a hierarchic view a top-level progenitor is able to divide in individuals belonging to different kind of populations whereas the subordinate families do not generate individuals of the top-level population. For example, individuals of a noble family can generate noble and non-noble offspring, while non-nobles can have only non-noble offspring. Instead in a non-hierarchic process the progenitor can generate offspring belonging only to its own family.

Thus, at the basis of the CSC theory is the existence of a minor subpopulation of cells that possesses the peculiar features of stem cells, whereas the remaining majority of the cells are more "differentiated" and do not have these properties.

In this landscape the CSCs are responsible for the growth of the tumor and thus to treat the disease should be sufficient to target this subpopulation. Indeed, this theory opened the doors to a new strategy of cancer treatment. In fact the main weakness of traditional chemiotherapy is that it is not target specific, i.e. kills all the cells that divide rapidly. This results in a lot of side effects like myelosuppression, the decreased production of blood cells, and alopecia, that is hair loss. Whereas,

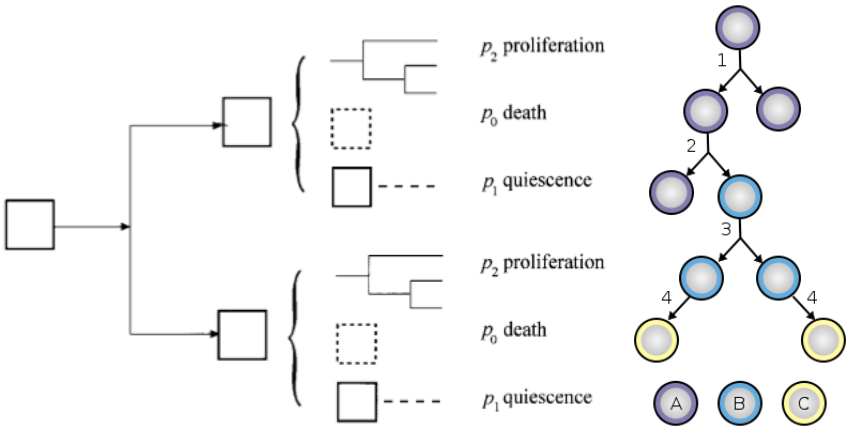

Figure 3: Difference between a non-hierarchic (left) and a hierarchic (right) model. Left: one progenitor divides generating two cell that can proliferate, die or be quiescent. Right: the individuals belong to different populations defined by different features, A can undergo symmetric (1) or asymmetric (2) division, B can undergo symmetric division (3) or generate a member of the C population (4) that is not able to divide.

according to this theory, the key consideration when devising therapeutic treatments should be the tumorigenic potential of the cells, so the driving trend in drug design should pose its strategy in targeting those cells only. In addition, cancer chemiotherapy efficiency is frequently impaired by tumor resistance, that is the reduction in effectiveness of a drug in curing a disease. This is strongly dependent on the exposition to the treatment and closely linked with the specificity of the drug, that is, the lower the specificity the greater the duration of exposure and hence the greater the risk of the development of resistance. A schematic depiction of the difference between traditional and CSC approach in cancer treatment is shown in figure 4.

The CSC theory was proved to be true in different kind of tumors: brain [14], breast [15], colon [16], ovary [17], pancreas [18], prostate [19] and melanoma [20]. The evidence of the existence of CSCs in human tumors is based on the creation of mice that are sufficiently immunodeficient to tolerate tumor growth of human tumor cells into them [21]. So one key objection to this model is the lack of an appropriate microenviroment because of the difference between mice and humans and of the lack of an intact immune system when evaluating the tumor-initiating

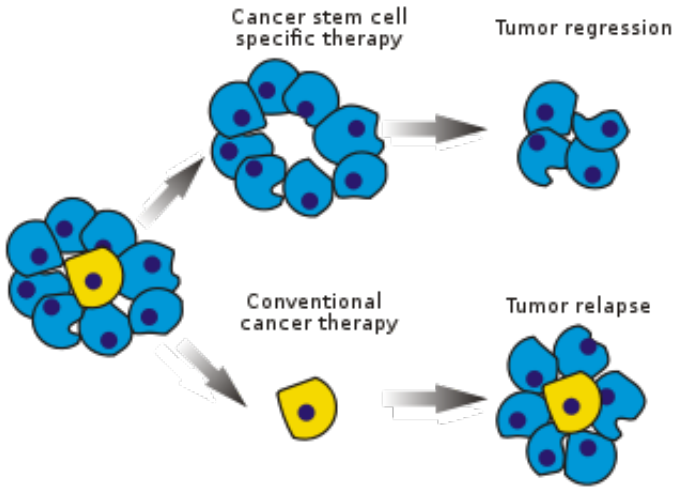

Figure 4: Difference between CSC specific and conventional cancer therapies

capacity of these human cancer cells. Thus, it is possible that the subpopulation of cells that appeared non-tumorigenic might actually be tumorigenic in the presence of the appropriate microenviroment. However, recent studies validated the existence of CSCs with different biological techniques [12].

A number of studies have investigated the possibility of distinguishing CSCs from the bulk of the tumor [21, 22]. This usually deals with the definition of the so-called biomarkers, that are indicators of a biological state which allows for the detection and isolation of a particular cell type. A constant problem in Biology is to find the best marker or the best combination of markers necessary to identify the subset of cells endowed with a specific quality.

In the last twenty years, physical and mathematical models played a crucial role in the comprehension of biological mechanisms [23]. Indeed, in the context of cell division and proliferation the theory of branching processes has been able to describe a wide range of phenomena. In fact this theory describes situations in which an entity exists for a time and then may be replaced by one, two, or more entities of a similar or different type.

The theory of branching processes is a well-developed and active area of research with theoretical interests and practical applications. It has made important con-

tributions to biology and medicine since Francis Galton considered the extinction of names among the British peerage in the nineteenth century [24]. More recently, branching processes have been successfully used to illuminate problems in the areas of molecular biology, cell biology, developmental biology, immunology, evolution, ecology, medicine, and others [23]. For the experimentalist and clinician, branching processes have helped in the understanding of observations that seem counterintuitive, to develop new experiments and clinical protocols, and have provided predictions which have been tested in real life situations. For the physicist, the challenge of understanding new biological and clinical observations has motivated the development of new theories in the field of branching processes.

The main goal of this thesis is to determine the kinetics of tumor growth. Starting from experiments *in vitro*, I will discuss a technique to analyze the data and study the behavior of observables in order to determine the evolution of clusters of melanoma cells. Six papers came out in the last two years showing the evidence of a CSC subpopulation in melanoma [25, 26, 27, 28, 29, 30, 20], and the research group with which I am working determined biomarkers to distinguish the CSCs from the bulk [31, 32]. The aim of this work is to determine the right evolution dynamics of the tumor within the context of Statistical Mechanics and the theory of branching processes, that is, to determine a model that fits the experimental data.

In chapter 1, I firstly discuss the nature of cell clusters and the format of the experimental data. Then I will develop a feasible way to compute the observables using imaging technique. I will show how everyday biological measurement can be performed in a systematic way with the use of percolation and clustering methods. A conversion factor between pixels and cells is computed in order to compare the results with biological observations and models based on Branching Process Theory.

Chapter 2 will be devoted to a test of the methods used. I will verify that clusters are randomly sparse and are not mutually interacting, showing that the distributions of distances between centers of mass of the clusters follow a random-like behavior. I will test the imaging technique used checking that measures of number of clusters

fall in the expected ranges. Further, I will show that measurements are not affected by the experimental setup, that is, a measurement of the volumes of clusters do not depends on the number of clusters, allowing to discard any possible interaction of clusters with the environment in which they are constrained.

Chapter 3 concerns a method to determine the isotropy of the clusters. I will address the question if the clusters follow a random-like growth. The inertia tensor represents a measure of the shape of a cluster: its eigenvalues define the elongation of the cluster along the diagonalization axes, while its maximum eigenvector defines its orientation. With the use of such tensor, I will compare the experimental results with the Eden model whose dynamics give rise to random-like clusters, showing that for the cell type used the growth is isotropic.

Having determined that clusters are mutually independent, I will perform a dynamical analysis of the cluster growth in the context of Branching Process Theory where independence between clusters is a basic feature. Therefore in Chapter 4 I will introduce the basic concepts of this theory, emphasizing the possible implementations of models that can be designed according to biological observations. Therefore, I will discuss a model that fits the Traditional Cancer Theory and a model based on the CSC hypothesis that keeps in account recent results in Biomedicine.

The last chapter deals with the comparison of experimental data with BP models. I will discuss the case of models inspired to the Traditional Cancer Theory and to the CSC Theory. The main goal is that experimental data can be understood only if we suppose the existence of two populations in Melanoma cells, in contrast with the hypothesis of TC Theory. This result open the way for new researches in the context of CSC Theory.

Chapter 1

A computational approach to data analysis

When modeling a biological system, it is of primary interest to understand the basic mechanism that drives the dynamics. Starting from experimental data *in vitro*, it is indeed possible to calculate different observables using imaging techniques and algorithms.

In the subsequent sections, I discuss data capture and analysis. The first section concerns the format of the data set, while in the following I design a suitable algorithm in order to access the observables of interest. The striking feature of the method implemented here is that it is general and can thus be used to study the behavior of different kinds of cells that form two-dimensional clusters.

1.1 Data sets

In this section the crystal violet technique to prepare the data sets *in vitro* is explained and the nature of the cells anlyzed is discussed.

Samples of Melanoma cells, originating from patients and frozen, are put for a number of days in culture, that is immersed in growth medium that facilitate the growth of the cells. Afterwards, the cells are disposed in different wells and solutions

containing crystal violet and formalin are used to simultaneously fix and stain cells grown in cell culture to preserve them and make them easily visible, since most cells are colorless.

The samples consist in sets of six wells, each one covered by violet spots representing cluster of cells and all prepared in the same condition. Figure 1.1 represents one of the samples. The main reason why it is necessary to prepare six wells is not only statistical but also experimental: it could happen that the cells are sometimes not fixed and the mixture of crystal violet and formalin spread on the well making difficult to distinguish the clusters.

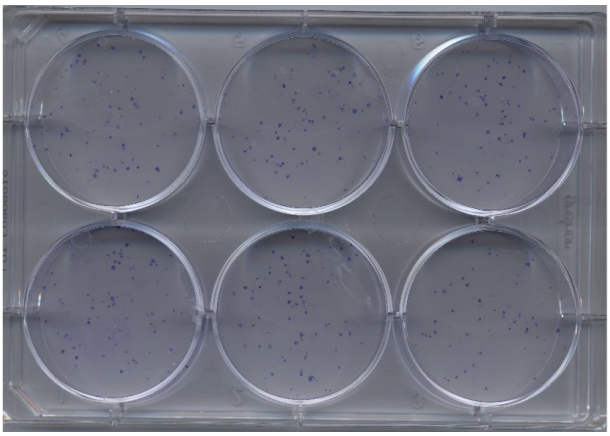

Figure 1.1: Here is shown an example of the data: each one of the six wells contains cells of Melanoma (line 39) in the wild type condition after 8 days growth.

In oncology it is possible to distinguish between different kinds of Melanoma cells. In order to understand these differences, it is essential to introduce the concepts of metastasis and biological marker.

Metastasis is the spread of a disease from one organ or part to another non-adjacent organ or part. This happens when the cancer cells, that form the primary tumor, acquire the ability to penetrate and infiltrate surrounding normal tissues in the local area, forming a new tumor. The newly formed "daughter" tumor in the adjacent site within the tissue is called a local metastasis.

Some cancer cells acquire the ability to penetrate the walls of lymphatic and/or blood vessels, after which they are able to circulate through the bloodstream (cir-

culating tumor cells) to other sites and tissues in the body. This process is known (respectively) as lymphatic or hematogeneous spread. After the tumor cells come to rest at another site, they re-penetrate through the vessel or walls, continue to multiply, and eventually another clinically detectable tumor is formed. This new tumor is known as a metastatic (or secondary) tumor. Metastasis is one of three hallmarks of malignancy (in contrast to benign tumors).

When tumor cells metastasize, the new tumor is called a secondary or metastatic tumor, and its cells are like those in the original tumor. This means, for example, that, if breast cancer metastasizes to the lungs, the secondary tumor is made up of abnormal breast cells, not of abnormal lung cells. The tumor in the lung is then called metastatic breast cancer, not lung cancer.

In genetics, cancers are distinguished by the so called line, that specifies the cell type. Melanoma cells belonging to the line 39 are obtained from a patient in a metastatic phase, while those belonging to the line 37 are obtained from the primary tumor. In the subsequent chapters the behavior of line 39 cells will be studied.

Another distinction between cancer cells is based on the biological markers. As said in the introduction, these markers are used to detect a biological state and thus to isolate a particular cell type. In Biology, it is said that a particular kind of cells (for example the Melanoma cells) express a certain marker, meaning that it is possible to detect this particular kind of cells. In this way it is possible to detect different populations among all the Melanoma cells.

The Cancer Stem Cell hypothesis opened the way for a lot of studies each suggesting a marker or a set of markers to detect the CSC population. Thus it is common use to distinguish Melanoma cells that express a certain marker, in facts a large number of markers has been proposed as good markers for Melanoma CSCs [21].

1.2 Data analysis

An efficient technique to analyze the data sets shown in figure 1.1 has not been designed yet. In this section, a suitable method to get informations on the clusters (for example on the shape and on the size) is discussed. This method is based on imaging techniques and on a percolation algorithm.

1.2.1 Data conversion

With a common scanner it is possible to obtain an image with very good resolution of the clusters (in the data analysis, it has been used a scanning resolution of 600 x 2400 dpi). Selecting circular section, an image for each of the six wells is obtained. In this operation, we should be careful in cutting out the shaded and the reflective areas.

We saved the image in ppm format. The ppm file is an ASCII file and allows for a simple manipulation of the information contained in a pixel. The ppm file can be opened with a simple text editor and contains:

- two lines that represent the file format and the filter used to produce the image
- a line containing two numbers that respectively define the number of columns and lines of the pixel lattice
- a line containing a number that represents the maximum color-component value that in the standard RGB scale is set to 255
- three ASCII decimal values for each pixel between 0 and the specified maximum value, starting at the top-left corner of the pixmap, proceeding in normal English reading order. The three values for each pixel represent red, green, and blue, respectively; a value of 0 means that color is off, and the maximum value means that color is maxed out (for example $(0, 0, 0)$ corresponds to black while $(255, 255, 255)$ corresponds to white).

1.2.2 Image conversion

The first goal to achieve is to convert the image obtained, that is a matrix of colors, in a boolean matrix where 1 corresponds to an element of a cluster, while 0 correspond to an empty pixel.

This is a very complicated task because gray and violet, that are the two colors that should be distinguished, are very "close" in the pixmap. In principle this is not a problem because it should be possible to distinguish a first set of colors to be treated like an element of a cluster and a second set of colors that correspond to empty pixels. But random noise and shadows must be taken into account, indeed if these two sets are disjointed in a region of the plate, they could overlap in another region. So a distinction between clusters and the background should not be based on the identification of these two sets of colors.

We executed this operation using a combination of imaging techniques. The basic feature is the edge detect "Difference of Gaussian" algorithm. It works by performing two different Gaussian blurs (a Gaussian blur acts on each pixel of the active layer or selection, setting its value to the average of all pixel values present in a radius defined) on the image, with a different blurring radius for each, and subtracting them to yield the result. This algorithm is very widely used in artificial vision and is pretty fast because there are very efficient methods for doing Gaussian blurs. The most important parameters are the blurring radii for the two Gaussian blurs. Increasing the smaller radius tends to give thicker-appearing edges, and decreasing the larger radius tends to increase the threshold for recognizing something as an edge. The details of this method are reported in Appendix 1 and the efficiency of this method can be appreciated in figure 1.2.

In this way, black clusters over a white background are obtained. Here should be noted that some clusters can be connected to the edges of the selected area. This means that these clusters are cut and can affect the measurement of observables, like for example centers of mass or volumes of the clusters. This is avoided simply

erasing these clusters.

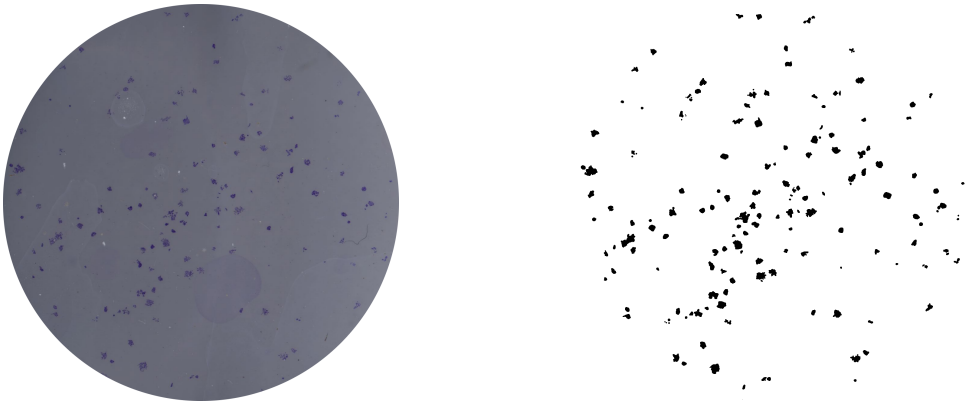

Figure 1.2: The images show the original section of the well (left) and its black and white conversion (right).

1.2.3 Cluster labeling

The pixmap obtained in this way contains black $(0, 0, 0)$ and white $(255, 255, 255)$ elements and can be simply converted in a boolean matrix B where 1 correspond to an occupied black site and 0 to an empty white site, thus

$$B = [\sigma_{i,j}] \quad \text{where } \sigma_{i,j} \in \{0, 1\}. \tag{1.1}$$

The interesting quantity that can be calculated at this point is the area covered by the clusters with a simple count of 1 and 0 in the boolean matrix, but for example the number of clusters and thier volumes cannot be calculated.

In order to distinguish between clusters it is necessary to assign labels. What we would like to have is an algorithm which gives all sites within the same cluster the same label and gives different labels to sites belonging to different clusters. The Hoshen Kopelman algorithm [33], widely used in percolation theory, allows a fast labeling of the clusters. The time complexity of this algorithm is linear and requires small computer memory size. In fact with this algorithm, it is possible to simply handle the corresponding matrices 1300×1300 of the wells.

The general idea of the Hoshen Kopelman algorithm is that we scan through the grid, from left to right and from top to bottom, looking for occupied sites and to the left and the top neighbors. To each occupied site we wish to assign a label corresponding to the cluster to which the site belongs. If the site has zero occupied neighbors, then we assign to it a cluster label we have not yet used (it is a new cluster, 1.3 top-left). If the site has one occupied neighbor, then we assign to the current site the same label as the occupied neighbor (they are part of the same cluster, 1.3 top-right and bottom-left). If the site has more than one occupied neighboring site, then we choose the lowest-numbered cluster label of the occupied neighbors to use as the label for the current site (1.3 bottom-right).

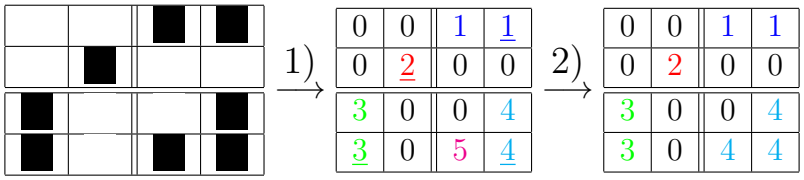

Figure 1.3: 1) A graphical sketch of the possible situation during cluster labeling. Considering the underlined numbers: Top-left: definition of a new cluster labeled with 2, top-right: the site belongs to the cluster already labeled with 1, bottom-left: the site belongs to the cluster already labeled with 3, bottom-right: the two neigbours are labeled with different numbers, thus the minor number (4) is assigned to the site. After this first step, $N(M) = M$ for $M \neq 5$ and $N(5) = 4$. 2) This second step represents the relabeling of the clusters using the information contained in N.

Furthermore, if these neighboring sites have differing labels, we must make a note that these different labels correspond to the same cluster. Thus we introduce an additional array, the labels of labels, and denote it as N. A good label for a site $\sigma_{i,j}$, say M, is characterized by $N(M) = M$ whereas a bad label has $N(M) = M'$, with M' the label to which that bad label turned out to be connected. Scanning the grid for the first time, all the connections are stored in the array N, that is when neighboring sites have different labels M_{max} and M_{min} then $N(M_{max}) = M_{min}$ and $N(M_{min}) = M_{min}$. Once finished, the good label for each cluster is found by the

following classification: given M the label of that site, then if $N(M) = M$ the label is good and we go to the next site, otherwise $N(M) = M'$ and we must check if M' is a good label, if not $N(M') = M''$ and we proceed until we find that $N(M^*) = M^*$ thus we set $N(M) = M^*$.

In this way, scanning the lattice once, an equivalence relation between two labels M_1 and M_2 is defined by

$$M_1 \sim M_2 \quad \text{if } N(M_1) = M_2 \tag{1.2}$$

and we can define as well the equivalence class of M,

$$[M] = \{M_i \in \Lambda' | M_i \sim M\} \tag{1.3}$$

given $\Lambda' = \{M_i\}$ the set of labels M_i. Therefore going through the lattice for a second time all the bad labels for the sites $\sigma_{i,j}$ are replaced by the good ones using the array N. In this step the labels in Λ', that do not follow a numerical order, are reordered, i.e. for an occupied site with original label M the good label at the root of the label tree is searched then replaced by an integer n for the n-th cluster found in the lattice (in english reading order) and stored in an array of new labels Λ,

$$\Lambda = \{n, n \in \{1, 2, ..., n_c\} | N_\sigma(\sigma_{i,j}) = n\} \tag{1.4}$$

where n_c is the total number of clusters and N_σ is the cluster respective label for the site. In this way the boolean matrix B that represents a well is converted in the cluster label matrix L where each occupied site is replaced by a number that labels the equivalence class to which it belongs, thus

$$B = [\sigma_{i,j}] \text{ where } \sigma_{i,j} \in \{0, 1\} \quad \xrightarrow{\text{H-K}} \quad L = [L_{i,j}] = \begin{cases} 0 & \text{if } \sigma_{i,j} = 0 \\ N_\sigma(\sigma_{i,j}) & \text{if } \sigma_{i,j} = 1 \end{cases} . \tag{1.5}$$

In this way, the number of clusters is immediately obtained and the labels assigned to the sites allows a simple calculation for example of the volumes of the clusters based on the count of matrix elements equal to the representative number of its equivalence class. The right panel of figure 1.4 shows the colored version of the well, obtained assigning a random color to each label of the clusters.

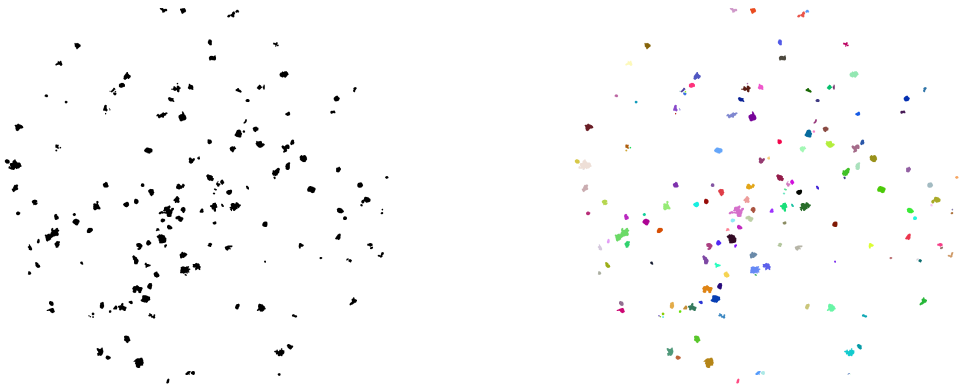

Figure 1.4: Here is shown the black and white image (left) and its coloured version (right). The colour of the clusters is determined choosing three random number between 0 and 255 and assigning them to the representative number of the equivalence class of a cluster, then printed on a file in ppm format.

1.2.4 Defining clusters of cells

With the Hoshen Kopelman algorithm, clusters can be separated labeling them with different numbers. However we have to underline that in this landscape we are using the definition of a cluster as a region of connected occupied sites (or black pixels), that is two occupied sites are said to be connected if there exists a path of occupied sites that connect them. The question that we address in this section is if a cluster of cells can be identified with the H-K definition of the cluster.

Consider the experimental protocol of preparing the sample: a certain number of cells is put and randomly scattered in the well and left in a growth medium for 8/10 days, then fixed with the crystal violet technique. In this way, a cluster is the

product of a series of divisions generated by one cell. Sometimes it happens that, after fixing, some cells are slightly separated from the cluster of which they are part of and the H-K algorithm counts them as different. Thus the H-K algorithm needs improvements in order to justify the identification of a biological cluster as a computed cluster.

What we need to do is to define a new equivalence relation between clusters that says that a "little cluster quite close to a big cluster" represents the same cluster.

Following this line, we designed an algorithm able to perform this task. The basic idea is that there exists a set Γ of big clusters surely generated by one cell and a set γ of smaller clusters that could be generated by one cell but also be part of an existing cluster. Thus at this point a first parameter must be defined, that is the threshold size S^* that determine the set to which the cluster belongs. Thus, given

$$C_k = \{L_{i,j} | L_{i,j} = k\} \tag{1.6}$$

the set of sites of the k-th cluster,

$$C_k \in \Gamma \quad \text{if} \quad |C_k| \geq S^* \tag{1.7}$$

$$C_k \in \gamma \quad \text{if} \quad |C_k| < S^* \tag{1.8}$$

Then scanning the lattice we look for a cluster in the neighbor of a certain radius r, that is identified as a coherence length, of the sites belonging to the γ clusters and we put them in the same equivalence class, i.e.

$$C_a \sim C_b \quad \text{if } \exists a \in C_a, b \in C_b \quad | \quad d(a,b) < r \tag{1.9}$$

where almost one between C_a and C_b belongs to the set γ. This clearly avoids the presence of an equivalence relation between two clusters in Γ that are biologically generated by two distinct initial cells. The method used to label the clusters is completely analogous to the one described in the preceding section. This algorithm

Figure 1.5: The figures show blowups of the colored images obtained with the Hoshen Kopelman algorithm (left) and the respective result of the second algorithm (right).

has been tested using different parameters and has brought good results for $S^* = 100$ px and $r = 8$ px. Figure 1.5 shows the effect of the algorithm on a well using these parameters.

1.3 Pixel conversion

In the preceding section I conducted the analysis using the pixel as unit of measurement. However in this case the natural unit of measurement is the cell, in order to reproduce the real composition of the clusters and to compare the experimental data with a growth model inspired to the theory of branching processes.

Thus it is necessary to determine a conversion factor between cells and pixels, that is defined by the ratio p of cells and pixels that constitute the clusters. We calculated the number of pixels counting the occupied sites of the lattice while we counted the corresponding number of cells using a microscope endowed with the resolution of a μm. Figure 1.6 shows a microscope image of a cluster and its respective counterpart in pixels. The computation is achieved using four small clusters from which we calculated the average of p, whereas large clusters contain a huge number of cells that is not easy to determine to the naked eye using the

microscope. The conversion factor is

$$p = 0.137 \pm 0.046 \text{ cells/px}$$

or equivalently a cell corresponds to 7/8 px.

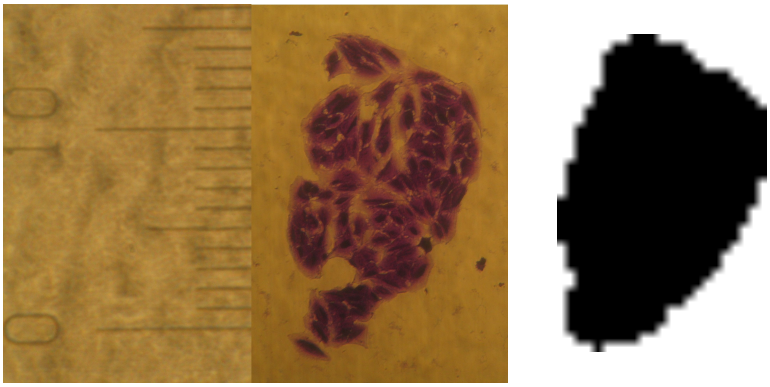

Figure 1.6: The left figure represents a photo of a cluster, the scale reported has a resolution of a μm. The right figure is the same image obtained with the scanner in pixel units.

Here it should be noted that the cluster label matrix $L_{i,j}$ contains spatial informations on the clusters, that get lost when converting observables in cell units. In fact, when dealing with geometrical properties of clusters, calculations will be carried out starting from $L_{i,j}$. Meanwhile when targeting dynamical properties in branching process context and in Biology field the cell represents the natural choice for the measurement unit.

Chapter 2

Calibration

In the preceding chapter we discussed a systematic way to convert the image of the wells and label clusters. Here the algorithm is implemented to discuss data on melanoma cells obtained from tumor in a non-methastatic stage (line 39). The interest of the subsequent discussion is not only biological, in facts this represents a way to test the algorithm. The emphasis is put here on the behavior of some observables that are useful to check the validity of the method implemented and are of ordinary interest in biological researches.

There are mainly two parameters that can be changed in this kind of experiments: time and density. In Biology the time scales of cell division dynamics are of the order of days. In fact the mean time in which a cell divide is usually between one and two days, but clearly depends on which kind of cell you are dealing with. The range in which time t can be varied for Melanoma cells is very narrow, because of a combination of experimental reasons. In facts according to the experimental protocol of crystal violet technique, 8 days must be waited to see enough big clusters, while 12 days are too much because clusters become so big that merge and clusters generated by different cells cannot be distinguished. Thus,

$$t \in [8 \text{ days}, 10 \text{ days}].$$

The second parameter is density ρ^*, that is the initial condition on the number of cells in a well. Also in this case there is a fundamental constraint, in fact a certain number of cells must not be exceeded in order to avoid a merging of the clusters after few days. There is clearly an inverse relation between time and the constraint on density, that is more is the time less must be the threshold number of initial cells. In our particular case the threshold value of initial concentration ρ^*_{max} can be set to

$$\rho^*_{max}(t) = \begin{cases} 250 \text{ cells/well} & \text{if } t = 8 \text{ days} \\ 150 \text{ cells/well} & \text{if } t = 10 \text{ days} \end{cases} . \tag{2.1}$$

This can be easily seen looking at the figure 2.1, where the wells in the two border situation mentioned above are shown. The problem of merging will be discussed further in the third section.

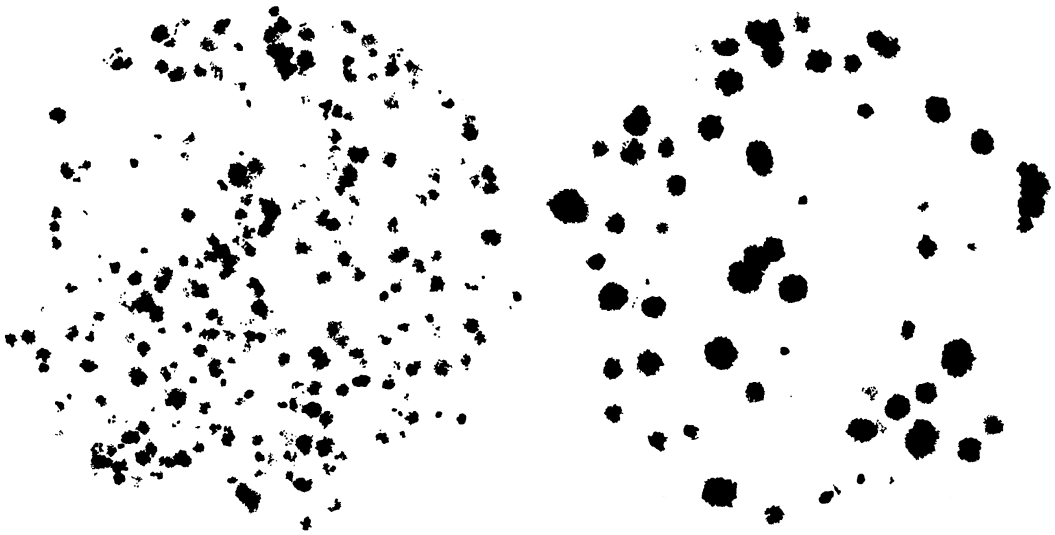

Figure 2.1: The left figure represents a sample of cells 39 wild type at 8 days with a density $\rho^* = 250$ cells/well and the right figure is a sample of the same kind of cells at 10 days with $\rho^* = 150$ cells/well. In these two wells situations of merging of clusters are seen.

Before going on discussing the results, some points that concern the experimental conditions and the "quality" of the data sets must be still considered. In facts when preparing the wells, a solution with cells is spread on the plate using a pipette and

in this process there are many errors that cannot be kept into account. Errors happen when preparing the exact concentration of cells in the solution and when dosing the right number of cells with the pipette. In the first step the number of cells thus statistical errors cannot be measured, while in the second step errors can be in principle estimated with a statistical approach (that we will adopt). However this procedure is strictly dependent on the precision of the operator that should be sure to exert an appropriate pressure on the stuff of the pipette when preparing each well. This problem is not trivial at all, because the operation is affected by the effort of the experimentalist and can give rise to a systematic error.

2.1 Sparseness of clusters

According to the experimental protocol, cells are supposed to be spread randomly on the well using the pipette. But nothing prevent the cluster to interact, in fact an attractive or repulsive force between them can exists and can affect the geometrical and dynamical aspects of the growth. This must be clearly verified and we are able to do this using the algorithm described in the preceding chapter.

The basic idea is that the average position of the initial cell that give rise to a cluster is located at the center of the cluster. Their center is defined as the center of mass of the cluster using the classical definition, that is, given (j, i) the coordinates of the occupied site in the cluster label matrix $L_{i,j}$ that is part of the k-th cluster, and C_k the set of sites of the k-th cluster, the coordinates of the center of mass are

$$x_{k,CM} = \frac{1}{|C_k|} \sum_{L_{i,j} \in C_k} j, \quad y_{k,CM} = \frac{1}{|C_k|} \sum_{L_{i,j} \in C_k} i \tag{2.2}$$

where $|C_k|$ is the number of sites that compose the k-th cluster. Note that the centers of mass of the clusters can be computed using experimental data because the algorithm allows a labeling of different clusters.

We computed in this way the average positions of the cells in the initial condi-

tion. In order to determine the presence of a possible interaction between clusters, the experimental data on the position of the initial cells must be compared with a random situation. This is achieved comparing the respective distribution of (euclidean) distances between the centers of mass of the clusters with the distribution $p_r(x) = p_r(x_{i,j} = x)$ of (euclidean) distances $x_{i,j} = ||\mathbf{x_i} - \mathbf{x_j}||$ between random points in a circle of radius r. The calculation of $p_r(x)$ is carried out in appendix B and gives

$$p_r(x) = \frac{2x}{r}\left(\frac{2}{\pi}\arccos\left(\frac{x}{2r}\right) - \frac{x}{\pi r}\sqrt{1 - \frac{x^2}{4r^2}}\right). \tag{2.3}$$

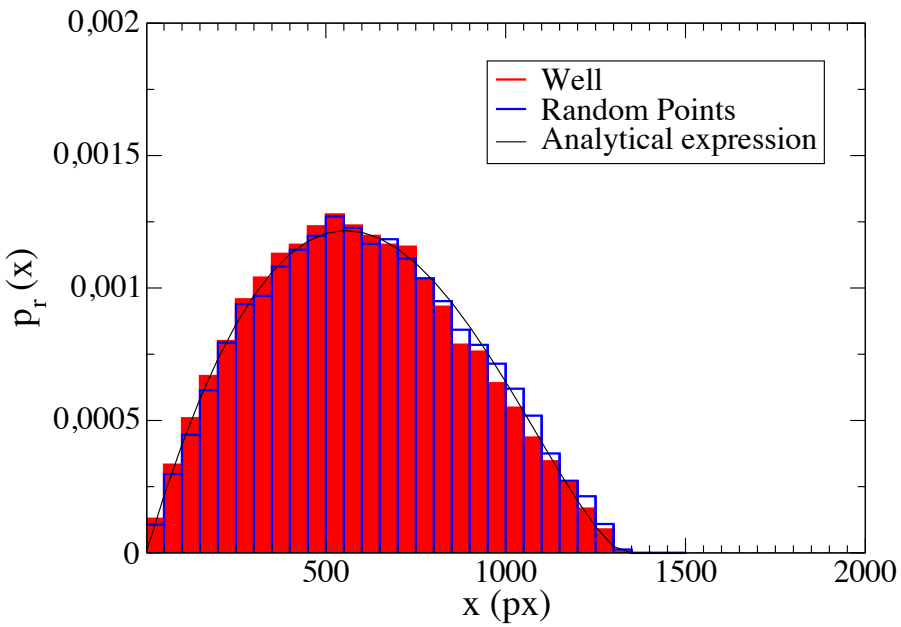

Figure 2.2: Here is shown the comparison between the experimental distribution (in red), the simulated one (in blue) and the analytical curve (in black) $p_r(x)$ for a sample of non-metastatic (line 39) ABCG2-negative cells.

We thus verified the sparseness of the clusters studying a sample for each kind of cell. Figure 2.2 shows the experimental distribution of $p_r(x)$ and the simulated distribution compared with the analytical expression of equation 2.3 for a sample

of non-metastatic ABCG2-negative cells. The simulated distribution is obtained trowing random points in a circular well, that is trowing two random numbers between 0 and the maximum number of rows R or columns C of the matrix that represents the pixel lattice (the matrix is squared, thus $R = C$), and accepting them when they fall in the circular well, that is when $\left(x - \frac{R}{2}\right)^2 + \left(y - \frac{R}{2}\right)^2 < \frac{R^2}{4}$ given (x, y) the coordinate of the random point in the grid in a reference frame where the origin is the bottom-left pixel. The good agreement of the data with the simulated and the analytical curves shows that the clusters are randomly distributed in the well.

2.2 Testing the code

In the preceding section we tested the sparseness of the cells in the initial condition showing that the distribution of distances between the centers of mass of the cluster $p_r(x)$ follows the behavior of random points in a circular section. Thus any correlation in the initial conditions and any effect due to a possible interaction between different clusters must be discarded.

We test now the validity of the algorithms used to label clusters, thus measurements of observables of the well will be now discussed. The image conversion technique is not at issue because the steps described in appendix A are kept under control using a graphical interface, thus if any error occur this can be detected and solved, or if background noise is too strong that cannot be removed the well is discarded.

A direct and efficient test of the method is based on the measurement of the number of clusters n_c in the wells that is determined by the largest number that labels the clusters. In facts, as said at the beginning of the chapter, between the two parameters that can be set it is density ρ^*, that is the starting number of cells displaced in a well (it is a pure number and has not the dimension of a physical density, it is a numerical density in a well not in a volume). Therefore a confirm of

the method used consists in checking the agreement between the number of clusters and the density. If n are the numbers that label the sites and defines the cluster to which they belong, the number of clusters is calculated as

$$n_c = \max_{n \in \Lambda} n \tag{2.4}$$

Note that this is possible because in the algorithms the labels of the clusters are in a numerical order. If cells do not die, we should expect that $n_c = \rho^*$, however some cells could die thus we expect that $n_c \lesssim \rho^*$ (The reason of the relation between n_c and ρ is that it is experimentally observed that few cells, not the majority, die).

Here should be noted that the density ρ^* that measures the number of cells in a well must be rescaled in order to obtain the number of cells in the circular section that will be denoted as ρ. The scaling factor is determined as the ratio of the selected area and the area of the well. Areas has been selected with fixed radius $r = 650$ px while wells have a radius $R = 800$ px, thus the conversion factor is $\alpha = \frac{\pi r^2}{\pi R^2} = \frac{650^2}{800^2} = 0.66$. Therefore the density defined in the experiments must be rescaled in our analysis, i.e. $\rho = \alpha \rho^*$.

There are two sources of errors in this method when defining ρ. The first one deals with the experimental protocol, in facts, as explained at the beginning of this chapter, there is an error when determining the density ρ^*. The second one appears when cutting the section of the well, i.e. when defining ρ, because it is possible that in some cases the fraction of initial cells and thus clusters that fall inside the circular section is not exactly α. This combination of effects is kept in account averaging over six wells (when possible, in facts it happens that some wells cannot be analyzed because of background noise) like those shown in figure 1.1. However in this way the errors are underestimated because of "systematic" errors on the initial density that are not measurable.

The fraction of area covered by clusters A_c can be used as an indicator of experimental errors. This does not depend on the algorithms used to label the clusters

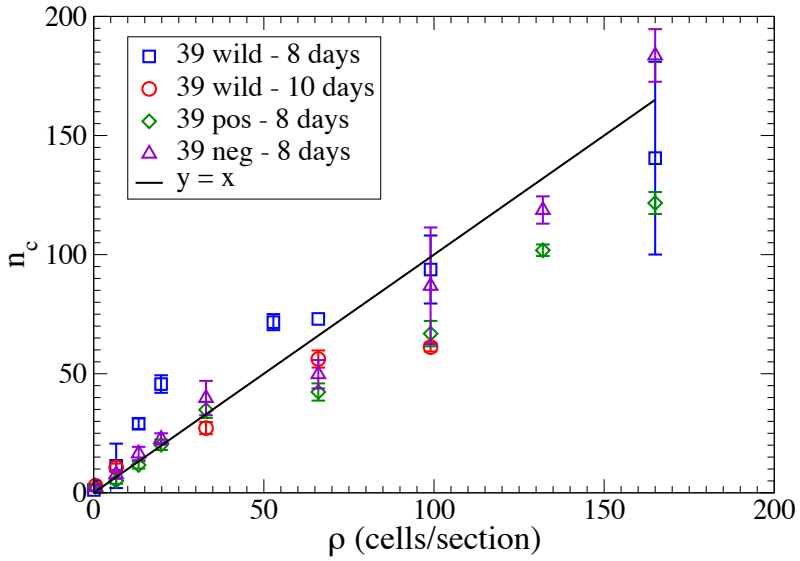

Figure 2.3: The graph shows the comparison between experimental data of $n_c(\rho)$ for the 39 wild type at 8 and 10 days, the 39 ABCG2 - positive and the 39 ABCG2 - negative with their theoretical upper bound $n_c = \rho$ (black curve).

because is determined just by counting the occupied sites in the pixel lattice. Therefore the functions $n_c(\rho)$ should somewhat reflects the behavior of $A_c(\rho)$.

Figure 2.3 shows the experimental data of $n_c(\rho)$ for four different data sets with the curve $n_c = \rho$ that represents their upper bound and figure 2.4 shows the compounding experimental results for $A_c(\rho)$. The cells analyzed all belong to the 39 line (non-metastatic) but differ in kind and time. "Wild type" are commonly defined as those cells that have not been treated and do not express a particular marker, they are the cells of the tumor obtained from the patient. "ABCG2 - positive" or "ABCG2 - negative" denote those populations of cells that respectively result positive or negative to the marker ABCG2, that is supposed to be a marker for the CSC subpopulation [31]. We analyzed these three kind of cells after a growth process of 8 days, while for the wild type we varied also the time studying the situation at 10 days.

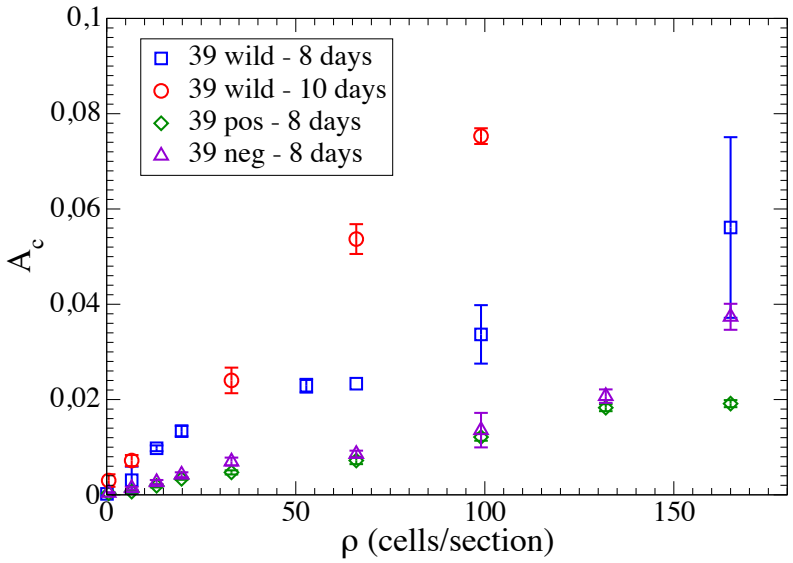

Figure 2.4: Here are shown the experimental data of $A_c(\rho)$ for the 39 wild type at 8 and 10 days, the 39 ABCG2 - positive and the 39 ABCG2 - negative.

The graph of $A_c(\rho)$ is helpful in understanding the existence of a systematic error, in facts it clarifies that the bump of the blue points in figure 2.3 is due to an experimental error, not to an error of the algorithms. Similar arguments are valid for the last violet and green point.

To summarize with the Hoshen-Kopelman algorithm the connected black areas as defined in the first chapter are labeled, but counting the clusters in this way would result in an overstimation of the number of clusters. This is evident looking at the wells, because separated connected regions are in some cases part of the same cluster. This has been solved introducing a clustering algorithm based on a "coherence length" that decreases the count on the number of clusters previously obtained, thus resulting in the constraint $n_c \lesssim \rho$ experimentally observed in figure 2.3. Figure 2.5 shows the graph of $n_c(\rho)$ for all the analyzed cases considering the measurements of n_c with and without the clustering algorithm in comparison with the expected theoretical upper bound. The difference is evident for the case of the 39

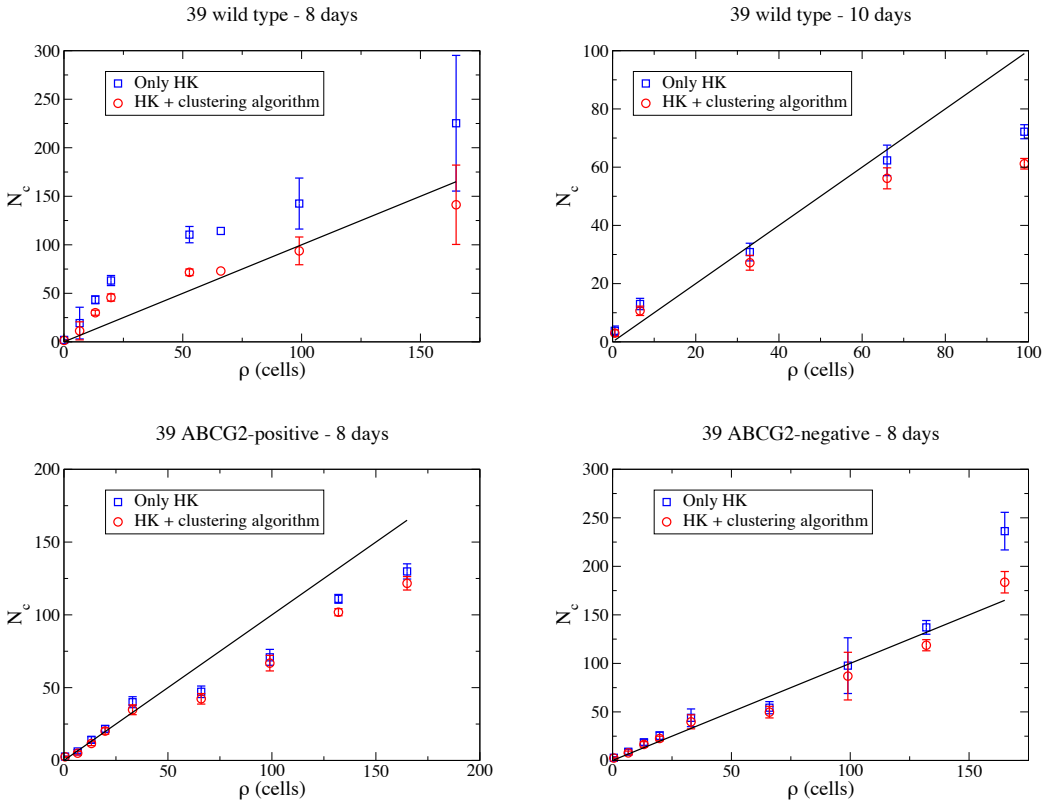

Figure 2.5: The graph represents the comparison between the efficiency of the H-K algorithm alone (blue points) and the one improved with the clustering technique (red points). The black curve represents the theoretical expected upper bound.

wild type at 8 days where the cells tend not to stay one close to each other as seen in the left panel of figure 2.1. In this case the clustering algorithm represents a better improvement of H-K, while for the other cases the difference is not well-marked, but the clustering algorithm will be crucial when calculating the cluster volumes of the clusters in the next section.

2.3 Independence of the growth on density

We excluded a possible interaction between clusters and tested the code. Next we will show that any dependence on the density can be discarded when calculating the volumes of clusters. Thus a possibility that the clusters are somewhat affected

during the growth by the constraint of being in a well will be discarded as well.

The volume V_n of the n-th cluster is calculated as the number of the elements of the cluster matrix $L_{i,j}$ labeled with a definite number n, i.e. given $C_n = \{L_{i,j} | L_{i,j} = n\}$ the set of sites of the n-th cluster then V_n is the cardinality of C_n,

$$V_n = |C_n|. \tag{2.5}$$

A distribution for each well is then obtained for different values of the density ρ and for the different kinds of cells. We used here for the distributions a logarithmic binning, against the common linear binning. This is useful when dealing with few datas and when small occurrences are extremely common whereas large instances are extremely rare, in fact this binning method is widely used when dealing with noisy tails (for example power-law and exponentials) as is the case here. The number of data for a well is few, indeed it is given by the number of clusters n_c and is not large enough because the density ρ^* must follow the constraint $\rho^* < \rho^*_{max}$ expressed by equation 2.1, and thus $n_c \lesssim \rho = \alpha\rho^* < \alpha\rho^*_{max}$. In a linear binning landscape, few data points and a corresponding theoretical probability law with rare large instances would result in a noisy tail making harder any interpretation of experimental data.

By definition logarithmic binning in a given base β means that the bin has a constant logarithmic (in the base β) width, thus the logarithm of the upper edge of a bin b_{i+1} is equal to the logarithm of the lower edge of that bin b_i plus the bin width δb. That is,

$$\log_\beta(b_{i+1}) = \log_\beta(b_i) + \delta b \Longleftrightarrow b_{i+1} = b_i\beta^{\delta b}.$$

The center of the bin is then plotted on the x-axis, thus

$$x_i = \frac{1}{2}(b_i + b_{i+1}) \tag{2.6}$$

The number of observation in a bin y_i is normalized by the with of the bin $\Delta b_i =$

$b_{i+1} - b_i$ they fall in when dealing with observables that assumes real values, resulting in

$$y'_i = \frac{y_i}{\Delta b_i}.$$

Instead, when dealing with integer observables, the number of observation y_i is normalized by the number of integers $\Delta \hat{b}_i = \lfloor b_i \rfloor$ that fall in the interval Δb_i when dealing with integers, resulting in

$$y'_i = \frac{y_i}{\Delta \hat{b}_i}.$$

This distinction is crucial when evaluating short intervals Δb_i.

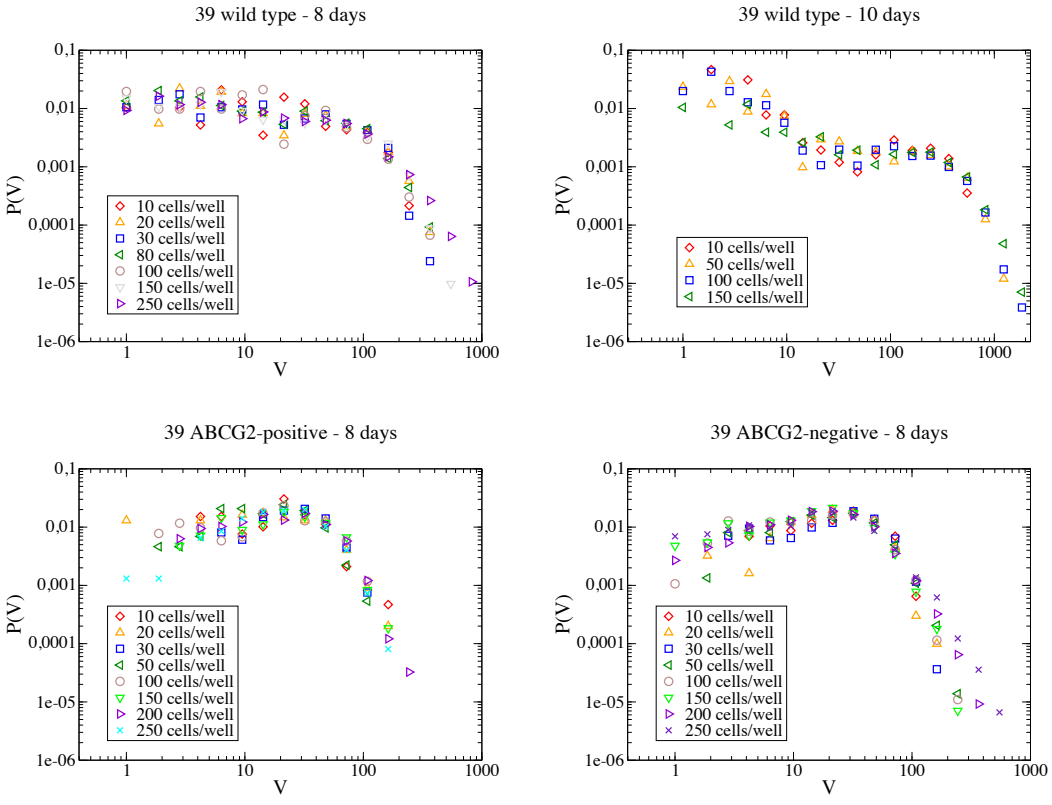

Figure 2.6: The graphs show the normalized distributions $P(V)$ of the volumes of the clusters at different densities for the four sets analyzed in log-log scale. The parameters used are $\beta = 1.3$, $\delta b = 1$, $b_0 = 1$. Here is evident that the distributions show a good matching.

The distributions obtained are renormalized in order to compare between the results obtained for different densities. Figure 2.6 shows the distribution for the four different kinds of cell studied at different densities. There is here a clear evidence of matching between the distributions at different densities, that prove the absence of any possible dependence on density, thus confirming that the cells do not feel the constraint of being in a well.

This also confirms that merging does not affect measurements. In facts, experiments at low densities ρ are not affected by merging and match with high density well, where merging can happen and corrupt the results.

Furthermore the data for a given cell type on volumes of clusters can be summed to get smoother curves, indeed here should be remarked that the amount of data for a well is less than ρ^*_{max}. A distribution is obtained for each cell type considering the data on all the volumes at different densities (see figures 2.7). From the plots, it is trivial that clusters at 10 days are bigger than at 8 days for the 39 wild type cells, meanwhile ABCG2-negative and -positive data overlap showing that sorting with this marker does not result in a difference in cell volumes at short times.

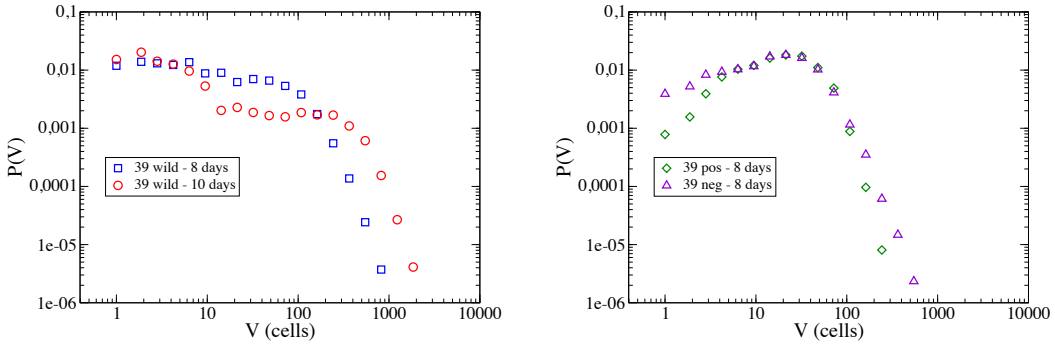

Figure 2.7: The graphs show the normalized distributions $P(V)$ of the volumes of the clusters obtained from all the data at different densities. In the left figure are compared the wild type distributions at different time steps, while in the right figure are compared the ABCG2-sorted data.

Chapter 3

Targeting the geometry

We widely discussed in the preceding chapter the independence of clusters, showing that clusters are randomly distributed in a well and that the distributions of volume of clusters are not affected by density, thus excluding any possible interaction with the environment. Now that a cluster do not affect its neighbors, we should ask if clusters grow randomly thus isotropically, or there exists a preferential direction of growth thus an anisotropy.

3.1 Parametrizing the anisotropy

An observable able to parametrize the degree of anisotropy is here discussed. A natural measure of the geometry of a cluster is described by the inertia tensor $I^{\mu,\nu}$ and for a system of particles with masses m_i and positions $\mathbf{r}_i$ is defined by

$$I^{\mu,\nu} = \sum m_i(|\mathbf{r}_i|^2\delta^{\mu,\nu} - r_i^\mu r_i^\nu) \tag{3.1}$$

where i labels the particle and μ, ν are the coordinate indexes. Here should be noted that $I^{\mu,\nu}$ depends on the location of the origin of the axes set and on the orientation of the axes with respect to the system of particles considered.

The natural moment of inertia of a system is about the axes passing through the

centre of mass r_{CM} because they minimize the moment of inertia with respect to the position of the axes,

$$\frac{\partial}{\partial x^\beta} I^{\mu,\nu} = \frac{\partial}{\partial x^\beta} \sum m_i (|\mathbf{r}_i - \mathbf{r}_{CM}|^2 \delta^{\mu,\nu} - (r_i^\mu - r_{CM}^\mu)(r_i^\nu - r_{CM}^\nu)) = 0. \tag{3.2}$$

This match with the well-known Huygens-Steiner theorem that states that

$$I^{\mu,\nu} = I_{CM}^{\mu,\nu} + \sum m_i (|\mathbf{r'}_i|^2 \delta^{\mu,\nu} - r_i'^\mu r_i'^\nu) \tag{3.3}$$

where $\mathbf{r}'_i = \mathbf{r}_i - \mathbf{r}_{CM}$ represent the positions with respect to centers of mass, thus implying that

$$I^{\mu,\nu} \geq I_{CM}^{\mu,\nu} \tag{3.4}$$

Therefore when all principal moments of inertia are distinct, the principal axes through center of mass are uniquely specified. If all the principal moments are the same, the system is spherically symmetric and any axis can be considered a principal axis, meaning that the moment of inertia is the same about any axis.

By the spectral theorem, since the moment of inertia tensor is real and symmetric, there exists a Cartesian coordinate system in which it is diagonal, having the form

$$I^{\mu,\nu} = \lambda^\mu \delta^{\mu,\nu} \tag{3.5}$$

where the coordinate axes are called the principal axes and the constants λ^μ are called the principal moments of inertia. The principal axis with the highest moment of inertia $\lambda_{max} = \max_\mu \lambda^\mu$ is sometimes called the figure axis.

The cluster is a 2-dimensional object and the inertia tensor for the k-th cluster I_k is calculated starting from the coordinates (j, i) of the cluster label matrix $L_{i,j}$ as

$$I_k = \begin{bmatrix} \sum\limits_{L_{i,j}=k} (j - x_{k,CM})^2 & -\sum\limits_{L_{i,j}=k} (j - x_{k,CM})(i - y_{k,CM}) \\ -\sum\limits_{L_{i,j}=k} (j - x_{k,CM})(i - y_{k,CM}) & \sum\limits_{L_{i,j}=k} (i - y_{k,CM})^2 \end{bmatrix} \tag{3.6}$$

where $(x_{k,CM}; y_{k,CM})$ are the coordinate of the center of mass of the k-th cluster defined in equation 2.2. The principal moments of inertia $\lambda_{k,M}$ and $\lambda_{k,m}$ (where M denote the maximum and m denote the minimum) and the corresponding principal axes for the k-th cluster $\mathbf{v}_{k,M}$ and $\mathbf{v}_{k,m}$ are found by diagonalizing the moment of inertia I_k.

For a symmetric matrix of the form

$$I = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \tag{3.7}$$

the eigenvalues are given by

$$\lambda_{M,m} = \frac{1}{2}\left(a + c \pm \sqrt{(a-c)^2 + 4b^2}\right) \tag{3.8}$$

and are associated with the eigenvectors

$$\mathbf{v}_{M,m} = \begin{pmatrix} 1 \\ \frac{\lambda_{M,m} - a}{b} \end{pmatrix}. \tag{3.9}$$

The anisotropy of the clusters is then parametrized by the ratio of the maximum and the minimum eigenvalue

$$E_k = \frac{\lambda_{k,M}}{\lambda_{k,m}}, \tag{3.10}$$

that is the "eccentricity" of the cluster, thus $E_k \geq 1$ and $E_k = 1$ when the cluster is spherically symmetric.

The anisotropy is not only defined by E_k because this parameter does not keep into account the orientation of the clusters with respect to the well. This is in fact determined by the eigenvectors $\mathbf{v}_{k,M}$ that define the orientation of the figure axis with respect to a fixed reference frame. A second parameter is then given by the angle θ that $\mathbf{v}_M$ form with the reference frame, as shown in figure 3.1. From the non-normalized expression of the eigenvectors 3.9, $v_{M,y}$ immediately gives $\tan\theta_k$

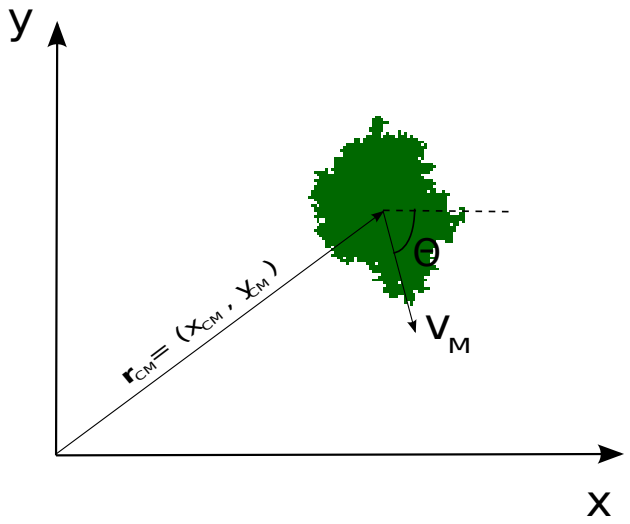

Figure 3.1: Here is a schematic depiction of $\mathbf{v}_M$ and θ for a given cluster with respect to a fixed reference frame. The position of the center of mass of the cluster is $\mathbf{r}_{CM} = (x_{CM}, y_{CM})$.

thus implying that

$$\theta_k = \arctan\left(\frac{\lambda_{k,M} - a}{b}\right) \tag{3.11}$$

3.2 Random clusters

In the preceding section we identified two parameter (E and θ) in order to define the geometry of a cluster. The main goal is now to detect if the dynamic follows a random behavior or if there exists a preferential direction of growth.

At this point, a definition of random cluster growth is needed. Most of the progress that has been made in the study of aggregation phenomena in the last few years has derived from numerical simulations carried out on models of growth mechanisms. Dynamical features enter here due to the irreversible nature of these systems; time has a direction. A simple model of cellular growth was proposed along these lines by Eden in 1961 to account for the tumor proliferation [34, 35].

The process begins with a nucleation site on a square lattice, then one of the empty sites next to the aggregate, that is defined by the perimeter sites, is chosen

randomly and added to the cluster. In this landscape, one of the sites next to the aggregate is occupied by a particle with probability $p_L = L^{-1}$ where L is the perimeter of the cluster. In this way a perimeter site connected to the cluster through more than a occupied nearest neighbor has more chance to be occupied. A large cluster is obtained after having repeated this procedure many times.

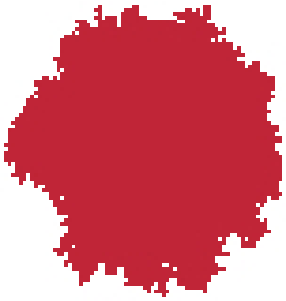

Figure 3.2: The image represents a cluster obtained with the eden model after 4000 time steps.

The Eden model has a random spherical symmetric dynamics, since there are no constraints on the direction of duplication of a site (that represents a cell). When a Eden cluster is examined (see Figure 3.2), it is found to be compact except for a few holes close to the surface, but its surface is found to be rather tortuous. Indeed, this model has been widely investigated in the fractal geometry field, in order to determine the behavior of his fractal surface.

The Eden model is here considered to mimic the geometry of random clusters. However a parameter must be set, the volume of the cluster V. Here should be remarked that in this model the anisotropy E is now a function of the volume, i.e. $E = E(V)$. Indeed as the volume of the cluster increase, the surface of the cluster get smoother and the volume get more isotropic.

A further observation concerns the orientation angle θ for random clusters obtained with this model. In facts simulated clusters do not have a precise orientation one respect to each other or with respect to a fixed reference frame because they

are completely independent. If no interaction between clusters is set, no orientation is predicted. There is no reason to expect something different from a uniform distribution.

3.3 Results

We gave a definition of random cluster according to the Eden model. The goal is now to compare experimental results with the ones obtained from random clusters. As said in the first section of this chapter, given a cluster on a square lattice the inertia tensor I_k is calculated for each cluster (labeled by k) as defined by equation 3.6 and values of eccentricity E_k and the orientation angle θ_k are calculated as well (Equations 3.10 and 3.11). Thus the distribution $P(E)$ for the eccentricity E and $P(\theta)$ for the orientation angle θ is obtained from experimental data.

The analysis in this chapter is restricted to the 39 wild type data that represent the main target of all this work. In fact we are interested in understanding the main feature of the tumors as found in nature, not treated with any marker. We analyzed data concerning the ABCG2-negative or -positive cells in the preceding chapters in order to test the confidence of the algorithm, in facts the morphology of these cells is different from the wild type ones.

Here a method to determine the distribution $P(E)$ for the simulated clusters is discussed. In fact it is not possible to obtain a form for this distribution only from the simulation because of the dependence of the eccentricities on the volumes, i.e. $E = E(V)$. At this point, we can have information only on the conditional probability $P(E|V)$ of the eccentricity E given the volume V. Therefore the basic idea is to calculate $P(E)$ using the law of total probability,

$$P(E) = \sum_V P(E|V)P(V). \tag{3.12}$$

Note that the volumes V assume integer values, because they are measured in cell

units, thus we have a sum not an integral.

To perform exactly this calculation it is necessary to have an analytical form of $P(E|V)$ and $P(V)$, because the sum runs over all the volumes V. Obtaining a simulated distribution $P(E|V)$ for each volume V is not possible. However calculating it till a maximum value for the volume V_{max} in this case makes sense because both $P(E|V)$ and $P(V)$ vanish for high value of V (see figures 3.3 and 3.5). In this way a precision is set. A rude estimate of the level of precision is given by the fitted distribution given by equations 3.17 and 3.19, that will be discussed further in the subsequent analysis. In facts, the ratio

$$k_{E,V} = \frac{P(E=1|V)}{P(E=1|V=1)} = V^{\beta} e^{-\gamma(V^{\delta}-1)} \tag{3.13}$$

that is a measurement of precision for $P(E|V)$ is less than 10^{-3} if $V \gtrsim 1.5 \cdot 10^4$ cells and the ratio

$$k_V = \frac{P(V)}{P(V=1)} = e^{-\frac{V-1}{V^*}} \tag{3.14}$$

that measures the precision for $P(V)$ is less than 10^{-3} if $V \gtrsim 750$ cells for 10 days data. Thus we should be able to simulate Eden dynamics till a volume $V_{max} = 1.5 \cdot 10^4$ to obtain a correct shape for the distribution $P(E)$. This takes very long times, therefore a combined fit is achieved for $P(E|V)$.

A similar statement concern the distribution of volumes, in facts in order to have a value of $P(V)$ for each volume V we should have a huge amount of data on volumes of clusters, that would imply the analysis of hundreds of 6-well sets. Therefore also for $P(V)$ a fit is achieved.

From the simulation, the number of occurrences $N(E|V)$ is obtained then renormalized according to

$$P(E|V) = \frac{N(E|V)}{\sum_E N(E|V)} \tag{3.15}$$

in order to obtain the distributions $P(E|V)$ at fixed values of V shown in figure 3.3.

The error bars are determined running different simulations, thus calculating the errors as a standard deviation for each measurement of $P(E|V)$ at a given value of E (and for a fixed volume V).

An analytical formula for $P(E|V)$ is obtained using a combined fit. This is achieved using *pyFitting* (available on the site https://github.com/gdurin/pyFitting), a python-based program that perform data fitting using non-linear square minimization, that is minimizing the cost function

$$H(\{\pi_i\}) = \sum_i \left(\frac{y_i^{theory} - y_i^{data}(\{\pi_i\})}{\sigma_i} \right)^2 \tag{3.16}$$

where $\{\pi_i\}$ are the parameters, y_i the function value, and σ_i the error on the data points. The reason of this choice is that it fits a number of curves simultaneously with parametric functions, thus we can fit with this program $P(E|V)$ using different values for V.

The curves obtained are fitted with the law

$$P(E|V) = \alpha V^{\beta} e^{-\gamma E V^{\delta}} \tag{3.17}$$

where the parameters are determined with the fit,

$$\alpha = 0.0251 \pm 4 \cdot 10^{-4}$$
$$\beta = 1.351 \pm 4 \cdot 10^{-3}$$
$$\gamma = 0.1390 \pm 8 \cdot 10^{-4}$$
$$\delta = 0.517 \pm 1 \cdot 10^{-3}$$

Figure 3.3 shows the plots of the conditional probability $P(E|V)$ as a function of E for different values of V with their relative fit. Figure 3.4 shows that all the data

for different V collapse on a single master curve if we make a change of variables

$$\begin{cases} y = \frac{P(E|V)}{V^{\beta}} \\ x = E \cdot V^{\delta} \end{cases} \tag{3.18}$$

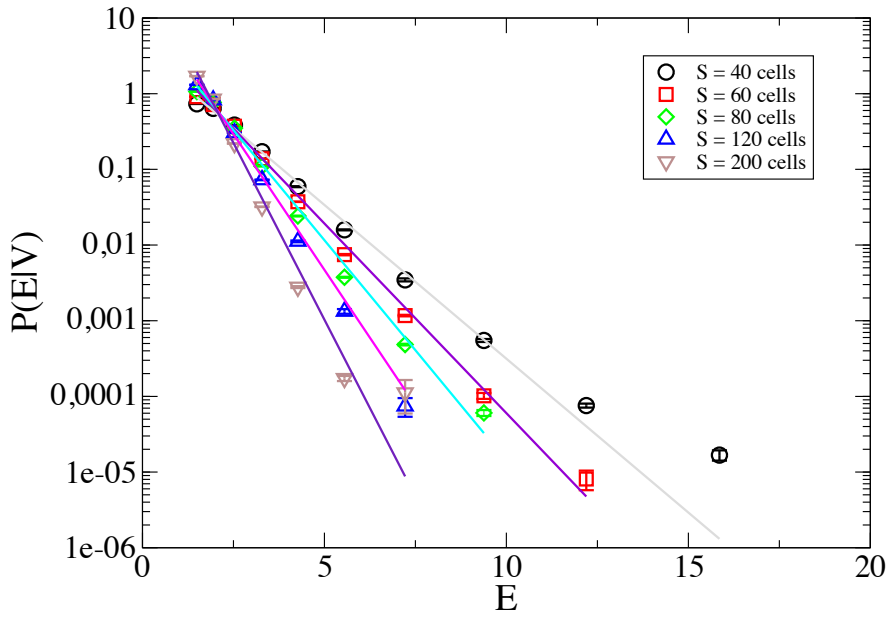

Figure 3.3: The plot shows the curves $P(E|V)$ (labeled by symbols) and the respective fit obtained with *pyFitting* (denoted with straight lines). Log-linear scale is set to emphasize the inverse exponential behavior of the curves.

Given the conditional probability $P(E|V)$, further discussions are needed for the distributions of cluster volumes $P(V)$. In the preceding chapter, the calculation of $P(V)$ has been achieved for different values of the density ρ showing that all the data collapse on a single curve and that there is no dependence on the density. Because of the low number of data, a smoother curve for $P(V)$ shown in figure 3.5 is obtained considering the data on all the volumes from the different wells.

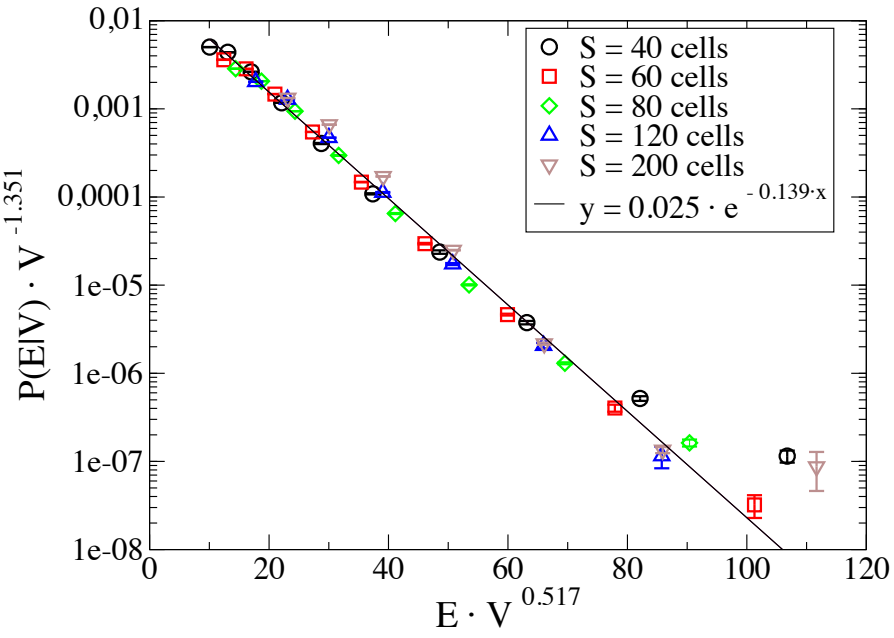

Figure 3.4: The graph represents the master curve (in log-lin scale) to which the data collapse. Symbols represent the simulated data while the straight line is the fitting curve obtained with *pyFitting*.

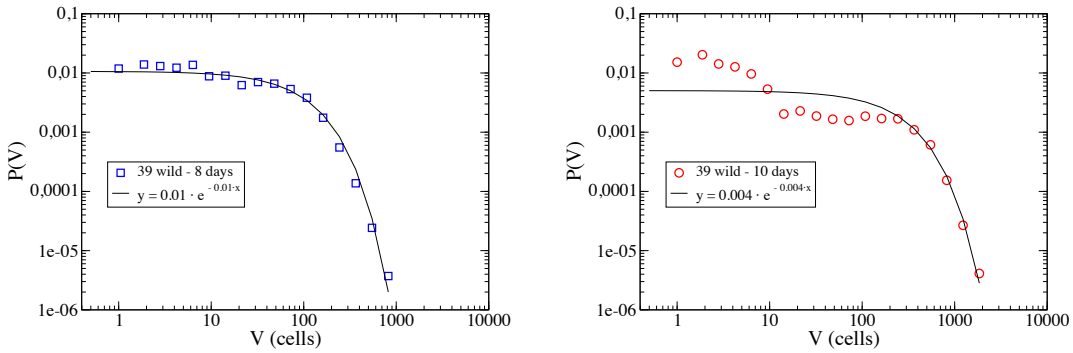

Figure 3.5: These plots represent the distributions of volumes of clusters for the wild type cells at 8 days (left figure) and at 10 days (right figure). Exponential fits are obtained for both curves.

The 8 days data set shows an exponential-like behavior, i.e

$$P(V) \sim e^{-\frac{V}{V^*}} \tag{3.19}$$

This is supposed to be true also for 10 day. Thus using *pyFitting* the data has been fitted with equation 3.19, giving

$$V_8^* = 96 \pm 2 \text{ cells} \tag{3.20}$$

$$V_{10}^* = 250 \pm 6 \text{ cells.} \tag{3.21}$$

In figure 3.5 experimental data are compared with the exponential fitted curves. As shown, the matching is strikingly good at 8 days, while the distribution $P(V)$ at 10 days displays a bump for low value of V that is not precisely reproduced by the fit. This will be further discussed in the next chapters, however here should be remarked that we are not trying at this point to explain why we should see a particular behavior for $P(V)$, because the goal is now to detect an approximation of the experimental curves.

At last the distribution $P(E)$ is determined numerically according to equation 3.12 using equation 3.17 for $P(E|V)$ and equation 3.19 for $P(V)$. Figures 3.6 show the results obtained for the normalized distributions $P(E)$ that satisfy $\sum_E P(E) = 1$. These figures show that experimental data display a good matching with random Eden cluster predictions.

Here should be remarked that error bars for experimental distributions cannot be determined because of the small number of experimental data. In principle they can be determined considering different sets of wells and calculating the standard deviation on the different sets of $P(E)$ for a given value of E. Instead if evaluating error bars for the random cluster distributions using error propagation, huge error bars are obtained because they keep in account of the errors in the fit of $P(E|V)$ and $P(V)$.

Figure 3.7 shows the experimental plots for the cumulative distribution $P_c(\theta)$ of the orientation angles θ. The graphs reflect a uniform distribution for θ, resulting in a random behavior for cluster orientations with respect to the well.

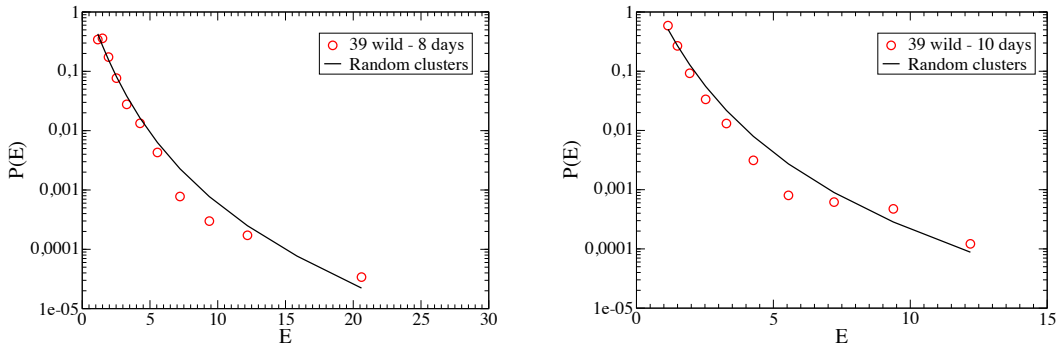

Figure 3.6: Here are shown the distributions $P(E)$ of the cluster anisotropy parameter E for the wild type cells at 8 days (left figure) and at 10 days (right figure). Black curves represent the distribution $P(E)$ for random Eden clusters having a volume distribution $P(V)$ (shown in figure 3.5). $P(E)$ is then obtained according to equation 3.12. The value for $P(E)$ are obtained using the logarithmic binning method (for decimal values) because of noisy tails due to the limited number of data.

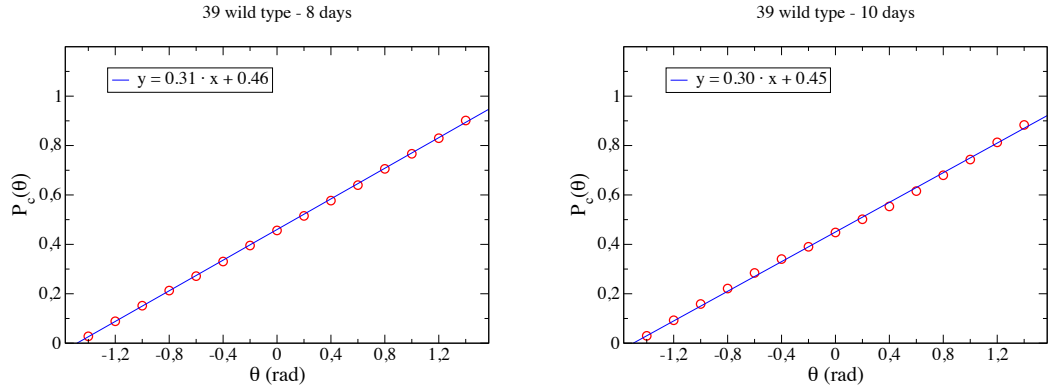

Figure 3.7: The graphs represent the cumulative distributions $P_c(\theta)$ of the orientation angles for the 39 wild type samples at 8 and 10 days. The linear fitting (blue lines) shows that experimental values are uniformly distributed ($\chi^2_{8,rel} = 0.0016$,$\chi^2_{10,rel} = 0.0095$).

Chapter 4

Branching process theory and models

The methods explained in the preceding chapters have been designed to understand the phenomenology of cell cluster growth. The main goals achieved for the cell type used are that cell clusters are completely independent and show random-like geometries. Here is addressed the question of how it is possible to determine a model to explain the dynamic of cluster growth.

I adopted an approach based on Branching Process Theory [39]. This theory describes processes of systems of particles (individuals, cells, molecules, etc.) which live for a random time and, at some point during lifetime or at the moment of death, produce a random number of progeny. Processes that assume the production of progeny at the terminal point of the parent entity's life-time are called the classical processes. They are usually sufficient for modeling populations of biological cells, genes, or biomolecules. Indeed branching processes are used to model reproduction, therefore this theory represents a natural choice for researches in modeling cell population dynamics.

One of the oldest branching processes ever considered was the process in which particles were male individuals bearing noble English family names. An ancestor in

such a process initiated a pedigree which might inevitably become extinct if all of the male descendants died without heirs. Is the extinction of a noble family name inevitable in the long run? How many generations will elapse before extinction occurs? These are typical questions asked about a process in which the number of progeny of an individual may be equal to zero.

A different type of question may be posed for processes in which the growth is assured by a sufficiently high proliferation rate. Then, the interesting parameter is the long-term growth rate and the size and composition of the population at a given time. This is typical of laboratory populations of biological cells, cultured with abundant nutrients and sufficient space. This is our case since cancer cell clusters grow indefinitely in the wells considered.

This theory, that has been firstly introduced in probability theory, has brought striking results in Molecular and Cell Biology [23], Evolution Theories and Medicine.

In this chapter, the mathematical formulation of the Branching Processes Theory will be reviewed, emphasizing how this theory can be applied to cell cluster analysis.

4.1 Analytical formulation of Branching Process Theory

In order to give a mathematical definition of a branching process, some basic definitions of probability theory must be pointed out.

A *stochastic process* with state space S is a collection of random variables $\{Z_t, t \in T\}$ defined on the same probability space (Ω, F, P). The set T is called its parameter set. If $T = \mathbb{N} = \{0, 1, 2, ...\}$, the process is said to be a discrete parameter process. If T is not countable, the process is said to have a continuous parameter. The index t represents time, and then one thinks of Z_t as the "state" or the "position" of the process at time t.

A *Branching Process* is a stochastic process in which each particle in the process

behaves identically as all other particles and independently of all other particles. This feature is usually called *branching property*. It implies that a process can be decomposed into subprocesses, which are identical with each other and with the entire process. This in probability theory results in the statement that subprocesses are independent and identically distributed "variables".

The classical approach starts from considering a branching process in which progeny are born at the moment of parent's death. This is more intuitive when developing the mathematical formulation, but it is not exactly our case, because cells do not die producing a random number of progeny, cells divide through mitosis, die or become quiescent, i.e. they continue to exist without proliferating or dying. However the situation is equivalent because a death individual that produce two offsprings is not distinguishable from a mitosis process.

A branching process is parametrized by a family of non-negative random variables $\{Z_t(\omega), t \geq 0\}$ defined on a common probability space Ω with elements ω, where $Z_t(\omega)$ is the number of particles in the process at time t and ω index the particular realizations of the process (here we use the notation Z instead of X for the random variables following the standard notation found in BP literature [23, 39, 40, 41]). When counting the cells in a cluster, we are measuring $Z_t(\omega)$, in facts following the classical approach the individuals existing at preceding times die leaving place to new offsprings, while in real situations cells existing at preceding times divides producing new cells.

The branching process is initiated at time $t = 0$ by a single ancestor particle, that in cell cluster analysis corresponds to one of the cell displaced on the well in the initial condition. Suppose that the life length of the ancestor is a random variable $\tau(\omega)$ and that the number of its progeny (produced at its death) is equal to $X(\omega)$. Each of the progeny can be treated as the ancestor of its own process, which is a component of our branching process. Then, the number of individuals present in the process at time t is equal to the sum of the numbers of the individuals present in all these subprocesses. This bookkeeping is correct for $t \geq \tau(\omega)$, i.e. after the

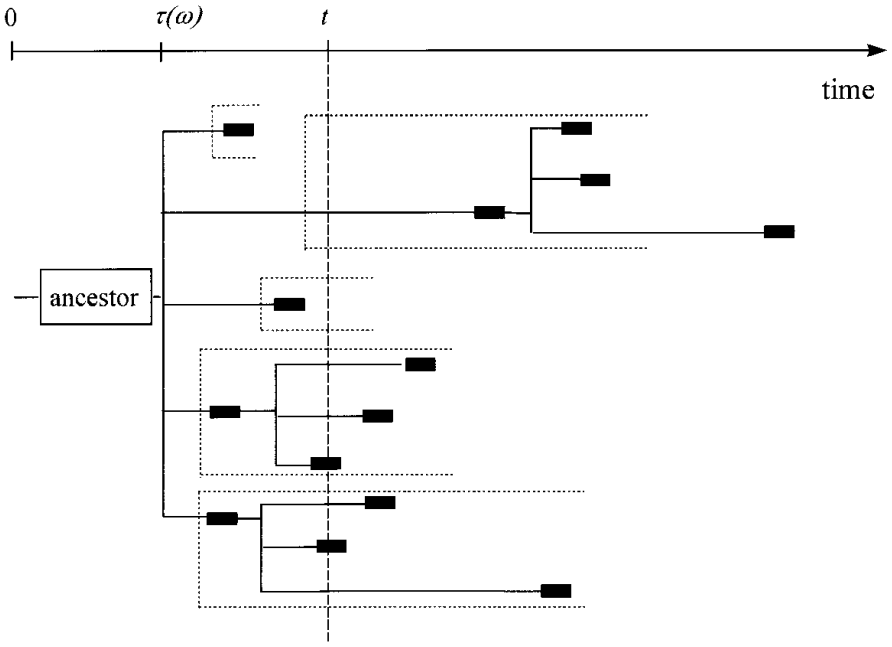

Figure 4.1: Decomposition of the branching process into subprocesses generated by the first-generation progeny of the ancestor. In the case depicted, the number of the first-generation progeny is equal to $X(\omega) = 5$. At time $t > \tau(\omega)$, the number of particles in the subprocesses generated by progeny $1, 2, 3, 4$, and 5 is equal, respectively, to $Z_t^{(1)}(\tau(\omega), \omega) = 0$, $Z_t^{(2)}(\tau(\omega), \omega) = 1$, $Z_t^{(3)}(\tau(\omega), \omega) = 0$, $Z_t^{(4)}(\tau(\omega), \omega) = 3$ and $Z_t^{(5)}(\tau(\omega), \omega) = 3$. The total number of particles in the process at time t is seven.

ancestor has died. Before the ancestors death, the number of particles is equal to 1. Summarizing,

$$Z_t(\omega) = \begin{cases} \sum_{i=1}^{X(\omega)} Z_t^{(i)}(\tau(\omega), \omega) & t \geq \tau(\omega) \\ 1 & t < \tau(\omega) \end{cases} \tag{4.1}$$

where $Z_t(\tau(\omega), \omega)$ denotes the number of individuals at time t in the process started by a single ancestor born at time $\tau(\omega)$, and the additional superscript (i) denotes the ith independent identically distributed (iid) copy. A schematic depiction of an example of branching tree is shown in figure 4.1.

The *branching property* that states that subprocesses are identical to the whole

process is expressed by

$$Z_t(\tau(\omega), \omega) = Z_{t-\tau(\omega)}(\omega) \tag{4.2}$$

that is the processes initiated by the progeny of the ancestor are independent and identically distributed as the ancestor.

Here we handle distributions of random variables (Z) using the probability generating function (pgf) of the distribution. It is the basic analytic tool employed to deal with non-negative random variables. In the subsequent, the argument ω will be dropped from the notation, although implicitly it is always existing.

Let Z be a non-negative random variable, such that $P(Z = i) = p_i$. We write $Z \sim \{p_i\}_{i\geq 0}$ and say that p_i is the distribution of Z.

Definition *(Probability generating function)* The pgf f_Z of a non-negative random variable Z is a function $f_Z(s) = \langle s^Z \rangle = \sum_{i=0}^{\infty} p_i s^i$ of a symbolic argument $s \in U = [0, 1]$ and we write $Z \sim f_Z(s)$.

We restrict ourselves to normalized non-negative probabilities, that is $\sum_i^{\infty} p_i = 1$ with $p_i \geq 0$, thus implying that $f_Z(1) = 1$. Most of the results that will be shown later rely on this arguments that are summarized by the pgf theorem [42].

Theorem 4.1.1. (The pgf theorem) *Let Z be a non-negative random variable with pgf $f_Z(s)$, then*

- *f_Z is non-negative and continuous with all derivatives on $[0, 1)$ and it is increasing and convex.*
- $p_k = \frac{1}{k!} f_Z^{(k)}(0) = \frac{1}{k!} \frac{d^k f_Z(s)}{ds^k}|_{s=0}$
- *the k-th factorial moment of Z, $\mu_k = \langle Z(Z-1)...(Z-k+1) \rangle$, is finite if and only if $f^{(k)}(1^-) = \lim_{s\to 1^-} f_Z^{(k)}(s)$ is finite. In such case, $\mu_k = f^{(k)}(1^-)$.*
- *if Z_1 and Z_2 are two independent non-negative random variables, $f_{Z_1+Z_2}(s) = f_{Z_1}(s) f_{Z_2}(s)$.*

- *if* $Z = \sum_{i=1}^{Y} X^{(i)}$ *where* Y *is a non-negative random variable and* $\{X^{(i)}, i \geq 1\}$ *is a sequence of iid non-negative random variables independent of* Y*, then* Z *has the pgf* $f_Z(s) = f_Y(f_{X^{(1)}}(s))$

4.2 Classifications of branching processes

In the preceding section, a mathematical formulation of branching process theory has been given. Here we discuss the parameters that define a given BP, enumerating all the different possibilities.

One of the important notions in the theory of branching processes is that of the *type space*. The type space S is the set, which can be finite, denumerable, or a continuum, of all possible varieties of particles included in the process, that is

$$S = \begin{cases} \{1\} & \text{single type} \\ \{1, 2, ..., N\} & \text{multi type} \\ \{1, 2, ...\} & \text{denumerable type} \\ \mathbb{R}, \mathbb{R}^+, [0, 1] & \text{continuous type} \end{cases}$$

Particles of a given type may produce particles of different types. Restrictions on type transitions, as well as on the type space, lead to differing properties of resulting processes.

The second parameter to be set is *lifetime* τ. It can assume a fixed value or can be a random variable. The simplest case is the so called Galton-Watson process in which lifetime is conventionally fixed to 1, that means that subprocesses are equally spaced in time. This will be discussed later in more details. A toy model to develop more complicated processes involve exponential lifetime distributions. This law for lifetime distributions is not well motivated by any biological assumption, however it leads to computable expressions. The Bellmann-Harris process is instead a more general model in which τ is a non-negative random variable and bring to an "age

dependent" dynamic (Ref. [23], page 65).

A branching process is also classified according to its degree of *criticality*, that is the asymptotic behavior of a process. In a BP a fundamental role is played by the "order parameter" $m = \langle X \rangle$ that represents the main progeny count of an individual. This concept is simply understood in the Galton-Watson process that will be discussed in this chapter, in which a relation between the mean number of particles $\langle Z_t \rangle$ at time t and m holds such that (it will be shown later in the case of the Galton-Watson process)

$$\langle Z_t \rangle \equiv m_t \sim m^t. \tag{4.3}$$

A rigorous approach can be found in Harris' "The Theory of Branching Processes" [40]. Therefore the criticality is defined by the value of m in a given process, in facts

$$\begin{aligned}
&\langle Z_t \rangle \to \infty \qquad \text{if } m > 1 \quad \text{(supercritical case)} \\
&\langle Z_t \rangle = 1 \qquad \text{if } m = 1 \quad \text{(critical case)} \\
&\langle Z_t \rangle = 0 \qquad \text{if } m < 1 \quad \text{(subcritical case)}
\end{aligned}$$

The parameter m is also connected to the probability of eventual extinction $q = q_{t\to\infty} < 1$, given $q_t = P(Z_t = 0|Z_0 = 1) = f_t(0)$ the extinction probability of a process at the time t (the notation $f_t(0)$ stands for $f_{Z_t}(0)$ having dropped Z because it is redundant). In facts it can be shown (Ref. [41], page 4) that q is the unique root in $[0, 1)$ of the fixed point equation $q = f(q)$ and

$$\begin{aligned}
&q < 1 \qquad \text{if } m > 1 \quad \text{(supercritical case)} \\
&q = 1 \qquad \text{if } m = 1 \quad \text{(critical case)} \\
&q = 1 \qquad \text{if } m < 1 \quad \text{(subcritical case)}
\end{aligned}$$

Further, $P(Z_t \to \infty \text{ as } t \to \infty|Z_0 = 1) = 1 - q$. Therefore a subcritical or critical process becomes surely extinct, while the extinction of a supercritical process is defined by q.

Here should be addressed some questions about the system we are interested in. A cell cluster is the product of a series of divisions that have been started by a single ancestor cell. What can we say about the type space, the lifetimes and the criticality? Some biological observations justify some choices, however assumptions are needed also in relation to the cell type and to the environment. For example a skin cell cluster can grow and then stop while a cancer cell cluster grow indefinitely, the growth can be favored if immersed in a nutrient solution or inhibited if immersed in other substances.

A distribution for the lifetimes τ cannot be determined *a priori*. However, according to biological studies, cells have an average duplication rate that measures the number of duplication in a day. Following this line, we will set a conventional fixed value for the lifetime $\tau = 1$, thus considering Galton-Watson processes. Observe here that a conversion factor between a unitary time step and the real time is not determined.

Here we will consider 39 wild type Melanoma cells immersed in nutrient solution. They display a high duplication rate and do not stop duplicating. This is a strong information about criticality because we see that $\langle Z_t \rangle$ is an increasing function of time t, thus implying that we are in the supercritical case. We should thus find that $\langle Z_t \rangle \sim m^t$ with $m > 1$.

The most interesting parameter in our particular case is the type space. In facts the distinction between the Traditional Cancer (TC) Theory and the Cancer Stem Cell (CSC) Theory relies on the number of particle considered. The first one involve a single population of cells, while the second one propose the existence of a subpopulation of cells that sustain one or more populations of cells.

Summarizing, we will study the growth of cancer cell clusters as Galton-Watson ($\tau = 1$) branching processes in the supercritical regime ($m > 1, q < 1$), varying the type space and the sets of probabilities $\{p_i^\alpha\}_{\alpha \in S, i \geq 0}$.

4.3 Single type process: the TC Theory

Traditional cancer theory states that all cancer cells are similar and can therefore be modeled by a single particle type. Therefore here is discussed the simple case of a single type population with $\tau = 1$ in the supercritical case. This shows how the mathematical approach defined in the first section of this chapter is applied to a specific situation. Furthermore, properties of this process provide intuitions about more complicated branching processes. We will see that even in this simple case we can obtain analytical results only in calculating the asymptotic behaviors, therefore simulations are inevitably needed. We first review the general approach, due to Galton and Watson, then we focus our attention on a specific model based on simple biological observation.

In considering $\tau = 1$, the continuous time index of Z_t can be replaced by a discrete index n that labels the number of generation. Therefore Z_n where $n = 0, 1, 2, ...$ are the particle counts in the n-th generation. Let $\{X_{n,k}\}_{n\geq 0, k\geq 1}$ be an array of nonnegative integer valued random variables that are i.i.d. (independent and identically distributed) with a probability distribution $\{p_i\}_{i\geq 0}$, where $X_{n,k}$ represents the number of progeny of the k-th particle existing in generation n. Let $Z_0 = 1$ be the starting number of progeny, therefore

$$Z_{n+1} = \begin{cases} \sum_{k=1}^{Z_n} X_{n,k} & \text{if } Z_n > 0 \\ 0 & \text{if } Z_n = 0 \end{cases} \tag{4.4}$$

According to the fifth point of theorem 4.1.1, the probability generating function f_{n+1} of Z_{n+1} is given by

$$f_{n+1}(s) = f_n(f_1(s)) = f_n(f(s)) = \underbrace{f(f(...f}_{\text{n times}}(s))) \tag{4.5}$$

having used the fact that $f_0(s) = s$ because $Z_0 = 1$. Note here that for an arbitrary

Z_0 we have that $f_0(s) = s^{Z_0}$, however we are modeling cell clusters formed by a single ancestor cell, i.e. $Z_0 = 1$. Thus we have that

$$\langle Z_1 \rangle = f'(1^-) = \sum_{j=0}^{\infty} j p_j \equiv m \tag{4.6}$$

and

$$\langle Z_n \rangle = f_n'(1^-) = [f_{n-1}(f_1(s))]'|_{s=1^-} = f_{n-1}'(1^-) f'(1^-) = ... = m^n \tag{4.7}$$

according to equation 4.3.

Here we considered average quantities, however interesting results concern the behavior of the random variables themselves. Recall that if $S_n = X_1 + X_2 + ... + X_n$ is the sum of i.i.d, random variables then the Law of Large Numbers states that S_n/n converges to a constant, namely $\langle X_1 \rangle$ [42]. A similar limiting theorem for branching processes exists. If Z_n represents the number of offspring after n generations, we have seen that the expected value of Z_n is m^n. Thus we can scale the random variable Z_n to have expected value 1 by considering the random variable

$$W_n = \frac{Z_n}{m^n}. \tag{4.8}$$

The analogous Law of Large Numbers for branching processes is formulated as follows (Ref. [23], page 45).

Theorem 4.3.1. *If* $0 < m < \infty$ *then there exists a random variable* W *such that* $\lim_{n\to\infty} W_n = W$ *with probability* 1.

However, unlike the case of the Law of Large Numbers where this limit is a constant, for a branching process the limiting value of the random variables W_n is itself a random variable. Other interesting results concern the supercritical case that corresponds to our situation [43].

Theorem 4.3.2. (*supercritical case) If* $m > 1$, $\sigma^2 = \langle (Z_1 - m)^2 \rangle < \infty$ *and* $Z_0 = 1$ *then*

- $\lim_{n\to\infty}\langle(W_n - W)^2\rangle = 0$
- $\langle W\rangle = 1, \sigma_W^2 = \langle(W - \langle W\rangle)^2\rangle = \sigma^2/(m^2 - m)$
- $P(W = 0) = q = P(Z_n = 0 \textit{ for some } n)$

The second condition is very important and states that in the supercritical case if fluctuations are finite, $\langle Z_n\rangle \sim m^n$ and $\sigma_{Z_n}^2 = \langle(Z_n - \langle Z_n\rangle)^2\rangle \sim \frac{\sigma^2}{(m^2-m)}m^n$ asymptotically ($n \to \infty$).

Observe that here we considered the simplest class of problems in the context of Branching Process Theory, however also in these cases only asymptotic behavior can be calculated for average quantities. This will be discussed further in the subsequent section.

4.3.1 A model for TC Theory

The basic feature of TC Theory is the existence of a unique population of cells, however we must define a specific process between the Galton-Watson-like processes able to fits our particular situation. Therefore some restriction will be considered according to biological observations. Indeed cells can divide through mitosis or die, but cannot produce two or more offsprings. Our process is thus defined by the set of probability $\{p_i\}_{i=0,1,2}$ (such that $\sum_i p_i = 1$) where p_0 corresponds to the probability of a cell death, p_1 is the probability of being quiescent (not dividing) and p_2 is the probability of a cell mitosis (see figure 4.2). The quiescence keeps track of two effects: the real quiescence of a cell, i.e. a cell do not divide in a time step, and a second one that deal with time fluctuations, that is varying p_1 we set the speed of the branching process.

The probability generating function is thus given by (cfr. equation 4.1)

$$f(s) = p_0 + p_1 s + p_2 s^2 \tag{4.9}$$

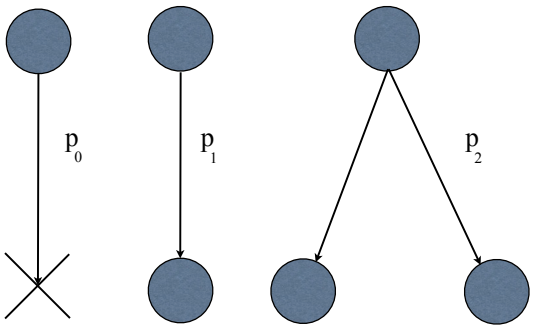

Figure 4.2: Here is a schematic depiction of the model based on the TC Theory: p_0 denotes the probability of cell death, p_1 is the probability of quiescence and p_2 represents the duplication probability.

and according to theorem 4.1.1

$$m = f'(1^-) = p_1 + 2p_2 = 1 - p_0 + p_2 \tag{4.10}$$

having used the fact that $\sum_i p_i = 1$. In the supercritical case ($m > 1$) we find that $\frac{p_0}{p_2} < 1$. This is obvious because, in a process of growth, duplication must be more likely than death.

Further the fixed point equation for the probability of eventual extinction $q \in [0, 1)$ is

$$q = p_0 + p_1 q + p_2 q^2 \tag{4.11}$$

that as solution $q = \frac{p_0}{p_2}$ in the permitted range of q if $\frac{p_0}{p_2} < 1$. Note here that a supercritical process implies that the fraction of clusters that stop growing in the asymptotic limit in the wells considered is less than one, i.e. $q < 1$.

Recall that we are interested in a set of branching processes, each one started from a single cell, that give rise to clusters. With the method implemented in the first chapter the number of cells in a well or in a given cluster can be calculated. According to the BP Theory, the average number of progeny of a single ancestor is

$$\langle Z_n \rangle = m^n \tag{4.12}$$

for sufficiently large n and corresponds to the volume of the cluster, i.e. $\langle V \rangle = \langle Z_n \rangle$. Further if we consider the entire well with a given rescaled density ρ (as defined in the preceding chapters) we should measure a number of cells in a well equal to

$$\langle Z_n^*(\rho) \rangle = m^n \rho. \tag{4.13}$$

The expected error when measuring these quantities is however huge as follow from theorem 4.3.2. In facts

$$\sigma_n^2 = \langle Z_n^2 \rangle - \langle Z_n \rangle^2 = \frac{\sigma^2}{m^2 - m} m^n = \frac{\sigma^2}{m^2 - m} \langle Z_n \rangle. \tag{4.14}$$

where $\sigma^2 = \langle Z_1^2 \rangle - \langle Z_1 \rangle^2$ is calculated using theorem 4.1.1, i.e.

$$\sigma^2 = f^{(2)}(1^-) + \langle Z_1 \rangle - \langle Z_1 \rangle^2 = 2p_2 + m - m^2 \tag{4.15}$$

Therefore combining 4.15 and 4.14 we find that

$$\sigma_n^2 = \left(\frac{2p_2}{m^2 - m} - 1 \right) \langle Z_n \rangle \tag{4.16}$$

A straightforward calculation shows that $\sigma_n^2 = 0$ only in the trivial deterministic case defined by $p_1 = 1$ and $p_0 = p_2 = 0$.

Note here that the behavior of Z_n is correct in the asymptotic limit, however the number of division n is limited by experimental reasons, in facts we saw that considering times greater than 10 days results in prohibitively large clusters.

Note also that here we have more information in experimental data, in facts we can compute also a distribution $P(Z_n = V)$ for the volumes of clusters as discussed in the preceding chapters. However this does not have any analytical counterpart, therefore simulations are needed to calculate the distribution.

4.4 Multi-type process: the CSC theory

The basic feature of the CSC theory is the existence of a subpopulation of cells able to sustain the growth of the tumor. We address here the question if a model in the context of multi-type branching processes is possible. First the basic concepts of multi-type branching processes will be reviewed, the behavior of which is a direct extension of the single-type case, then some biological and mathematical results will be discussed in order to develop a model.

The main goal here is to define the behavior of a population of individuals of k types. Despite the case of a single-type BP in which the first nucleation site is a cell of the single type considered, here an initial condition must be defined, because cells of different types give rise to different processes. Consider an ancestor particle of type i that after one time step die producing a random number of progeny particles of k types as our initial condition. Thus in the second generation we have k different subprocesses with k different dynamics and distribution of this subprocess depends only on the type of the ancestral particle.

Here should be emphasized that multi-type BP theory that will be exposed in this section is developed assuming that the type of the first nucleation site is known. This will be kept in account supposing that the initial condition is not random. However such assumption is not obvious in real situation. In facts when cells are displaced in a well, it is not trivial to know the type of a given cell.

The total number of particles at time n in the process started by an ancestor of a fixed type constitute a random vector $\mathbf{Z}_n = (Z_n^1, ..., Z_n^k)$, where the components $\{Z_n^{(i)}\}_{1 \leq i \leq k}$ denote the number of particles of type i in the n-th generation and whose distribution depends on the type of the ancestral particle of the process.

Denote with T the set of all k-dimensional vectors whose components are non-negative integers. Let $\{\mathbf{e}_i\}_{1 \leq i \leq k}$, denote the vector whose i-th component is 1 and whose other components are 0. The multitype (or vector) Galton-Watson process is a temporally homogeneous vector-valued Markov process $\mathbf{Z}_0, \mathbf{Z}_1, \mathbf{Z}_2, ...$ whose states

are vectors in T.

If $\mathbf{Z}_0 = \mathbf{e}_i$, then $\mathbf{Z}_1$ will have the generating function

$$f_1^i(s_1, ..., s_k) = \sum_{r_1,...,r_k}^{\infty} p_i(r_1, ..., r_k) s_1^{r_1} \cdots s_k^{r_k} \quad |s_1|, ..., |s_k| \leq 1 \tag{4.17}$$

where $p_i(r_1, ..., r_k)$ is the probability that an object of type i has r_1 children of type $1, ..., r_k$ of type k. In general, if $\mathbf{Z}_n = (r_1, ..., r_k) \in T$, then $\mathbf{Z}_{n+1}$ is the sum of $r1 + ... + r_k$ independent random vectors, r_1 having the generating function $f^{(1)}$, ..., r_k having the generating function $f^{(k)}$. If $\mathbf{Z}_n = 0$, then $\mathbf{Z}_{n+1} = 0$.

The generating function of $\mathbf{Z}_n$, when $\mathbf{Z}_0 = \mathbf{e}_i$, will be denoted by $f_n^i(s_1, ..., s_k) = f_n^i(\mathbf{s})$ where $1 \leq i \leq k$ and $n = 0, 1, ...$ The vector $(f_n^1(\mathbf{s}), ..., f_n^k(\mathbf{s}))$ will be frequently denoted by $\mathbf{f}_n(\mathbf{s})$. Here the multi-type counterpart of equation 4.5 is expressed by the subsequent theorem [23].

Theorem 4.4.1. *The generating functions f_n^i are functional iterates, defined by the relations*

$$\begin{aligned} f_n^{i+1}(s) &= f^i(f_n^1(\mathbf{s}), ..., f_n^k(\mathbf{s})) \qquad && n = 0, 1, 2, ... \\ f_n^0(\mathbf{s}) &= s_i && i = 1, 2, ..., k. \end{aligned} \tag{4.18}$$

In vector form

$$\mathbf{f}_{n+N}(\mathbf{s}) = \mathbf{f}_n(\mathbf{f}_N(\mathbf{s})) \quad n, N = 0, 1, 2, ... \tag{4.19}$$

The main progeny count m of the single type BP is replaced by the matrix $\mathbf{M} = [m_{i,j}]$ whose components are the expected number of progeny of type j of a particle of type i, i.e.

$$m_{i,j} = \langle Z_i^j \rangle_{\mathbf{Z}_0 = \mathbf{e}_i} = \frac{\partial f_1^i}{\partial s_j}(1, ..., 1) \quad i, j = 1, ..., k. \tag{4.20}$$

Using the chain rule in 4.18 we obtain $\langle \mathbf{Z}_{n+1} \rangle = \mathbf{Z}_n \mathbf{M}$ and iterating we get the

analogous of equation 4.7 for the multi-type case, i.e.

$$\langle \mathbf{Z}_{n+N} \rangle = \mathbf{Z}_N \mathbf{M}^n \tag{4.21}$$

The Frobenius-Perron theorem [23] is the most useful tool when dealing with powers of non negative and irreducibles matrixes (i.e. such that $\mathbf{M}^N$ is positive for some positive integer N). In facts it demonstrate that $\mathbf{M}$ has a unique positive eigenvalue ρ that is greater in absolute value than any other eigenvalue. Further, ρ corresponds to positive right and left eigenvectors $\mu = (\mu_i)$ and $\nu = (\nu_i)$, respectively, which are the only non-negative eigenvectors and $\mathbf{M}^n$ can be rewritten as

$$\mathbf{M}^n = \rho^n M_1 + M_2^n \quad n \in \mathbb{N} \tag{4.22}$$

where $\mathbf{M}_1 = (\mu_i \nu_j)$ with normalization $\sum_i \mu_i \nu_i = 1$, $\mathbf{M}_2$ is such that $\mathbf{M}_2 \mathbf{M}_1 = \mathbf{M}_1 \mathbf{M}_2 = 0$ and $\mathbf{M}_2^n = O(\alpha^n)$ for some $\alpha \in (0, \rho)$. Therefore $\mathbf{M}^n$ can be approximated by the n-th power of the maximum eigenvalue for enough large n, thus implying that

$$\langle \mathbf{Z}_n \rangle \sim \rho^n \mathbf{Z}_0 \mathbf{M}_1. \tag{4.23}$$

Here ρ plays the role of the "order parameter" played by m for the single-type BP. The supercritical case is indeed defined by $\rho > 1$.

Further the theorem 4.3.2 defined for the single type supercritical case has a multi type generalization [23].

Theorem 4.4.2. *Suppose that the process is positively regular with $\rho > 1$ and that all the second moments of progeny distributions are finite. Then, the random vectors $\mathbf{Z}_n/\rho^n$ converge with probability 1 to a random vector $\mathbf{W}$. Vector $\mathbf{W}$ is nonzero except for trivial cases of all covariance matrices $\mathbf{V}_i = Cov(\mathbf{Z}_1 | \mathbf{Z}_0 = \mathbf{e}_i)$ being zero or $\mathbf{Z}_0 = 0$. If $\mathbf{W}$ is nonzero, then with probability 1 its direction coincides with that of ν, the left eigenvector of $\mathbf{M}$.*

The probability of extinction has a multi-type generalization too, but in this case

it depends on the type of the ancestral particle. Therefore there is an extinction probability q_i for each process in which the nucleation site is a particle of type i, i.e. $q_i = P(\mathbf{Z}_n = 0 \text{ for some n } |\mathbf{Z}_0 = \mathbf{e}_i)$ where $1 \leq i \leq k$. Denoting $\mathbf{q} = (q_1, ..., q_n)$ and using the rule that for example $\mathbf{q} \geq 0$ means that $q_i \geq 0 \ \forall i$, a theorem is enunciated as follow [23].

Theorem 4.4.3. *Suppose that the process is positively regular (that is* $\mathbf{M}$ *is irreducible) and not singular (which would mean that each object has exactly one progeny). If* $\rho \leq 1$*, then* $\mathbf{q} = 1$*. If* $\rho > 1$*, then* $\mathbf{q} \in [0, 1)$ *and* $\mathbf{q}$ *satisfies the equation* $\mathbf{q} = \mathbf{f}(\mathbf{q})$*.*

4.4.1 A two population model

Here a two type branching process is considered. The goal is to show the asymptotic behavior of such a model. This case gives an insight into features of a two population dynamics in order to design a possible model able to describe the CSC theory.

The basic feature of a CSC model is the existence of a subpopulation of cells responsible of the growth of the tumor, that means that some cells sustain their own cell-type population and are able to give rise to cells belonging to a different population. Such a population will be denoted with S because of their stem-like features, while the second population of "common" cancer cells will be denoted with C.

As said in the preceding sections, the cell can duplicate, die or be quiescent. However if the existence of such a population is postulated, their cells cannot die otherwise the tumor will stop growing definitely. Meanwhile the second population could die. Therefore we consider a first population of cells S that has probability p_0 of being quiescent, p_1 of dividing in one C-cell and one S-cell and probability p_2 of dividing in two S cells. Instead the C-cells can die, be quiescent or duplicate in two C-cells. A scheme of this process is shown in figure 4.3.

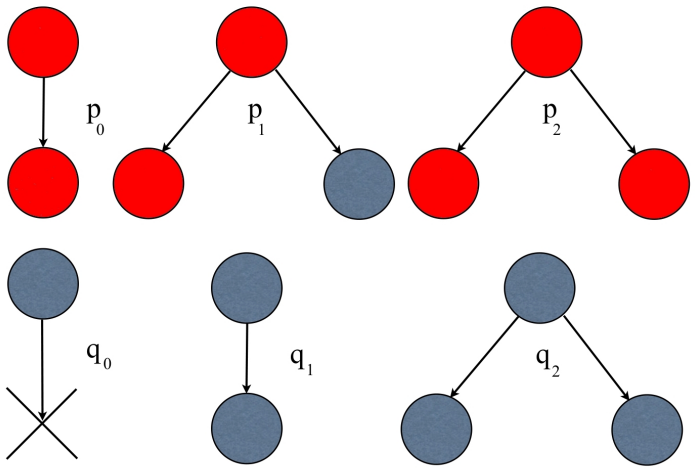

Figure 4.3: Here is a schematic depiction of the two-type toy-model for the CSC Theory.

The generating functions are given by (see equation 4.17)

$$f_1^S(x_S, x_C) = p_0 x_S + p_1 x_S x_C + p_2 x_S^2 \tag{4.24}$$

$$f_1^C(x_S, x_C) = q_1 x_C + q_2 x_C^2 \tag{4.25}$$

The behavior of this model is determined by the matrix $\mathbf{M}$ defined in equation 4.20, that is

$$\mathbf{M} = \begin{bmatrix} p_0 + p_1 + 2p_2 & p_1 \\ 0 & q_1 + 2q_2 \end{bmatrix} = \begin{bmatrix} 1 + p_2 & p_1 \\ 0 & 1 + \delta \end{bmatrix} \tag{4.26}$$

having denoted $\delta = (q_2 - q_0)$ and having used the normalization constraints $\sum_i p_i = \sum_i q_i = 1$. Note here that the case in which C-cells do not die is achieved simply setting $\delta = q_2$ ($q_0 = 0$).

In this case the matrix is not irreducible, thus neither the Frobenius-Perron theorem either theorem 4.4.2 apply. Therefore the number of cells must be computed by iteration using

$$\langle \mathbf{Z}_n \rangle = \mathbf{Z}_0 \mathbf{M}^n \tag{4.27}$$

We immediately find that

$$\mathbf{M}^n = \begin{bmatrix} (1+p_2)^n & B_n \\ 0 & (1+\delta)^n \end{bmatrix} \tag{4.28}$$

where B_n is computed observing that

$$\begin{aligned} \mathbf{M}^n &= \begin{bmatrix} (1+p_2)^n & B_n \\ 0 & (1+\delta)^n \end{bmatrix} = \\ &= \begin{bmatrix} (1+p_2)^{n-1} & B_{n-1} \\ 0 & (1+\delta)^{n-1} \end{bmatrix} \begin{bmatrix} (1+p_2) & p_1 \\ 0 & (1+\delta) \end{bmatrix} = \mathbf{M}^{n-1}\mathbf{M} \end{aligned} \tag{4.29}$$

that define the recursion relation $B_n = p_1(1+p_2)^{n-1} + (1+\delta)B_{n-1}$ with $B_1 = p_1$. This is solved analytically and in the case $p_2 \neq \delta$ we obtain

$$\mathbf{M}^n = \begin{bmatrix} (1+p_2)^n & \frac{p_1}{p_2-\delta}[(1+p_2)^n - (1+\delta)^n] \\ 0 & (1+\delta)^n \end{bmatrix} \tag{4.30}$$

For $p_2 = \delta$, we find that $B_n = \frac{p_1 n(n-1)}{2(1+\delta)}(1+\delta)^n$.

If we consider the case of a C nucleation cell, i.e. $\mathbf{Z}_0^C = (0,1)$, we find the trivial case of a single branching process with main progeny count $m = (1+\delta)$, as we should expect. A non trivial dynamics is found when the ancestor is an S-cell, i.e. $\mathbf{Z}_0^S = (1,0)$. In this case we find that

$$\langle \mathbf{Z}_n \rangle = \left((1+p_2)^n, \frac{p_1}{p_2-\delta}[(1+p_2)^n - (1+\delta)^n] \right) \tag{4.31}$$

and denoting F^S and F^C respectively the asymptotic fraction of S-cells and C-cells at a given time n we find that

$$F^S = \lim_{n\to\infty} \frac{\langle Z_n^S \rangle}{\langle Z_n^S \rangle + \langle Z_n^C \rangle} = \begin{cases} \dfrac{p_2-\delta}{p_2+p_1-\delta} & \text{if } p_2 > \delta \\ 0 & \text{if } p_2 \leq \delta \end{cases} \tag{4.32}$$

$$F^C = \lim_{n\to\infty} \frac{\langle Z_n^C \rangle}{\langle Z_n^S \rangle + \langle Z_n^C \rangle} = \begin{cases} \dfrac{p_1}{p_2 + p_1 - \delta} & \text{if } p_2 > \delta \\ 1 & \text{if } p_2 \leq \delta \end{cases} \tag{4.33}$$

This result shows that the there is a relation between the probabilities of the two processes or between the main progeny count of the two population. In facts if $p_2 \leq \delta$, the C population is overwhelming and the effects of the S population cannot be seen if we look at a given sample where cell populations are indistinguishable. The existence of two population is detectable only if $p_2 > \delta$ that corresponds to the situation in which the cells of the S population duplicate more than the cells of the C population.

Note here that the procedure achieved in this section is completely general and can be applied to every two population hierarchic branching processes, i.e. where a C-cell do not duplicate in one or more S-cells, or analogously whenever dealing with upper triangular matrixes. In facts we see that the case in which a C-cell do not die is obtained setting $q_0 = 0$ and thus $\delta = q_2$. Further if we consider the possibility that an S-cell divide in two C-cells with probability p_3, the dynamic is simply solved with the substitution $p_1 \mapsto p_1 + 2p_3$ in the preceding equations. Otherwise Frobenius-Perron theorem and theorem 4.4.2 provide a simple recipe to solve the dynamic of non-hierarchic two-type BP.

4.4.2 A CSC model

A growing number of papers appears on Biological reviews in the last years showing the evidences of a subpopulation of senescent cell in tumors [45, 46, 47, 48]. Cellular senescence is the phenomenon by which normal cells stop proliferating, usually after about 50 cell divisions in vitro, however they remain metabolically active. This last feature is the one that specify a difference from a dead cell. Such a process is called Hayflick phenomenon and the Hayflick limit is the number of times a normal cell population will divide before it stops [26].

Senescence can be kept into account introducing a time index for cells. When

cells are too old, that is the time index counts a certain number of divisions, then they stop dividing.

Therefore I consider a "three"-type model. The S-cells belong to the immortal CSC population, thus we suppose that they divide in two C-cells with probability (wp) p_0, one S-cell and one C-cell wp p_1, two S-cells wp p_2 and remain quiescent wp p_3. The C-cells that are offsprings of an S-cell has a lifetime $k = 1$, meaning that they are just born. The C population is constituted by "common" cancer cells that remain quiescent wp q_1 or divide wp q_2 and labeled with a time index k that keep track of the number of divisions of the cell, thus a k divisions old cell when duplicate give rise to two "older" cells with a time index equal to $k + 1$. Instead if the C-cell do not divide, the time index remains constant. When a C cell gets too old, that is his division time k reach a certain value M, when divides his offspring is constituted by two senescent D-cell that will not divide anymore.

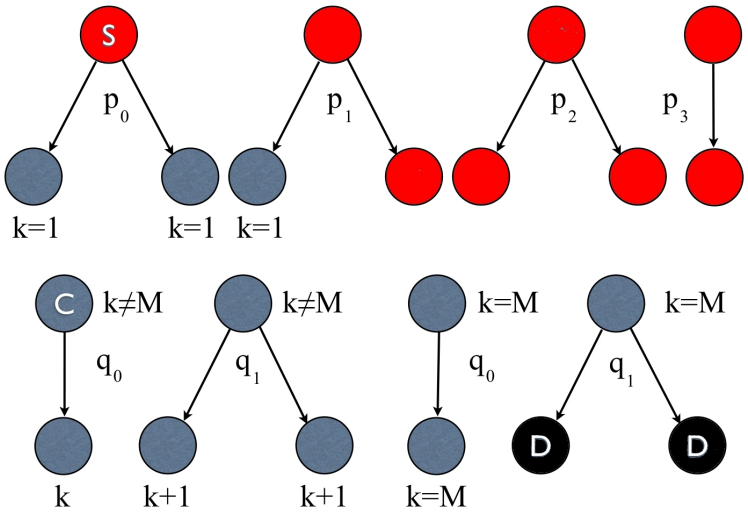

Figure 4.4: Here is a schematic depiction of the the CSC model implemented.

This is not exactly a three-type model, as said before, in facts there is the S and the D population and M different kind of C cells. This is therefore an $(M+2)$-type model. The generating functions are computed using equation 4.17 and the matrix $\mathbf{M}$ follows from equation 4.20. Therefore we get the upper triangular block diagonal

matrix (on the left of the matrix the respective cell types of the row considered are listed, having denoted $\{C^{(k)}\}_{1\leq k\leq M}$ the C population with division time k)

$$\mathbf{M} = \begin{bmatrix} (p_1+2p_2+p_3) & 2p_0+p_1 & 0 & ... & & \\ 0 & q_0 & 2q_1 & 0 & .... & \\ 0 & 0 & q_0 & 2q_1 & 0 & ... \\ ... & & & ... & & \\ .... & & & & q_0 & 2q_1 \\ 0 & .... & & & 0 & 1 \end{bmatrix} \begin{matrix} S \\ C^{(1)} \\ ... \\ ... \\ C^{(M)} \\ D \end{matrix} \tag{4.34}$$

The Frobenius-Perron theorem and theorem 4.4.2 do not apply in this case too, therefore we have to compute the number of cells by iteration. Denote $1+\epsilon = p_1+2p_2+p_3 = 1+(p_2-p_0)$, that corresponds to the main progeny count for the S population (i.e. $m_S = 1+\epsilon$) and thus $2p_0+p_1 = 1-\epsilon-p_3$. In this way the parameter ϵ define the capability of the S population of sustaining the proliferation, that is for $\epsilon > 0$ all the cells do not stop proliferating, otherwise the BP will surely end. In our case the process do not end, therefore $\epsilon > 0$ will be assumed.

The n-th power of $\mathbf{M}$ can be computed as in the case of the two-type process considered in the preceding subsection imposing a condition analogous to 4.29. Denoting $\{B^{(n)}_{i,j}\}_{n\in\mathbb{N},2\leq i,j\leq M+2}$ the entries i,j of the upper triangular matrix $\mathbf{M}^n$ diagonal excluded, we get

$$B^{(n)}_{i,j} = \begin{cases} (1-\epsilon-p_3)(1+\epsilon)^{n-1} + q_0 B^{(n-1)}_{12} & \text{if } i=1, j=2 \\ 2q_1 B^{(n-1)}_{1,j-1} + q_0 B^{(n-1)}_{1,j} & \text{if } i=1, j\geq 3 \\ q_0^{n-1} 2q_1 + q_0 B^{(n-1)}_{i,j} & \text{if } i\geq 2, j=i+1 \\ 2q_1 B^{(n-1)}_{i,j-1} + q_0 B^{(n-1)}_{1,j} & \text{if } i\geq 2, j>i+1 \end{cases} \tag{4.35}$$

with the boundary conditions

$$B_{i,j}^{(1)} = \begin{cases} (1-\epsilon-p_3) & \text{if } i=1, j=2 \\ 0 & \text{if } i=1, j\geq 3 \\ 0 & \text{if } i\geq 2, j=i+1 \\ 2q_1 & \text{if } i\geq 2, j>i+1 \end{cases} \tag{4.36}$$

The most interesting case is the one in which the nucleation site is an S cell, the other cases being processes that will inevitably end. In facts in the CSC theory the S cells are the responsibles of the growth of the tumor. We are thus interested in $\langle \mathbf{Z}_n \rangle = \mathbf{Z}_0^S \mathbf{M}^n = (1,0,...,0)\mathbf{M}^n$, that corresponds to the first row of $\mathbf{M}^n$, when $\epsilon > 0$. Observe that we should require that $1-\epsilon-p_3 = 2p_0+p_1 > 0$ and this is not achieved only in the trivial case $p_0 = p_1 = 0$, that is an S cell do not divide in a C cell. In this case the solution is given by

$$\begin{aligned}
\langle Z_n^S \rangle =& (1+\epsilon)^n \\
\langle Z_n^{C^{(k)}} \rangle =& \frac{1-\epsilon-p_3}{1+\epsilon-q_0}\left(\frac{2q_1}{1+\epsilon-q_0}\right)^{k-1}\left[(1+\epsilon)^n - q_0^n\right] \\
& - \frac{1-\epsilon-p_3}{1+\epsilon-q_0}\left(\frac{2q_1}{q_0}\right)^{k-1} q_0^n(1-\delta_{1k}) \sum_{j=0}^{k-2} \binom{n}{k-1-j} \left(\frac{q_0}{1+\epsilon-q_0}\right)^j \\
\langle Z_n^D \rangle =& \frac{1-\epsilon-p_3}{1+\epsilon-q_0} 2q_1 \left(\frac{2q_1}{1+\epsilon-q_0}\right)^{M-1}\left[\frac{(1+\epsilon)^n-1}{\epsilon} - \frac{1-q_0^n}{1-q_0}\right] \\
& - \frac{1-\epsilon-p_3}{1+\epsilon-q_0} 2q_1 \left(\frac{2q_1}{q_0}\right)^{M-1} \sum_{j=0}^{M-2} \binom{n}{M-j} \left(\frac{q_0}{1+\epsilon-q_0}\right)^j
\end{aligned} \tag{4.37}$$

that reduce in the asymptotic limit ($n \to \infty, \epsilon > 0, q_0 < 1$) to

$$\begin{aligned} \langle Z_n^S \rangle &= (1+\epsilon)^n \\ \langle Z_n^{C^{(k)}} \rangle &\simeq \frac{1-\epsilon-p_3}{1+\epsilon-q_0} \left(\frac{2q_1}{1+\epsilon-q_0} \right)^{k-1} (1+\epsilon)^n \\ \langle Z_n^D \rangle &\simeq \frac{1-\epsilon-p_3}{\epsilon} \left(\frac{2q_1}{1+\epsilon-q_0} \right)^{M} (1+\epsilon)^n \end{aligned} \tag{4.38}$$

having used the fact that for $n \to \infty$

$$\frac{1}{(1+\epsilon)^n} \sum_{j=0}^{k-2} \begin{pmatrix} n \\ k-1-j \end{pmatrix} \left(\frac{q_0}{1+\epsilon-q_0} \right)^j \sim \frac{1}{(k-1)\Gamma(k-1)} \frac{n^{k-1}}{(1+\epsilon)^n} \to 0 \tag{4.39}$$

The total number of C cells is computed as $\langle Z_n^C \rangle = \sum_{k=1}^{M} \langle Z_n^{C^{(k)}} \rangle$ and in the asymptotic limit we obtain

$$\langle Z_n^C \rangle \simeq (1-\epsilon-p_3) \frac{\left[1 - \left(\frac{2q_1}{1+\epsilon-q_0} \right)^{M-1} \right]}{(1+\epsilon) - (2q_1+q_0)} (1+\epsilon)^n \tag{4.40}$$

The number of cells grow exponentially with n for all cell type, because $\epsilon > 0$, thus the fraction of cells for a given cell type remains constant during the time. The asymptotic fraction of cells F^T of cell type $T \in \{S, C, D\}$ give informations on the composition of a given sample of cells and is computed as

$$F^T = \lim_{n \to \infty} \frac{\langle Z_n^T \rangle}{\sum_i \langle Z_n^i \rangle}. \tag{4.41}$$

Further constraints must be set to match with experimental biological results. In fact, in general cases the number of divisions after which a cell become senescent is about 50. Thus M is supposed to be enough large that we can consider $M \to \infty$. Therefore two cases must be distinguished: the first one in which the S cells are more likely to duplicate than the C cells that corresponds to the condition $2q_1 <$

$1+\epsilon-q_0$, the second one is the opposite situation and corresponds to $2q_1 > 1+\epsilon-q_0$. Summarizing we find that for $n \to \infty$, and after $M \to \infty$ under the conditions of non trivial dynamic $\epsilon > 0, 1-\epsilon-p_3 > 0$

$$
\begin{aligned}
F^S &\simeq \begin{cases} \dfrac{\epsilon - q_1}{(2-p_3)-(1+q_1)} & \text{if } 2q_1 < 1+\epsilon-q_0 \\ 0 & \text{if } 2q_1 > 1+\epsilon-q_0 \end{cases} \\
F^C &\simeq \begin{cases} \dfrac{(1-\epsilon-p_3)}{(2-p_3)-(1+q_1)} & \text{if } 2q_1 < 1+\epsilon-q_0 \\ \dfrac{\epsilon(1+\epsilon-q_0)}{2q_1[(2q_1+q_0)-(1+\epsilon)]+\epsilon(1+\epsilon-q_0)} & \text{if } 2q_1 > 1+\epsilon-q_0 \end{cases} \\
F^D &\simeq \begin{cases} 0 & \text{if } 2q_1 < 1+\epsilon-q_0 \\ \dfrac{2q_1[(2-p_3)-(1+q_1)]}{2q_1[(2q_1+q_0)-(1+\epsilon)]+\epsilon(1+\epsilon-q_0)} & \text{if } 2q_1 > 1+\epsilon-q_0 \end{cases}
\end{aligned}
\tag{4.42}
$$

In the case of high proliferation rate of S cells ($2q_1 < 1+\epsilon-q_0$) according to the two population model of the preceding subsection, the existence of D cells cannot be seen while the fractions of S and C cells are defined by ϵ, p_3 and q_1. Instead if the C cells have an higher proliferation rate than S cells, we are not able to see S cells, leaving place to a population of high fraction of C and D cells, defined by ϵ, p_3 and q_1 . However this is true if the sample is "small", in facts the limit $M \to \infty$ is not completely exact. We should say that in the first case it is not likely to see D cells and in the second case it is not likely to see S cells.

Chapter 5

Numerical simulations and results

In the preceding chapter, we designed models in the context of BP theory according to experimental results in Biology. We address here the question if these models reproduce the results obtained using the methods developed in the first chapters. We will consider the case of Melanoma 39 wild type samples in order to avoid any possible effect of the marker. However, even if we deal with a particular kind of cell, this method is completely general and could be extended to other different cell types.

Here we should emphasize that in our experimental data observables are computed in cells unit, thus the results can be directly compared with BP models. However BP analytical results can be achieved only in the asymptotic limit, therefore simulations are inevitably needed. In facts with our experimental setup we are able to determine a shape for the distribution of cluster volumes. This is a strong instrument because it gives information on all the volume composition of a given sample.

We will first consider the case of a single-type BP, that corresponds to the TC theory, then we will discuss the case of a CSC model.

5.1 TC model

Here we will consider the single-type TC model described in the preceding chapter in comparison with the experimental data. Here we address the question if a TC model is able to fits experimental measurements. We will find that such a model is not suitable, however it represents an interesting starting point to understand the dynamics and a powerful instrument to verify the existence of two population.

What we can effectively compute analytically is the scaling behavior of the average cluster size, defined by equation 4.12, i.e.

$$\langle Z_n \rangle \sim m^n \tag{5.1}$$

and the number of cells in a well defined by equation 4.13, i.e.

$$\langle Z_n^*(\rho) \rangle \sim m^n \rho. \tag{5.2}$$

However the process cannot be univocally fixed because both the average number of divisions n and the main progeny count m are not known in these equations.

Thus we should look at the distribution of cluster volumes discussed in the last section of the third chapter to plug this gap. Following this line $\langle Z_n^{well}(\rho) \rangle$ counts the number of cells in a well but does not keep track of the effective volumes reached by the clusters. Therefore a sample constituted by many separate cells or a sample of few big clusters cannot be distinguished by such a measure.

Further the expressions of equation 5.1 and 5.2 are almost useless for our purposes. In facts they keep track of the fact that some cells or clusters could disappear because of cell death during all the branching process. This is because of the dependence of m on p_0, i.e. $m = 1 + (p_2 - p_0)$. The average expressed by such equations is computed on all the cluster sizes and clusters that disappear because of cell death count as a cluster composed by 0 cells. Instead when measuring cluster volumes or the total number of cells in a well only averages on all the existing clusters can

be computed. Therefore the measurable quantity is the average cluster volume in a sample

$$\langle Z_n \rangle_{well} = \sum_{i=1}^{n_c} Z_n^i, \tag{5.3}$$

where n denote the number of divisions and n_c the number of clusters, and the average number of cells in a well

$$\langle Z_n^*(\rho) \rangle_{well} = \langle Z_n \rangle \rho. \tag{5.4}$$

These value cannot be predicted with analytical calculation but only computed with simulations.

The goal is thus to find the model that fits the distributions of cluster volumes at 8 and 10 day and that match with the predictions of equation 5.2. The parameters that must be varied are the probabilities $\{p_i\}_{i=0,1,2}$ that define the mean progeny count, i.e. in this process $m = 1 + (p_2 - p_0)$, and the average numbers of divisions n, that is a function of time ($n = n(t)$). Thus time enters in equation 5.2 and define the distribution of cluster volumes.

The theoretical distribution $P(V)$ of cluster volumes V is determined with a simulation of the process. The process is simulated on a sample of 10^6 cells, then the distribution is computed using the logarithmic binning method (even if here it is not necessary, but the comparison of the results will get across). The distribution will be peaked on higher values of V as p_2 approach to 1, that is $m \to 2$, and will be a δ function peaked on 2^n when $p_2 = 1$, that is the limiting deterministic case. Meanwhile when $m \to 1^+$, $P(V)$ will be peaked on low values of V.

In right figure 5.1 we show different curves obtained varying the set of probabilities $\{p_i\}_{i=0,1,2}$ with fixed n. In none of these cases the experimental bump in $P(V)$ for low values of V after 10 days shown in the left figure 5.1 can be detected. However this represents a track of the existence of two populations, the majority of which duplicate faster than a second little population that give rise to small volume

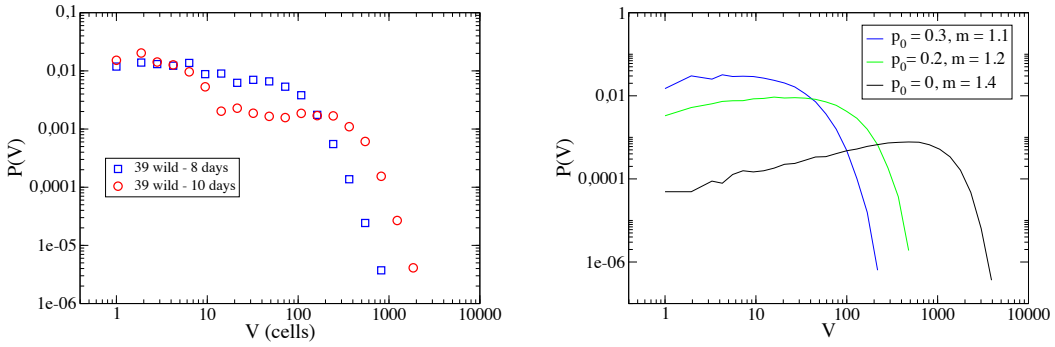

Figure 5.1: Left: the experimental curves for normalized probabilities $P(V)$ at 8 and 10 days obtained from the sample of Melanoma 39 wild type cells. Right: the curves represent the normalized distribution $P(V)$ of cluster volumes for different set of parameters $\{p_i\}_{i=0,1,2}$ fixing the number of divisions $n = 20$.

clusters.

Therefore we consider two population, that will be conventionally denoted as S and C, that divide with two different sets of probabilities $\mathbf{p}^{S,C} = \{p_i^{S,C}\}_{i=0,1,2}$ as those defined for the TC single-type process (cfr. figure 4.2) and exist in the sample with two different concentration $\alpha_{S,C}$, such that $\alpha_S = 1 - \alpha_C$. This system can be defined as a two-type BP where the vector $\mathbf{Z}_0$, that define the initial condition, is a random variable and his distribution is defined by a binomial distribution, that is it is an S cell with probability α_S otherwise it is a C cell. In this case the matrix $\mathbf{M}$ is diagonal and is obtained by equation 4.26 setting $p_1 = 0$.

The curves have been fitted simulating the dynamic, that is varying $\alpha_S, p_0^{S,C}, p_1^{S,C}$ and the average number of divisions n while the constraints $\alpha_C = 1 - \alpha_S$ and $\sum_i p_i^{S,C} = 1$ fix the other parameters. Figure 5.2 shows that a process with parameters $\mathbf{p}^S = (0, 0.93, 0.07)$, $\mathbf{p}^C = (0.2, 0.4, 0.4)$ and $\alpha_S = 0.05$, $\alpha_C = 0.95$ fits the experimental curves for the cluster volume distribution. After 8 days the average number of time steps is $n_8 = 21$ while after 10 days we have that $n_{10} = 27$. Here should be emphasized that n does not correspond to the average number of divisions of the cells.

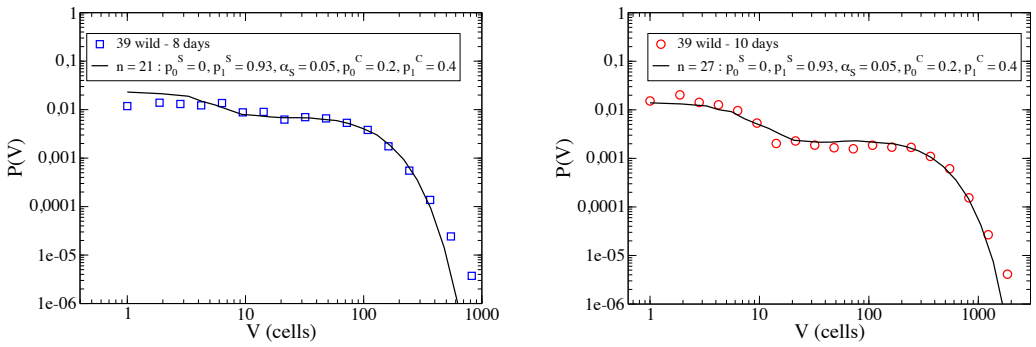

Figure 5.2: Here are shown the experimental distributions of volumes at 8 (left) and at 10 (right) days compared with the simulated curves. Both curves are obtained with the same sets of parameters $\mathbf{p}^S = (0, 0.93, 0.07)$, $\mathbf{p}^C = (0.2, 0.4, 0.4)$ and $\alpha_S = 0.05$, $\alpha_C = 0.95$. The only parameter from which they differ is the number of divisions n.

Further, these parameters should fit the data for the average cluster volume in a sample $\langle Z_n \rangle_{well}$ and the average number of cells in a well $\langle Z_n^*(\rho) \rangle_{well}$ obtained in the experiments. The predicted values for $\langle Z_n \rangle_{well}$ and $\langle Z_n^*(\rho) \rangle_{well}$ are computed running simulations on a sample of 10^6 cells, whose 5% evolves according to $\mathbf{p}^S$ and 95% evolves according to $\mathbf{p}^C$ and using equations 5.3 and 5.4.

Here we note that the average volumes $\langle Z_n \rangle_{well}$ and the distribution of volumes $P(\langle Z_n \rangle = V)$ are the most powerful instruments in cluster analysis, because they are not affected by experimental errors on density, whereas the number of cells in the well $\langle Z_n^*(\rho) \rangle_{well}$ depends by definition on the initial density ρ.

The results are shown in figures 5.3. The discrepancies for the 10-day data are due to the narrow range that can span the time t, in facts at high densities the clusters are so big that overlap while at low densities the samples count few data. Moreover clusters are enough big that are likely to be cut when selecting the circular area and erasing the black regions connected to the boundaries. This explain also why 10-day data underestimate the theoretical predictions for $\langle Z_n^*(\rho) \rangle_{well}$.

Summarizing we showed that the cell samples considered show two population dynamic features, whereas a TC theory alone do not explain the experimental re-

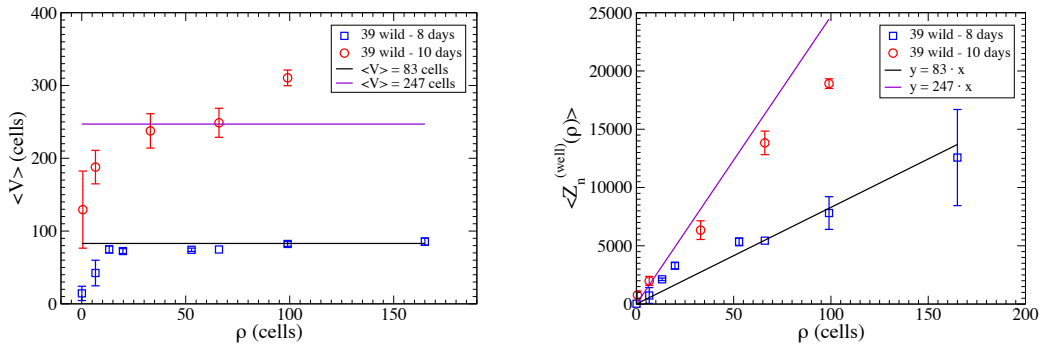

Figure 5.3: These two graphs represents respectively the average volume of clusters $\langle Z_n \rangle$ (left) and the average number of cells in a well $\langle Z_n^*(\rho) \rangle$ (right) as a function of density ρ. Symbols correspond to experimental measures, whereas straight lines represent theoretical results for the set of parameters that fits the distribution of volumes $P(\langle Z_n \rangle = V)$ shown in figure 5.2.

sults. Further the set of probabilities determined would be in agreement with the CSC hypothesis, that there exists a little subpopulation of cells endowed with the feature of being immortal ($p_1^S = 0$).

5.2 Towards a CSC theory

The experimental data show the existence of two independent populations in Melanoma cell samples. Hence, if the CSC hypothesis is considered, we should ask if a two population hierarchic model is able to fit these data. However when considering more than a single population we must deal with the problem of defining a starting condition, because multi-type BP are strongly dependent on. Furthermore a BP theory with random nucleation cell type has not been designed yet. Therefore assumptions are inevitably needed.

A reasonable approach consists in considering the case of a stationary starting condition, supposing that the cells of the patient have reached a steady state. Whatever this is not obvious, not least because the cells of the patient interact with the environment from which they are removed. If we consider such initial condition, the

type of the cells displaced in the wells must be determined according to the stationary cell fractions and the set of probabilities specific of a given cell type define the cluster growth process thus determining the cluster volume distributions.

Furthermore in the context of a multi-type BP the number of parameters is large. In the case of a CSC model, we have 5 free parameters, whereas the others are determined by the normalization conditions. This case is completely different from the two population non-hierarchic model discussed in the preceding section, indeed in this case the fit of the double exponential decay observed in the cluster volume distribution is achieved fitting separately the two independent BP.

Hypothesis on the initial conditions define strong constraints in the parameter space. Consider the case of a stationary initial condition in the context of the CSC model. Restrictions must be imposed if the existence of two population is tracked in the experiments. First of all, if we suppose that the fraction of D cells is not null, we should see many single cells in the well resulting in a high peak centered in $V = 1$ in the cluster volume distributions. This is not what we see, therefore we must consider the case of an high proliferation rate of the S cells ($2q_1 < 1 + \epsilon - q_0$).

Conclusion and outlook

In this thesis we studied Melanoma cell clusters in petri dishes with a Statistical Mechanics approach. The concepts developed represent a collection of methods to approach cell cluster analysis. This research is motivated by an increasing interest in the study of aggregation phenomena. Aggregates are the results of complex dynamics of a number of individuals. In the DLA model the patterns observed are the results of a diffusive system, whereas cell clusters forms as a consequence of cell division processes.

This topic deserves great attention in Biology, indeed a great challenge for scientists is the comprehension of tumor proliferation dynamics. The recent CSC theory, that support the existence of a minor subpopulation of cells endowed with stem-like features responsible of the proliferation of the tumor, suggested new therapies in cancer treatments and raised a great debate in this context. Conversely the traditional theory is based on a single population hypothesis. Understanding which are the real dynamics of tumor growth would be a striking discovery and a positive boost for improved cancer therapies.

To this purpose, we studied sets of cell colonies in petri dishes at different growth stages. We designed a systematic approach to the calculation of experimental observables developing an imaging technique. Experimental observations performed in Biology are improved with a computational method that allows exact measurements.

We developed methods to investigate static properties of clusters in order to test the confidence of the experimental setup and to verify the independence of cluster growth processes. We found that Melanoma cells form sparse independent clusters

and that the proliferation do not depends on the concentration of cells in the region in which they are confined. We studied the cluster geometries in order to discuss if the proliferation follows an isotropic trend. The experimental data are compared with random like clusters simulated with the Eden model, showing that the growth process is isotropic for Melanoma cells.

The randomness of the proliferation is the basic hypothesis in Branching Process Theory, where individuals behaves identically as all other individuals and independently of all other individuals. We reviewed the basic concepts of this theory and developed models to interpret experimental measurements. We enlighten double population dynamics in cell cluster growth and we detected one type of cells that constitutes the minority of all the cells for colonies of Melanoma cells. Both these results give credit to the CSC theory.

Summarizing in this project we discussed a possible method to approach cell cluster analysis. We studied geometrical aspects of cell clusters involving concepts of Classical and Statistical Mechanics, as well as Complex Systems models. Afterwards we studied the dynamics of cell proliferation in the context of Branching Process Theory.

This thesis represents a modest contribution towards the comprehension of the proliferation of cancer cells. It would be interesting to prove the existence of a hierarchic structure in cancer cell populations. This would indeed represents an additional striking confirmation of the CSC theory.

An interesting and natural expansion of this work would be the development of a graphical user interface to simply handle images of 2D compact clusters. The methods discussed in this thesis are based on physical concepts and would be very useful for non-physicists that would be able to extract useful informations for example from cell aggregates. Indeed we developed a systematic framework to assess properties of compact clusters, such as mutual independence and randomness in spatial growth, and to perform measurements, such as number of clusters and cluster volumes, useful to understand the dynamical behavior of these systems.

Appendix A

Technical details of the image conversion method

The image conversion method consists in a series of steps that are achieved in our research with the GNU image manipulating program GIMP. It is possible to implement most of them in a code but we preferred to keep them under control with a graphical interface.

- It happens that the well is not "clean" because of dust that settle and then fix with the solution of crystal violet and formalin or because of spots formed by evaporated water. This is solved blurring the impurities in the image.

- Thus a first edge detect is achieved with the DoG (difference of Gaussian) technique setting $r_1 >> r_2$ and $r_2 = 1$ px with the Invert option in order to obtain light violet spots on a grey-white background. At this point it is not possible to simply convert the violet spots in black spots using the threshold because of grey shades on the background, thus better improvement are needed.

- The contrast is increased using the Colorize option, where we fixed Hue = 240 and Saturation = 100.

- A second edge detect is then run without Invert, obtaining bright yellow spots on a black background that can be now removed selecting the black (there are no shades at this point) and filling the same area with white color. Now all the yellow spots are converted in black ones using the threshold option setting a threshold value of 0.

- At last all the cluster connected to the contour of the selection are cancelled. This is achieved to avoid any error when counting for example the volumes of the clusters or observables that involve the geometry that are clearly affected by cuts due to the selection considered.

Appendix B

Distribution of distances between random points

Consider now a circle A of radius r. Assume that two points are randomly thrown in A and we are interested in probability distribution function of the Euclidian distance x between them $P_r(x)$. To solve this problem we use Crofton fixed points theorem [17], that shows a way how to evaluate definite integrals without performing direct integration.

Theorem B.0.1. (Crofton fixed points theorem, 1885) *Let n points ξ_i, $i = 1, 2, ..., n$, be randomly distributed on a domain S and let H be some event (property) that depends on the positions of these points. Let $S' \subset S$ such that δS is a part of S not in S'. Then the following relation can be used to find the probability of certain point arrangements*

$$dP(H) = n(P(H|\xi_1 \in \delta S) - P(H))S^{-1}dS. \tag{B.1}$$

Let P denote the probability that two points are separated by a distance between x and $x + \Delta x$ (see Figure B.1). P_1 denotes the same probability given that one of the points is on the circumference of the circle. Thus, in our case B.1 simplifies to

$$dP = 2(P_1 - P)\frac{dA}{A}, \tag{B.2}$$

where A is the area of the circle, i.e. $A = \pi r^2$ and $dA = 2\pi r dr$. Observe Figure B.1, where illustration of the problem is presented. When one point is on the circumference dA, for two points to be separated by x another point must be exactly x distance away. This implies it should reside on a section of an annulus. When dx is infinitesimally small the area of the annulus is $2\phi x dx$, where ϕ is readily found to be $\arccos(x/2r)$. Thus, P_1 can be found as

$$P_1 = \frac{2x dx \arccos(x/2r)}{\pi r^2}. \tag{B.3}$$

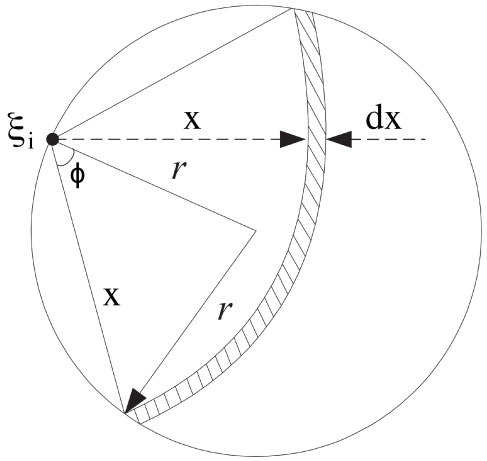

Figure B.1: Here is a schematic depiction of the circular section with radius r. The angle $\phi = \arccos(x/2r)$ is shown and ξ_1 denotes one point on the circumference used in calculating P_1.

Substituting B.3 in B.2 we get

$$dP = \left(\frac{2x dx \arccos(x/2r)}{\pi r^2} - P \right) \frac{4dr}{r}. \tag{B.4}$$

Rearranging terms and integrating both sides we have

$$Pr^4 = \frac{4x^2dx}{\pi}\int \frac{2r}{x}\arccos\left(\frac{x}{2r}\right)dr \tag{B.5}$$

$$= \frac{4x^2dx}{\pi}\left(2r^2\arccos\left(\frac{x}{2r}\right) - xr\sqrt{1-\frac{x^2}{4r^2}}\right) + C \tag{B.6}$$

where C is the integration constant. For $r = \frac{l}{2}$ two points have to fall on the circumference diametrically across and this event has probability 0. Therefore, for $r = \frac{l}{2}$ we shouldd have $P = 0$. Substituting this into B.6, we get $C = 0$, thus the probability distribution function of the distances x between random points is given by [18]

$$p_r(x) = \frac{2x}{r^2}\left(\frac{2}{\pi}\arccos\left(\frac{x}{2r}\right) - \frac{x}{\pi r}\sqrt{1-\frac{x^2}{4r^2}}\right) \qquad 0 < x < 2r. \tag{B.7}$$

Mean and variance are computed using the probability distribution function $p_r(x)$ resulting in

$$< x >= \frac{128r}{45\pi} \qquad \sigma^2 = r^2 - \left(\frac{128r}{45\pi}\right)^2 \tag{B.8}$$

Bibliography

[1] Schmelzer J., Rpke G., Mahnke R. *Aggregation phenomena in complex systems* 1 ed. (Wiley-VCH, 1999).

[2] Lin M. Y., Lindsay H. M., Weitz D. A., Ball R. C., Klein R., Meakin P. *Universality in colloid aggregation* (Nature 339, 360 - 362, 1989).

[3] Witten T. A., Pincus P. A. *Structured Fluids: Polymers, Colloids, Surfactants* (Oxford University Press, USA 2004).

[4] S. R. Forrest, T. A. Witten *Long-range correlations in smoke-particle aggregates* (Jr 1979 J. Phys. A: Math. Gen. 12 L109).

[5] T. A. Witten, L. M. Sander *Diffusion-Limited Aggregation, a Kinetic Critical Phenomenon* (Phys. Rev. Lett. 47, 14001403, 1981).

[6] Tamas Vicsek *Fractal Growth Phenomena* 2 ed. (World Scientific Pub Co Inc, 1992).

[7] Bernhard Kramer *Advances in Solid State Physics, Volume 41* (Springer,2010).

[8] Klein A. M., Doupé D. P., Jones P. H., Simons B.D. *Kinetics of cell division in epidermal maintenance* (Physical Review E 76, 021910, 2007).

[9] Antal T., Krapivsky P. L. *Exact solution of a two-type branching process: clone size distribution in cell division kinetics* (J. Stat. Mech. P07028, 2010).

[10] Cesar Cobaleda C., Vicente-Dueas C., Sanchez-Garcia I. *Cancer Stem Cells* (Encyclopaedia of Life Sciences. John Wiley & Sons, 2007).

[11] Bonnet Dominique, John E. Dick *Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell.* (Nature Medicine 3, 730 - 737; 1997).

[12] Jeffrey M. Rosen, et al. *The Increasing Complexity of the Cancer Stem Cell Paradigm.* (Science 324, 1670; 2009).

[13] Mani S.A, Guo W., Liao M.J., Eaton E.N., Ayyanan A., Zhou A.Y., Brooks M., Reinhard F., Zhang C.C., Shipitsin M., Campbell L.L., Polyak K., Brisken C., Yang J., Weinberg R.A. . *The epithelial-mesenchymal transition generates cells with properties of stem cells.* (Cell, 133(4):704-15, 2008).

[14] Singh S.K., Clarke I.D., Terasaki M., Bonn V.E., Hawkins C., Squire J., Dirks P.B. *Identification of a cancer stem cell in human brain tumors* (Cancer research 63 (18): 58218, 2003).

[15] Al-Hajj M., Wicha M.S., Benito-Hernandez A., Morrison S.J., Clarke M.F. *Prospective identification of tumorigenic breast cancer cells* (Proceedings of the National Academy of Sciences of the United States of America 100 (7): 39838, 2003).

[16] O'Brien C.A., Pollett A., Gallinger S., Dick J.E. *A human colon cancer cell capable of initiating tumour growth in immunodeficient mice* (Nature 445 (7123): 10610, 2007).

[17] Zhang S., Balch C., Chan M.W., Lai H.C., Matei D., Schilder J.M., Yan P.S., Huang T.H., Nephew K.P. *Identification and characterization of ovarian cancer-initiating cells from primary human tumors* (Cancer research 68 (11): 431120, 2008).

[18] Li C., Heidt D.G., Dalerba P., Burant C.F., Zhang L., Adsay V., Wicha M., Clarke M.F., Simeone D.M. *Identification of pancreatic cancer stem cells* (Cancer research 67 (3): 10307, 2007).

[19] Lang Sh., Frame F., Collins A. *Prostate cancer stem cells* (J. Pathol. 217 (2): 299306, 2009).

[20] Civenni G., Walter A., Kobert N., Mihic-Probst D., Zipser M., Belloni B., Seifert B., Moch H., Dummer R., van den Broek M., Sommer L. *Human CD271-Positive Melanoma Stem Cells Associated with Metastasis Establish Tumor Heterogeneity and Long-Term Growth* (Cancer Res. 71 (8): 3098109, 2011).

[21] Caterina La Porta *Cancer Stem Cells: Light and Shadows.* (Stem Cell, Regenerative Medicine and Cancer, pp.513-525; 2010).

[22] Caterina La Porta *Cancer Stem Cells: Lessons From Melanoma.* (Stem Cell Rev and Rep 5:6165; 2009).

[23] Kimmel M., Axelrod D. E. *Branching Processes in Biology* 1 ed. (Springer, New York 2002)

[24] H. W. Watson and Francis Galton *On the Probability of the Extinction of Families* (The Journal of the Anthropological Institute of Great Britain and Ireland, Vol. 4, (1875), pp. 138-144).

[25] Dou, J., Pan, M., Wen, P., Li, Y., Tang, Q., Chu, L., et al. *Isolation and identification of cancer stem cell-like cells from murine melanoma cell lines.* (Cell molecular immunology, 4, 467472; 2007).

[26] Schatton, T., Murphy, G. F., Frank, N. Y., Yamaura, K., Waaga- Gasser, A. M., Gasser, M., et al. *Identification of cells initiating human melanomas.* (Nature, 451, 345349; 2008).

[27] Fang, D., Nguyen, T. K., Leishear, K., Finko, R., Kulp, A. N., Hotz, S., et al. *A tumorigenic subpopulation with stem cell properties in melanomas.* (Cancer research, 65, 93289337; 2005).

[28] Klein, W. M., Wu, B. P., ZHao, S., Wu, H., Klein-Szanto, A. J., and Tahan, S. R. *Increased expression of stem cell markers in malignant melanoma.* (Modern pathology, 20, 102107; 2007).

[29] Hadnagy, A., Gaboury, L., Beaulieu, R., and Balicki, D. *SP analysis may be used to identify cancer stem cell population.* (Experimental cell research, 312, 37013710; 2006).

[30] Keshet, G. I., Goldstein, I., Itzhaki, O., Cesarkas, K., Shenhav, L., Yakirevitch A., et al. *MDR1 expression identifies human melanoma stem cells.* (Biochemical and biophysical research communications, 368, 930936; 2008).

[31] Monzani E, Facchetti F, Galmozzi E, Corsini E, Benetti A, et al. *Melanoma contains CD133 and ABCG2 positive cells with enhanced tumorigenic potential.* (Eur J Cancer 43: 935946; 2007).

[32] Rouzbeh Taghizadeh, Minsoo Noh, Yang Hoon Huh, Emilio Ciusani, Luca Sigalotti, Michele Maio, Beatrice Arosio, Maria R. Nicotra, PierGiorgio Natali, James L. Sherley, Caterina A. M. La Porta *CXCR6, a Newly Defined Biomarker of Tissue-Specific Stem Cell Asymmetric Self-Renewal, Identifies More Aggressive Human Melanoma Cancer Stem Cells* (PLoS ONE 10/2010 Vol. 5-12).

[33] Dietrich Stauffer, Amnon Aharony *Introduction to percolation theory*, 2nd edition (Taylor & Francis, London 1994), page 110.

[34] Tamàs Vicsek *Fractal growth fenomena*, 2nd edition (World Scientific, Singapore 1989), page 186.

[35] Paul Meakin *Fractals, scaling and growth far from equilibrium* (Cambridge University Press, Cambridge 1998), page 184.

[36] Antonio Bru, Juan Manuel Pastor, Isabel Fernaud, Isabel Bru, Sonia Melle, Carolina Berenguer *Super-Rough Dynamics on Tumor Growth* (Physical Review Letters 11/1998 Vol. 81-18).

[37] Solomon Herbert *Crofton's Theorem and Sylvester's Problem in Two and Three Dimensions* (Standford University, 1978).

[38] Blumenfeld D. *Operations Research Calculations Handbook* (CRC Press, 2001).

[39] Grinstead C. M., Snell J. L. *Introduction to Probability* 2 rev. ed. (American Mathematical Society, 1997)

[40] Harris T. E. *The Theory of Branching Processes. Die Grundlehren der Mathematischen Wissenschaften 119* (Springer, Berlin 1963).

[41] Krishna B. Athreya, P. E. Ney *Branching Processes* (Dover Publications, 2004).

[42] Feller W. *An Introduction to Probability Theory and Its Applications*, 3rd Edition (Wiley, 1968).

[43] Harris, T. E. *Branching processes* (Ann. Math. Statistics 19 474494, 1948)

[44] Hayflick L., Moorhead PS *The serial cultivation of human diploid cell strains* (Exp Cell Res 25 (3): 585621, 1961).

[45] Caterina A. M. La Porta, Stefano Zapperi, James P. Sethna *Senescent Cells in Growing Tumors: Population Dynamics and Cancer Stem Cells* (Plos Computational Biology, 2012)

[46] Collado M., Gil J., Efeyan A., Guerra C., Schuhmacher A.J., Barradas M., Bengura A., Zaballos A., Flores J.M., Barbacid M., Beach D., Serrano M. *Tumour biology: Senescence in premalignant tumours* (Nature 436, 642, 2005).

[47] Thomas R. Yeager, Sandy DeVries, David F. Jarrard, Chinghai Kao, Stephen Y. Nakada, Timothy D. Moon, Reginald Bruskewitz, Walter M. Stadler, Lorraine F. Meisner, Kennedy W. Gilchrist, Michael A. Newton, Frederic M. Waldman, Catherine A. Reznikoff *Overcoming cellular senescence in human cancer pathogenesis* (Genes Dev. 1998 Jan 15;12(2):163-74, 1998).

[48] Collado M., Serrano M. *Senescence in tumours: evidence from mice and humans* (Nature Reviews Cancer 10, 51-57, 2010).

Tesi di Laurea di: Massimiliano Maria Baraldi
Relatore: Prof. S. Caracciolo
Correlatore: Dott. S. Zapperi
Correlatore: Dott.ssa C. La Porta
Codice PACS 87.17.-d

STATISTICAL METHODS IN CELL CLUSTER ANALYSIS

La formazione di aggregati in natura è il risultato della dinamica complessa di una molteplicità di particelle. Tali fenomeni sono di ordinario interesse nello studio dei Sistemi Complessi e in Biofisica e la loro comprensione ha portato ad importanti sviluppi nella fisica dello Stato Solido [1].

L'oggetto d'indagine di questa tesi consiste in cluster di cellule tumorali a diversi stadi di crescita. Nella prima parte della tesi si propone un metodo sistematico per l'analisi e la caratterizzazione di cluster bidimensionali compatti, mentre nella seconda parte si adotta la teoria dei "Branching Processes" (BP) [2] per interpretare la dinamica di crescita degli aggregati cellulari.

Questa ricerca è motivata dal recente sviluppo di una teoria interpretativa per la crescita tumorale, nota come Cancer Stem Cell (CSC) Theory [3, 4]: secondo questa la massa tumorale è composta da un ristretta popolazione di cellule, dette cellule staminali tumorali, in grado di proliferare a lungo e quindi responsabili dello sviluppo del tumore, mentre la maggioranza delle cellule tumorali sono dotate di un limitato potenziale di proliferazione. Questa teoria sta riscuotendo crescente successo all'interno della comunità scientifica in contrapposizione alla teoria tradizionale che prevede l'esistenza di una singola popolazione cellulare nei tumori [5].

In questo lavoro di tesi, discutiamo dei metodi di analisi per la crescita di cluster di cellule tumorali. La tecnica adottata coinvolge concetti di Meccanica Statistica e Sistemi Complessi con particolare attenzione alla teoria BP.

Nel Capitolo 1, presentiamo la tipologia di dati che si intende analizzare e discutiamo un metodo sistematico per il calcolo delle osservabili sperimentali. La tecnica che abbiamo sviluppato per lo studio di cluster cellulari è del tutto generale e può essere adattata a diverse tipologie di cluster bidimensionali.

I dati sperimentali dei quali disponiamo consistono in cluster di cellule tumorali su capsule di Petri circolari. Per ottenere questi dati, sono state disposte delle cellule sopra le capsule e ognuna di esse ha dato origine ad un processo evolutivo in grado di formare cluster. Quindi un'istantanea dello stadio evolutivo degli aggregati cellulari è stata ottenuta mediante la tecnica biologica del cristal-violetto in grado di fissare e colorare le cellule.

Abbiamo ottenuto un'immagine digitale della capsula con un semplice scanner, che poi abbiamo modificato in modo tale da ottenere cluster neri su sfondo bianco, eliminando il rumore di fondo con una tecnica di edge-detect. L'immagine ottenuta è convertita in una matrice booleana dove 1 corrisponde ad un sito occupato da un cluster e 0 ad un sito vuoto. Dopodiché abbiamo introdotto un algoritmo [6] in grado di etichettare ogni sito occupato con un numero rappresentativo del cluster di appartenenza, ottenendo una matrice dove 0 definisce un sito vuoto, mentre gli altri interi etichettano i cluster. In questo modo le osservabili possono essere facilmente calcolate a partire da tale "matrice di etichette".

Nel Capitolo 2 si discutono le problematiche relative al setup sperimentale e al metodo di calcolo descritto nel capitolo precedente. Lo scopo di questa parte del lavoro consiste nel verificare che i cluster non interagiscono tra loro e possono essere trattati come entità indipendenti. A tal fine è necessario effettuare delle misure che tengano conto dell'insieme delle particelle all'interno delle capsule.

Abbiamo verificato che non sono presenti interazioni fra i diversi aggregati cellulari, studiando la distribuzione delle distanze tra i baricentri dei cluster e confrontando il risultato con un sistema di particelle distribuite casualmente in una regione circolare delle stesse dimensioni di quella considerata.

La tecnica adottata per il calcolo delle osservabili è stata discussa effettuando delle misure sul numero di cluster e mettendole in relazione con il numero di cellule originariamente presenti. Come ci si aspetta biologicamente il numero di cluster risulta mediamente inferiore al numero di cellule di

nucleazione.

Ci siamo infine chiesti se fosse presente una possibile interazione delle cellule, e preciò dei cluster, con le pareti delle capsule di Petri nelle quali sono vincolate. Per rispondere a questo quesito sono state messe a confronto le distribuzioni rinormalizzate dei volumi dei cluster, ottenute per diversi valori del numero di cellule iniziali e si osservato che tutte danno origine allo stesso tipo di distribuzione.

Si è quindi appurato che i cluster sono indipendenti tra di loro. Il Capitolo 3 riguarda invece l'evoluzione geometrica dei vari aggregati, al fine di stabilire se il processo è isotropo oppure è possibile identificare una direzione preferenziale di crescita. Abbiamo definito, a partire dal calcolo del tensore di inerzia, delle misure che descrivono l'eccentricità e l'orientazione di oggetti bidimensionali. Si è verificato che i cluster seguono una crescita casuale in accordo con le previsioni del modello di Eden [7] e che la distribuzione degli angoli di orientazione è uniforme.

I risultati ottenuti suggeriscono che le cellule sono entità indipendenti tra loro. Questa è un'ipotesi fondamentale se intendiamo studiare la dinamica evolutiva dei cluster utilizzando modelli ispirati alla teoria BP. Questa teoria descrive infatti la dinamica stocastica di una popolazione di individui che seguono le stesse regole evolutive ma che si comportano indipendentemente tra loro.

Nel Capitolo 4 si introducono i concetti fondamentali della teoria. Si approfondisce in modo particolare il caso di processi a tempo discreto (processi di Galton-Watson) e si espongono i principali risultati che possono essere ottenuti nel regime asintotico. Questa limitazione introduce inevitabilmente la necessità di simulazioni che verranno sviluppate per comparare i modelli con i dati sperimentali. Si discute quindi un modello a singola popolazione coerente con la teoria tradizionale per la crescita tumorale e un modello a più popolazioni basato sui recenti sviluppi della teoria CSC [8].

Il Capitolo 5 riguarda il confronto dei suddetti modelli con le misure effettuate a partire dai cluster di cellule tumorali, utilizzando il metodo descritto nel primo capitolo. Analogamente alla teoria della percolazione, abbiamo analizzato le distribuzioni dei volumi dei cluster, poiché tengono in considerazione della composizione globale dei campioni analizzati. Secondo l'analisi

condotta, il modello "tradizionale" risulta inadatto a riprodurre i risultati sperimentali e si presenta la necessità di supporre l'esistenza di due popolazioni. Infatti tale ipotesi è in accordo con le misure ottenute ed è coerente con una teoria CSC nella quale esiste una ristretta sottopopolazione di cellule con probabilità di morte nulla. Infine discutiamo le problematiche che si presentano nel supporre l'esistenza di una gerarchia tra le due popolazioni cellulari. Se fosse possibile infatti verificare in un futuro la correttezza di quest'ipotesi, si otterrebbe un ulteriore conferma della teoria CSC.

Riferimenti bibliografici

[1] Schmelzer J., Rpke G., Mahnke R. *Aggregation phenomena in complex systems* 1 ed. (Wiley-VCH, 1999).

[2] Kimmel M., Axelrod D. E. *Branching Processes in Biology* 1 ed. (Springer, New York 2002).

[3] Bonnet Dominique, John E. Dick *Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell* (Nature Medicine 3, 730 - 737; 1997).

[4] Jeffrey M. Rosen, et al. *The Increasing Complexity of the Cancer Stem Cell Paradigm* (Science 324, 1670; 2009).

[5] Cesar Cobaleda C., Vicente-Dueas C., Sanchez-Garcia I. *Cancer Stem Cells* (Encyclopaedia of Life Sciences. John Wiley & Sons, 2007).

[6] Dietrich Stauffer, Amnon Aharony *Introduction to percolation theory* 2 ed. (Taylor & Francis, London 1994).

[7] Tama‘s Vicsek *Fractal growth fenomena* 2nd edition (World Scientific, Singapore 1989).

[8] Caterina A. M. La Porta, Stefano Zapperi, James P. Sethna *Senescent Cells in Growing Tumors: Population Dynamics and Cancer Stem Cells* (Plos Computational Biology, 2012)