Stochastic modelling of dynamical systems in biology

Pellin, Danilo

Chapter 1

Introduction

Statistics in computational biology

In several research contexts, technological advancement allows the investigation of processes with a level of details previously inconceivable. Examining phenomenon at microscopic level rather than macroscopic, on the one hand provides a better understanding of the underlying mechanisms but, on the other, often requires the modeling approach to be able to take into account for the potential impact that intrinsic heterogeneity might be present. In biological research for example, the emerging usage of *single-cell* protocols, such as single-cell RNA/DNA sequencing represents a promising tool to shed light on dynamics occurring within and between cells but, at the same time, unravel an unexpected diversity in cell sub-populations, so far considered homogeneous.

In many different fields, the higher resolution (in time, space or any other domain) provided by novel instruments is usually accompanied by an increment of the amount of data generated, giving birth to what is now known as the *big data* era. Time has come that individuals and mainly organizations have to tackle the problem of how to process large amounts of data in support of their respective needs and operations, aiming at improving their handling and response efficiency. Contrary to popular belief, the combination of detailed information and big data is not a *panacea* and the need for modelling and inference remains. New methods and algorithms are required to enable the exploitation of the available massive, multi-dimensional, multi-source, time varying information available. It is not possible to answer questions about complex systems relying solely on numbers, without a profound knowledge of the issue, perspective and sometimes intuition.

In modern science, to conduct a thorough investigation about a relevant scientific hypothesis, is often necessary that different disciplines interact and collaborate with each other. A research project metaphorically represents a *meeting point* where scientists bring their own baggage, share methods and ideas to reach a common final goal. In many cases, interdisciplinarity becomes essential. Neuroscience, bioengineering, data science, bioinformatics are all examples. In computational biology different disciplines, such as biology, genetics, genomics, mathematics, statistics, computer science, chemistry join forces to make new predictions or to discover new biology. This knowledge exchange is not always easy. Researchers are asked to learn terminologies and concepts they are not familiar with, in order to be able to share research questions with colleagues with different backgrounds.

But, what is the statistics contribution in computational biology? Statistics is concerned with the use of data in the context of uncertainty, and most of the biological processes are driven by a stochastic component. The majority of the fundamental steps necessary to address a biologically relevant question needs statistical expertise and reasoning. Crucial is the initial correct *translation* of the biological goal in a statistical problem. Strictly connected is the design of an appropriate experiment, able to guarantee a sufficient amount and quality of collected data, crucial to provide strength and reliability of findings. At the modelling stage, several aspect must be considered, such as the possible presence of underlying dependences, hidden unobservable entities and temporal dynamics among others. Finally, the representation of the results by means of intuitive visualization tools has a central role in current research, especially when the dimensionality is large and incisive messages must be communicated to users without a quantitative expertise.

Many biological processes can be modelled as dynamical systems involving sequences of biological events. Finding a feasible way to model their evolution represents the strong contribution that statistical theory can give in understanding the mechanisms underlying complex systems. In this thesis, statistical models and inferential procedures for the analysis of data derived from experiments performed with the most recent and advance techniques, will be presented. Two fundamental biological processes such as hematopoiesis in humans and the mechanisms underlying the information exchange occurring between connected neurons will be investigated.

Modelling human hematopoiesis

Over the past decade, gene therapy has proved its potential as a next generation therapy for many diseases with unmet clinical need. Gene therapy can be exploited to overcome a cellular defect due to a mutated gene, providing a fully functional copy of it or to equip target cells with a new cellular function through genetic engineering. Most clinical approaches are based on the delivery of exogenous DNA molecules by viral vectors using, when stable gene transfer is needed, retrovirus- or lentivirus-derived systems. This fascinating process begins after the virus enters the cell and its RNA genome is reverse transcribed into a double-stranded DNA. DNA contains at its termini particular sequences specifically recognized by the integrase, an enzyme produced by virus leading the integration process. Depending upon the virus type, pre-integration complexes, composed by a mixture of viral and host enzyme and proteins, enter the nuclei of non-dividing cells through the nuclear pore (e.g., human immunodeficiency virus, HIV) or wait until the nuclear membrane dissolves during cell division (e.g., Moloney murine leukemia virus MoMuLV). Once the pre-integration complex associates with the host chromosome, viral integrase catalyzes the insertion of the viral sequences into the host DNA. It is worth noting that integration site selection is a stochastic, albeit not completely random, process.

A positive side effect of treating cell with virus derived vector is that integration site coordinates can be used as marker to track the fate of individual cells after re-infusion in patients body. This innovative methodology, named in-vivo clonal tracking, exploits the fact that all cells that result from proliferation and differentiation of a corrected cell will carry identical markings. With the advent of the *sequencing era* and the refinement of antibodies based sorting protocols, it has been possible not only to increase the size of molecular markers simultaneously traceable, but also to estimate the amount of cells, over multiple lineages at different stage of maturation, deriving from individual transplanted cells.

The data at our disposal regards three patients recruited in a gene therapy clinical trial for Wiskott–Aldrich syndrome (WAS), a rare X-linked recessive blood disorder characterized by a severe immunodeficiency. Despite being efficacious and holding a life-long curative potential, integrating vector platforms come with an inherent risk of inducing insertional mutagenesis, consisting in the activation of oncogenes by the nearby integration of a vector. By means of integration site analysis on patients bone marrow and peripheral blood samples performed at different time points during the follow-up period, it is not only possible to assess the clonal evenness generated by treated hematopoietic stem/progenitor cells and to detect dangerous, abnormal clonal expansion, but also to study how this particular cells sub-population continuously replenish all classes of blood cells. Clearly, this data set represents an unprecedented opportunity to interrogate hematopoiesis in humans. Through appropriate statistical modelling, able to realistically mimic single clone dynamics, it would be possible to improve our knowledge about several open questions, such as the structure of hematopoietic tree and lineage specific dynamics.

We propose to model individual clone evolution over time as an multivariate continuous time Markov process, where each component corresponds to the number of cells of a specific type generated by a re-infused marked cell. Clone trajectories are then determined by a random sequence of cellular events, grouped in three main categories: cell duplication, death and differentiation. It is known that any Markov process can be defined by means of a *master equation*, a formulation widely used in many applied contexts in which process state space has a direct, physical interpretation. Master equation is strictly connected to the Kolmogorov forward equation and similarly consists in set of ordinary differential equations (ODEs) describing the time evolution of the probability of a system to occupy each possible state configuration. Despite analytical exact solutions of the master equation are possible to calculate only for special, simple cases, for systems of arbitrary complexity it can be exploited to retrieve important information about the behaviour of process characterizing indexes, such as moments. In this thesis moments equations will be used as starting point to develop method-of-moments procedures that allows us make inference on cellular event rates governing the hematopoiesis differentiation process in humans, as well as perform structural learning.

Modelling spontaneous neurotransmitters release at synaptic level

The summation of synaptic inputs mediated by a neuron and the impact that it produces on the activity of a neuronal network is one of the most cutting-edge topics in neuroscience. A real comprehension is limited by our approximative knowledge about the neurophysiological mechanisms underlying the generation of synaptic signals, either evoked by action potentials or spontaneous. Several studies prove that synaptic transmission at chemical synapses has a quantal nature. Chemical synapses of both the central and the peripheral nervous systems are able to exchange information thanks to the release of discrete amount of neurotransmitters, also known as synaptic quanta, packed in anatomical elements represented by synaptic vesicles. Both the action potential evoked and the spontaneous release process share a common pool of synaptic vesicles and this duality of transmission based on the same source of quanta explains the complex interaction between these two processes. Although much is known about the molecular elements involved in synaptic vesicles dynamics and the cascade of maturation steps, still little is clear about the process as a whole, how it is modulated in response to different stimuli and its physiological relevance. Most importantly, still missing is a comprehensive model able to establish a connection between vesicle cycle and neurotransmitters release measured at synaptic level.

From a practical perspective, quanta discharges are detected by means of the analysis of postsynaptic membrane voltage or current fluctuations that correspond to exocytosis, the mechanism by which vesicle membrane fuses with and becomes part of the outer cell membrane, causing its neurotransmitter load to be expelled and bind to postsynaptic receptors. In order to guarantee the conformational stability of the presynaptic terminal, exocytosis is followed by the retrieval of synaptic vesicles by endocytosis. These two biological processes occur at different and dedicated areas of the terminal, where vesicles come into contact with specific membrane proteins complexes. To reach the release site where exocytosis occurs, vesicles are transported over a moderately small distance by motor proteins. Once vesicles arrive in close proximity to release site, they come into contact with tethering factors that can restrain them in a docking status and enhance their fusion. The study of spontaneous quanta release offers the opportunity to study neuron communication at the level of its most elementary operating systems. The most popular approach to study the statistics of quanta release is to analyse the distribution of inter-quanta (inter-event) intervals. Strong divergence from the Poisson nature initially postulated has been accumulated in the years, encouraging the formulation of alternative signal generating models, primary motivated by the necessity to find a reasonable explanation for the presence of bursting episodes followed by long periods of synaptic inactivity.

In this thesis, a new comprehensive model underlying the generation of spontaneous quanta release time series is proposed. It has been derived from an extensive review of the literature available on each of the component contributing to determine the vesicle lifespan, with particular attention for those reporting statistical consideration and analysis. Through an appropriate model for vesicles motion, motivated and supported by experimental data published in various works, it is possible to mimic the characteristics shape of inter-events intervals distribution. By means of dedicated quantile-based inferential procedures and a set of experimental recordings regarding spontaneous activity in primary cultures of rat hippocampal neurons, we compare the reliability of different alternative models of vesicle dynamics currently debated in the neuroscientific literature.

Thesis outline

In the following chapters of this thesis, research questions will be *in primis* introduced from a biological perspective. Dedicated statistical models are described, along with their assumptions and novelties. Particular efforts is devoted to the development of suitable inferential procedures that, in combination with data collected from real experiments, provide estimation for biologically relevant parameters. Two main biological problems will be investigated. The second, third and fourth chapters of this thesis are focused on the study of hematopoiesis in humans. The fifth chapter studies of the basics of neuronal communication. A description of their content follows.

In the **second chapter**, the transition rate matrix of the Markov process modelling clone evolution has been defined through a set of linear function involving event rates and sub-population sizes. This formulation allows the analytical derivation of two coupled sets of ODEs for expected lineage specific cell counts and their variance-covariances, useful to developed a method-of-moments estimation procedure. In order to favour computational efficiency, ODEs solutions have been calculated by means of Euler's method, yielding an exact representation of moments dynamics under a frequent sampling time setting. Estimation of rates requires to limit the parametric space to non-negative values. To take into account this restriction, a least square based objective function has been optimized by means of an iterative quadratic programming approach. To encourage realistic solutions with to a progressive loss of differentiation potential moving from the top of hematopoietic hierarchy towards the committed lineages, we introduced a sparsity-inducing penalization terms in the objective function, in order to force a subset of differentiation rates to be zeros. The smoothly clipped absolute deviation (SCAD) penalty has been preferred to the least absolute shrinkage and selection operator (LASSO) since, at the cost of an additional tuning parameter, it can produce sparse solutions and approximately unbiased estimations for large coefficients.

The assumptions of linearity on both rates and sub-population sizes imply an unfeasible possibility of unlimited clone growth. To overcome this limitation, in the **third chapter**, transitions probabilities are defined as polynomial functions of cell counts. By considering a quadratic relationship for death event probabilities, a logistic behaviour for lineages growth curves is obtained. Logistic function is a common model in wide range of fields and is conceived to describe the self-limiting growth of a biological population. It is particularly suitable for modelling clones evolution in patients, since it is characterized by an initial stage of an approximately exponential growth, corresponding to the reconstitution of hematopietic heritage, followed by a steady state equilibrium. From an inferential perspective, the presence of polynomial terms leads process moments to be described by an infinite system of ODEs, approximated by an finite system under specific distributional assumptions. A solution is calculated by means of a numerical approach, rather than approximated as proposed in the second chapter, allowing for a remarkable improvement in terms of rates estimation performance at the cost of additional computational effort. Based on the analysis of a gene therapy data set, both the models shown in the second and third chapters are consistent with recently published results derived from studies performed on non-humans primate.

In the **fourth chapter**, the last dedicated to the investigation of human hematopoiesis, various aspects regarding the activity of distinct transplanted stem cells have been analysed and commented from a biological perspective. The presence of clonal repopulating waves has been assessed, along with some degree of heterogeneity within the progenitor lineages and the contribution of different progenitor classes during early and late post-transplant phases. From a clinical point of view, the ability to provide a stable, long-term, hematopoietic output by manipulated cell through viral vectors protocols has been verified, supporting further development and extension of gene therapy applications for other diseases.

In the **fifth chapter**, a comprehensive model for the generation of spontaneous quanta release occurring at the level of a single synapse is described. The model considers the recorded events time series as the superimposition of an unknown, but fixed, amount of renewal processes, each corresponding to vesicle specific fusion history. A connection between vesicle cycle, simplified to a cascade of three maturation steps, and the observed sequence of release events is hypothesized. Since the pool of vesicle involved is small, asymptotic results for inter-events intervals distribution cannot be exploited. As an alternative, we propose an estimation procedure based on both Monte Carlo simulation and a quantile-based scoring function.

Experimental data sets analysis confirms that a small amount of vesicle are likely to contribute to spontaneous signal generation and suggests a short duration for both vesicle exo- and endocytosis.